The Acoustics of Place of Articulation in English Plosives

Daniel Timothy Pio Denis McCarthy

A thesis submitted for the degree of

Doctor of Philosophy

School of Education, Communication and Language Sciences

Newcastle University

June 2019

Abstract

This thesis investigates certain aspects of the acoustics of plosives' place of articulation that have not been addressed by most previous studies, namely:

1. To test the performance of a technique for collapsing $F2_{onset}$ and $F2_{mid}$ into a single attribute, termed $F2_R$. Results: $F2_R$ distinguishes place with effectively the same accuracy as $F2_{onset}$+$F2_{mid}$, being within ±1 percentage point of $F2_{onset}$+$F2_{mid}$ at its strongest over most of the conditions examined.
2. To compare the strength of burst-based attributes at distinguishing place of articulation with and without normalization by individual speaker. Results: Lobanov normalization on average boosted the classification of individual attributes by 1.4 percentage points, but this modest improvement shrank or disappeared when the normalized attributes were combined into a single classification.
3. To examine the effect of different spectral representations (Hz-dB, Bark-phon, and Bark-sone) on the accuracy of the burst attributes. The results are mixed but mostly suggest that the choice between these representations is not a major factor in the classification accuracy of the attributes (mean difference of 1 to 1.5 percentage points); the choice of frequency region in the burst (mid versus high) is a far more important factor (13 percentage-point difference in mean classification accuracy).
4. To compare the performance of some traditional-phonetic burst attributes with the first 12 coefficients of the discrete cosine transform (DCT). The motivation for this comparison is that phonetic science has a long tradition of developing burst attributes that are tailored to the specific task of extracting place-of-articulation information from the burst, whereas automatic speech recognition (ASR) has long used attributes that are theoretically expected to capture more of the variance in the burst. Results: the DCT coefficients yielded a higher burst classification accuracy than the traditional phonetic attributes, by 3 percentage points.

For my mother, Denise, and for Claire, Niall, Emer, and Anna.

Acknowledgements

Many thanks to Jalal Al-Tamimi for his supervision over the course of this work. In particular a big thank you for creating the scripts that extracted the data for the pilot study and main study and for providing the random-forest script. Thank you to Ghada Khattab for reading the chapters and providing feedback on them. Thank you to Damien Hall and Danielle Turton for your annual feedback on my progress. Thank you to my examiners Rachel Smith and Kai Alter for your advice and insight. Thank you to the Economic and Social Research Council for funding me. Thank you to everyone who volunteered to participate in my study – your data have been fascinating to me in ways you'll probably never know! Thank you Jasmine, Hana, Abdulkareem, Caitlin, Yiling, Noura, Na, Hajar, Hanh, Ella, Knight, and Maha for your warmth and friendship – I'm lucky to have met such terrific people.

And lastly thank you Mammy, Claire, Niall, Emer, and Anna for your love. This is for ye.

Contents

List of Figures

List of Tables

Chapter 1: Overview

1.1 Aims

The overall aim of the thesis is to investigate certain aspects of the acoustics of place of articulation in plosives that have not hitherto been examined by most previous studies. Within this overall aim are contained four specific aims:

1. To test the performance of a technique for collapsing $F2_{onset}$ and $F2_{mid}$ into a single attribute, termed $F2_R$. The development of this technique has been inspired by the observation that the frequencies of $F2_{onset}$ and $F2_{mid}$ tend to be closely correlated (the slope in a regression plot between the two typically being between ca. 0.4 and 0.75).
2. To compare the strength of burst-based attributes at distinguishing place of articulation with and without normalization by individual speaker. This is done because normalization by individual speaker of formant frequencies has been used widely whereas normalization of aperiodic events such as the burst has been less widely studied.
3. To examine the effect of different spectral representations (Hz-dB, Bark-phon, and Bark-sone) on the accuracy of the burst attributes. This allows us to examine whether and to what extent an auditorily-oriented acoustic approach can aid the classification accuracy of attributes.
4. To compare the performance of some traditional burst-based attributes with the first 12 coefficients of the discrete cosine transform (DCT). The motivation for this comparison is that phonetic science has a long tradition of developing burst attributes that are tailored to the specific task of extracting place-of-articulation information from the burst, whereas automatic speech recognition (ASR) has long used attributes that function on all kinds of spectral speech envelope and which are theoretically expected to capture more of the variance in the burst than the attributes that have been traditionally used in phonetic science. Hence the DCT features can serve as a benchmark against which to compare the performance of other burst features that have been developed in the history of phonetics.

The data collected for the present study consist of the speech of 20 speakers of British English reading sentences that were designed to sound natural while containing plenty of plosives (an average of ca. 14 plosives per sentence). These tokens, along with the preceding and following segments, were labelled and segmented manually with narrow, broad, and orthographic transcription, and a tier for comments on the typicality of tokens. These tokens numbered 7,147,

of which 6,284 were brought forward for analysis (taps and glottal stops, for example, were not analysed). The material features the consonants /pb td kg/ in a wide variety of phonetic contexts. In terms of statistics, the present study employs discriminant analysis for classification of small numbers of attributes (three or less) and random forests for larger combinations of attributes. Other statistics employed include mixed-effects modelling, Wilks's Lambda, with the McNemar test being used for determining whether the difference in two classifications is statistically significant.

1.2 General Discussion

Classifying the place of articulation of plosives correctly is a vexatious task, and the present study is far from being the final word on the topic. There has been a long history of looking at the acoustics of plosives' place of articulation in phonetics. One theme that will emerge from our exploration of this history is the sheer number of acoustic attributes that has been developed over the decades for extracting place-of-articulation information from the release burst. Most such attributes are tailored to the specific task of extracting features from the burst. This approach, which has been the dominant one in phonetics, is known as a knowledge-based approach (Abdelatty Ali et al., 2001; Suchato, 2004). This is a style of research that emphasizes the development of attributes based on the researcher's prior knowledge of a particular acoustic phenomenon. This approach is but one approach to the burst that can be taken; and this thesis – perhaps more than most previous studies on this topic in the discipline of phonetics – seeks to engage with attributes for the burst developed from a different perspective, that of the statistically-driven and data-driven approach that dominates in automatic speech recognition (ASR). To this end, the present study will compare the performance of some of the burst attributes traditionally used in phonetics with the first 12 coefficients of the discrete cosine transform (DCT), a feature set that underpins the MFCC front end used widely in automatic speech recognition (Davis and Mermelstein, 1980).

Seeking ideas outside of phonetics from automatic speech recognition is just one direction in which one can look. Another direction is hearing science. This thesis investigates spectral representations that have not been utilized by most previous studies of plosives' place of articulation in the field of phonetics, namely the Bark-phon and Bark-sone spectra. These representations roughly approximate some (though by no means all) of the aspects of how the amplitude and frequency of sound are represented in the auditory periphery. It is hoped that the use of such representations, though far from being a perfect simulation of the auditory periphery, will stimulate a greater interest in the use of auditorily-oriented acoustic representations in phonetic science than appears to be the case at present.

This attempt to incorporate insights from different disciplines also runs through the literature review which, in addition to presenting a detailed history of the burst and formant attributes used in phonetics, also presents the research of those who have approached the problem of plosive place of articulation from a more auditorily-oriented angle. In particular, we will explore the results of studies that have investigated speech in noise (Miller and Nicely, 1955; Li et al., 2010), modified the amplitude of the burst peak and studied listeners' responses (Kapoor, 2010; Kapoor and Allen, 2012), and modelled the effect of forward masking in plosive-vowel syllables (Xie, 2013) and compared the deterioration in listeners' responses when the burst peak is removed relative to the F2 transition (Cvengros, 2011). It is hoped that this bringing together of diverse perspectives will stimulate greater interest in the findings of researchers outside of traditional phonetics who have studied some of the same phenomena as ourselves.

Life is a game of trade-offs, and the present study is no exception. One trade-off that is pervasive in pattern recognition in general – whether speech, handwriting, or imagery – is that between maximizing classification accuracy on the one hand and minimizing the number of attributes on the other. This theme will run through Chapters 5 and 7. In general, the smaller the number of attributes used, the easier it is to know exactly what information the attributes are capturing. This tension between attribute accuracy and attribute interpretability is one factor that will help to explain the different trajectories that phonetics and automatic speech recognition have tended to take over their histories, as revealed by the comparison between the traditional-phonetic burst features and the 12 DCT coefficients (Chapter 7). The phonetician is typically aiming to develop attributes whose interpretation is straightforward and which thus aid our understanding of the acoustics of particular speech sounds. For example, spectral tilt, which subtracts the intensity of the most intense spectral component in the high-frequency region of the burst from the most intense spectral component in the mid-frequency region, is readily interpretable. The 12 DCT coefficients mentioned above are also interpretable though arguably not to the same extent (Section 2.3.1.9) as traditional-phonetic attributes such as spectral tilt.

This trade-off between attribute accuracy and attribute minimization / interpretability is a theme we will encounter in the study of the second and third formant transitions (Chapter 5). $F2_{onset}$ and $F2_{mid}$ are strongly correlated with each other and mash together information about both the plosive and the vowel and yet there have been surprisingly few sustained attempts to merge them into a single attribute. This is all the more surprising when one considers the early history of the study of plosive place of articulation, in which the locus theory (Delattre et al., 1955) sought to abstract from the observed $F2_{onset}$ frequencies something more reflective of

place of articulation, $F2_{locus}$. Hence the present study develops a procedure for exploring the space in which $F2_{locus}$ lies, known as $F2_R$. Although this attribute can be regarded as an exploration of a hypothetical, abstract frequency, it is perhaps better thought of as a means of collapsing $F2_{onset}$ and $F2_{mid}$ into a single attribute. The comparison of $F2_R$'s accuracy with that of $F2_{onset}$ + $F2_{mid}$ can thus be seen as another instance of the theme in this study of the tension between minimizing the number of attributes and maximizing classification accuracy.

Following the presentation of this study's results, Chapter 8 will scrutinize the extent to which the study succeeded in addressing its four main aims, and in the process will discuss the study's limitations so that future research may avoid them. Another respect in which Chapter 8 is geared towards future research is in its discussion of a method for improving $F2_R$. The aim of this presentation will be to stimulate interest in an approach that uses syllable-length averaging of F2 to adjust the frequencies in an F2 contour and, in so doing, show that coarticulation might not be as bewilderingly complex a phenomenon as is sometimes emphasized.

1.3 Structure

The present thesis consists of nine chapters (including the present one). In the next chapter the acoustics of plosives' place of articulation is introduced along with the most important previous research that has been conducted on the topic (Chapter 2, Introduction and Literature Review). The chapter begins by presenting the acoustics of plosives' place of articulation in a general way, introducing the reader to formant transitions and the release burst and the variation that these exhibit due to place of articulation (2.1). After this general introduction, the formant transitions (Section 2.2) and the burst (2.3) are each explored in greater detail. Within each of these sections there are subsections dealing with specific topics, e.g. the formant section begins with a discussion of the 1950s locus theory (2.2.1). Although this theory was formulated over 60 years ago, it is extremely important for the present study as it will form the backbone of the approach to formant information developed in Chapter 5, known as $F2_R$. $F2_R$ can be thought of a method for exploring the locus theory. The review of the locus theory will be followed by a review of the study that seemed to undermine the insight of the locus theory, namely Öhman (1966), who found that in a VCV sequence V1 has a considerable influence on $F2_{onset}$ such that the CV transition does not in fact point to a single frequency (2.2.2). The review of previous formant-based techniques continues with a review of locus equations (2.2.3), including the role of F3 in identifying place, and ends with an examination of some other formant-based research, such as the measurement of formants in the burst and the difference in frequency between F2 and F3 (2.2.4).

The second part of the literature review (2.3) is concerned with the release burst. The overarching theme of this section is that there has been longstanding uncertainty in the field of phonetics as to what attributes to use on the burst for extracting its place-of-articulation information, with the result that a menagerie of attributes for the burst have been developed. Consequently this section begins by examining a large selection of these attributes (2.3.1). The section also includes a discussion of the discrete cosine transform, a technique for maximizing the variance captured from a spectral envelope that has been widely used in ASR acoustic models but not in phonetic science. Due to its strong variance-capturing potential, it is argued that this feature set should serve as a benchmark against which to compare the burst attributes that have been developed in the history of phonetic science, a comparison which will be presented in Chapter 7.

Given the longstanding uncertainty in phonetics over what attributes to use on the release burst, the remainder of the literature review on the burst is concerned with gaining a clearer picture of what information in the burst is likely to be the most important. Thus a wide variety of perceptual studies will be reviewed, including: experiments in which syllables containing plosives are played to listeners with and without bursts and/or transitions (2.3.2 and 2.3.3), or in differing degrees of background noise (2.3.4 and 2.3.5), or with the burst peak amplified or attenuated (2.3.6), as well as a simulation of forward masking from the burst onto the formant transitions and how this might be affecting the perceptual prominence of $F2_{onset}$ (2.3.7). Taken together, these studies can help us understand what information in the burst is most important, i.e. the most robust and consequently least likely to be lost under real-life acoustic conditions. The chapter ends with a discussion of vowel normalization and how it might be applied to plosives (2.3.8).

Chapter 3 (Pilot Study) summarizes the findings of a preliminary small-scale study that was conducted prior to the main study. The principal aim of the study was to compare the performance of a variety of acoustic attributes, the weakest of which would be culled prior to the main study. The overall finding of the chapter is that for the formants, there is one attribute ($F2_R$) that is much more performant than the others, whereas for the burst-based attributes there are several approaches (TiltTotaldB, HiPeakdB, CoGdB, AllPeakHz) that perform within a few percentage points of each other. These findings motivate the approach of the main study, in which one formant-based technique ($F2_R$) will be examined extensively, whereas for the burst a wider range of attributes will need to be compared.

Chapter 4 (Methodology) discusses the design of the main study, including the theoretical motivation (Section 4.1), data collection (4.2), annotation and transcription (4.3), formant measurements (4.4), and burst measurements (4.5).

Chapter 5 (Formant Frequencies) presents the results for the formant-based attributes and corresponds to Aim 1 of this thesis. It begins with an examination of VCV sequences (Section 5.1). This examination was prompted by Öhman (1966) who found a considerable influence of V1 on $F2_{onset}$. However, due to the small scale of Öhman's study and the artificial nature of his nonce VCV sequences, the present study revisits VCV sequences with a larger number of speakers and with real speech to quantify the influence of V1 relative to V2 on $F2_{onset}$. The results of this examination will motivate the decision for the rest of Chapter 5 not to incorporate the effect of V1 in the design of $F2_R$.

Section 5.2 presents the rationale for collapsing $F2_{onset}$ and $F2_{mid}$ into $F2_R$, namely feature minimization. 5.3 introduces the statistic employed for measuring statistical significance in this thesis, and also sketches out the $F2_R$ formula with reference to the 1950s locus theory. With this formula in hand, what follows is the meat of the formant study (5.4).

This section examines the performance of $F2_R$ under a wide variety of conditions; over each of these conditions the aim is to compare the accuracy of $F2_R$ with $F2_{onset}$ + $F2_{mid}$ (Aim 1 of the thesis). 5.4.1 begins by presenting the results for $F2_R$ on CV transitions for /b d g/ and /p t k/, followed by the results for an analogous attribute on the F3 information, $F3_R$. 5.4.2 compares four kinds of methods of normalizing formant frequencies and brings forward the best of the four for the remainder of the chapter. In 5.4.3 the classifications for front vowels and back vowels is run separately in order to quantify its improvement on classification accuracy, and 5.4.4 compares front- and back-vowel contexts in greater depth. 5.4.5 quantifies the difference in classification accuracy of VC formant transitions relative to CV transitions. 5.4.6 compares the classification accuracy of formant distances (F2 – F1, F3 – F2) relative to $F2_R$. 5.4.7 investigates the $F2_R$ of pre-schwa tokens and contrasts the results with those for non-schwa tokens. 5.4.8 examines a more elaborate version of $F2_R$ that incorporates time in addition to frequency. 5.4.9 discusses the role of *c* in the $F2_R$ formula, while 5.4.10 integrates VC and CV information and quantifies its improvement of classification accuracy over using VC or CV alone.

Chapter 6 (The Burst) is concerned with the release burst and corresponds to Aims 2 and 3 of the present thesis. It begins by taking a bird's eye view of the burst using spectral slices that average together hundreds of tokens, which provides a sense of what a typical bilabial, alveolar, and velar burst look like (Section 6.1). Following this the burst attributes to be compared in the main study are presented (6.2). To gain a further feel for the data, mixed-effects modelling is used to model the variation in the burst that is caused by the following vowel's backness and stress (6.3).

Following this the burst attributes are compared and combined under a wide variety of conditions (6.4). This section begins by comparing the attributes when all tokens are inputted to a single classification. These results are then compared with the results when the voiced /b d g/ and voiceless /p t k/ bursts are classified separately. Likewise, a classification is examined in which the attributes are classified on prevocalic and non-prevocalic tokens separately. The difference in results between these various classifications indicate whether voicing and/or the following segment affect the acoustics of bursts sufficiently to warrant separate classification (6.4.1-6.4.3). Sections 6.4.4 to 6.4.7 address Aim 2 of the present thesis, namely to compare burst attributes with and without normalization by individual speaker. 6.5 is principally concerned with Aim 3 of the thesis, namely to compare the accuracy of the attributes from the three kinds of spectral representations (Hz-dB, Bark-phon, and Bark-sone).

Chapter 7 addresses Aim 4 of the thesis, namely to compare the traditional burst-based attributes of phonetic science with the DCT coefficients used widely in ASR acoustic models. The aim of this comparison is to get a clearer sense of where the traditional burst-based attributes perform relative to the DCT approach, an approach that was specifically designed to maximize the variance captured from a spectral envelope. The chapter compares these two attribute types to a third, cruder attribute type in which 12 samples from the burst at 12 equidistant frequencies are used as attributes. This third attribute group serves as a kind of null hypothesis in that it represents the information in the burst without any kind of acoustic attribute design (other than the decision of how many samples to take from the envelope).

The chapter begins by explaining the theoretical rationale for these different feature groups (Section 7.1). It continues by examining the classification accuracy of four attributes as more and more of the channels are removed from the spectrum. The result is a spectrum that becomes sparser and sparser. The aim is to identify how sparse the number of frequency channels in the burst can be made before the classification accuracy begins to deteriorate appreciably (7.2). These results are then used to decide how many samples were to be taken from the burst for the spectral-sample group of attributes mentioned above.

Section 7.3 presents the main results of the chapter. The subsequent section quantifies the improvement to classification accuracy yielded by burst duration, an attribute which relatively few previous studies seem to have utilized.

The chapter ends by combining the attributes from Chapter 5 with those from Chapters 6 and 7, which yields an overall picture of the accuracy of the entire set of attributes used in the present study (7.5). The trade-off between classification accuracy and feature minimization is discussed (7.6).

Chapter 8 (Discussion) brings together the results of the thesis and evaluates the extent to which the present study has fulfilled each of its four main aims. This evaluation includes highlighting the study's limitations. The chapter ends by sketching a method for improving $F2_R$ in future research following an analogy from the visual system (8.7).

Chapter 9 (Conclusions) summarizes the major findings of the present study and makes a few recommendations for future research in the area.

Chapter 2: Introduction and Literature Review

2.1 Introduction

2.1.1 The Aims of the Present Study

This study addresses certain gaps in the study of the acoustics of plosives' place of articulation. It does this using classification experiments to statistically compare the strength of various acoustic attributes at identifying plosive consonants' place of articulation, so as to ascertain which of the attributes (as well as which *combination* of the attributes) does this with the fewest errors. Along the way many attributes and compound attributes will be tested. The language examined is English (British English to be precise). English has six plosive consonants, namely /p t k b d g/[1]; we are concerned with how to acoustically distinguish /p b/ from /t d/ from /k g/.

Here are the four main aims of the study:

1. $F2_{onset}$ and $F2_{mid}$ tend to be strongly correlated with each other, which suggests that they could be combined into a single attribute. This study compares the classification accuracy of an attribute that collapses these attributes into a single attribute ($F2_R$) with the existing technique of keeping them as separate attributes, to see whether collapsing the attributes into the one attribute is viable.
2. Normalizing acoustic features by individual speaker is widely practiced on formant frequencies, but it remains an open question whether normalizing burst-based attributes would result in an improvement to the classification accuracy of such attributes, and if so, to what extent.
3. The fields of phonetics and ASR have tended to use different spectral representations. This study investigates the performance of attributes on three spectral representations (Hz-dB, Bark-phon, and Bark-sone) to see whether this has an affect on the classification of (burst-based) attributes.
4. There is a long history in phonetics of making tailor-made attributes specifically for the burst. However, there appears not to have been much comparison of the performance of such attributes with the kind of features used in automatic speech recognition. This is important because there are theoretical reasons to expect DCT coefficients (a technique which underpins the MFCC front end widely used in automatic speech recognition) to

[1] The phoneme /t/ may be realized as a glottal stop [ʔ] in certain contexts. The present thesis does not investigate this place of articulation; see Section 4.3.1 for details.

capture a larger amount of spectral variance than alternative approaches. This study compares the performance of the two attribute types on the burst.

2.1.2 Essential Concepts

Three important terms have just been encountered: plosive consonant, place of articulation, and acoustic attribute. What do each of these mean?

A plosive is a type of consonant in which there is a complete closure at some place in the vocal tract, formed by two articulators making contact (Laver, 1994: 205), which prevents air from exiting through the mouth. When the closure is oral, it is accompanied by another closure: the soft palate stays raised against the back wall of the pharynx, which prevents the air from escaping through the nasal cavity (ibid.). The result is that air pressure builds up in the mouth behind the obstructing active articulator. When the closure is released, the disparity in air pressure between the air behind the closure and in front of the closure is resolved by the air behind moving rapidly through the opening closure.

Place of articulation refers to the location of this closure. In /p b/ the closure is made by closing the lips (hence they are termed bilabial consonants); in /t d/ the closure is made by the tip or blade of the tongue touching the alveolar ridge (termed alveolar consonants); and in /k g/ it is made by the body of the tongue touching the velum (termed velar consonants, though the closure is in fact often located in front of the velum at the palate if the following segment requires the tongue body to be further forward, such as /i e j/). Although some languages have plosives at other places of articulation – Laver (1994: 206) recognizes no less than eight chief places of articulation for plosives – bilabial, alveolar/dental, and velar are the three most frequent places of articulation for plosives in the world's languages (Ladefoged and Maddieson, 1996: 43).

The plosives of English come in pairs: /p b/, /t d/, /k g/. Phonologically the plosives /p t k/ are termed voiceless, /b d g/ voiced. In a voiceless consonant the vocal folds are brought apart from one another (abducted), which means they cannot vibrate, whereas in a voiced consonant they are relatively close together, which permits them (in principle) to vibrate, i.e. to produce voicing.[2] In the present study we are not concerned with how to distinguish plosives by voicing, but rather by place of articulation. As such we are concerned with how to distinguish /p b/ from /t d/ from /k g/.

[2] In practice, English 'voiced' stops often lack voicing for aerodynamic reasons (termed devoicing), but that does not change the fact that these consonants are usually voiced in the sense of containing a voicing gesture (adduction of the vocal folds). For further discussion, see Docherty (1992) for British English and Lisker and Abramson (1964) and Davidson (2016) for American English.

An acoustic attribute (also termed a feature by some writers, e.g. Lyon (2017: 419-426)) is any measurement of the speech signal believed to be capable of distinguishing one segment from another. The number of ways of measuring a signal is infinite, and this study will only concern itself with a selection of candidate attributes that seem the most promising.

2.1.3 Acoustic Events

The plosive has been defined with reference to its articulation: an active articulator touches a passive articulator, creating a closure that is subsequently released. This set of articulatory events gives rise to several acoustic events. The two most important of these events for identifying place of articulation are the release burst and the formant transitions, but we will lay out all of the events.

Look at the syllable [mag] in Figure 1.1 below. As we glance at the image from the bottom upwards, a series of dark ridges can be seen running from left to right. These regions where the concentration of sound energy is high relative to neighbouring frequencies are termed formants (Liberman et al., 1954: 1).

Nothing in the known universe can move from one position to another instantly, and the articulators are no exception. Notice that towards the end of the [a] (in the blue frame of Figure 1.1), the acoustic influence of the [g] on the [a] gradually increases, as we can see from the changing frequencies of the formants. This part of a vowel in which the formants are changing due to the influence of a neighbouring consonant is known as the vowel's formant transitions. The formant transitions thought to be relevant to recognizing a consonant's place of articulation are the second and third formants (F2 and F3):

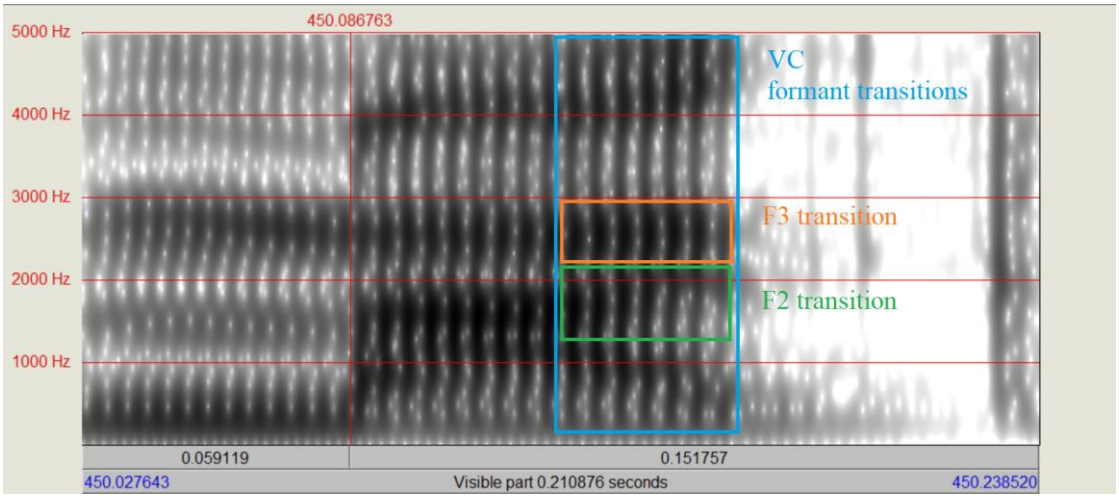


Figure 2.1: Spectrogram of the syllable [mag] from the word *magnifying*, uttered by a female speaker from Tyneside.

The movement of the formants in the second half of the vowel is highlighted in blue ("VC formant transitions"). Note the movement of the second formant (termed "F2 transition") and the movement of the third formant ("F3 transition"): these two formants are thought to be particularly important for identifying place of articulation.

Similarly, the formants of a vowel *after* a consonant also show some acoustic influence by the consonant (Figure 2.2). In this study we will term the formant transitions in the vowel *preceding* the plosive the "VC transitions" and the transitions in the vowel *after* the plosive the "CV transitions". Formant transitions are widely believed to carry information that identifies a consonant's place of articulation (Stevens, 1998, Chapter 7; Stevens et al., 1999; Abdelatty Ali et al., 2001; Suchato, 2004: 25). They are one of the two acoustic events examined in this study.

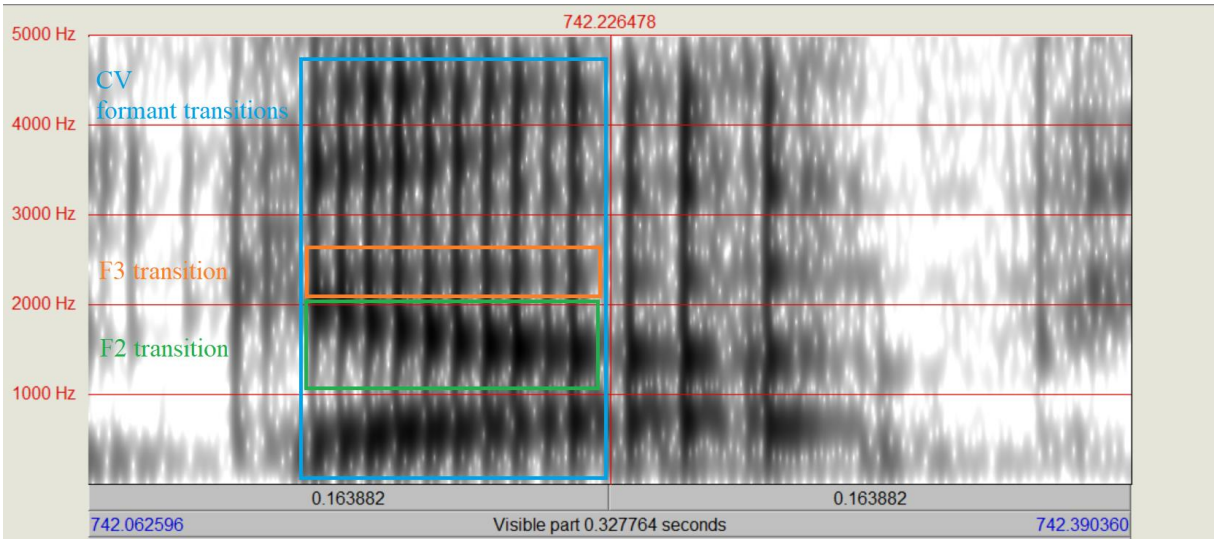


Figure 2.2: Spectrogram of the syllable [gap] from the word *gap*, uttered by a male speaker from Sunderland. The movement of the formants in the first half of the vowel is highlighted in blue ("CV formant transitions"). Note again the movement of the second formant ("F2 transition") and third formant ("F3 transition").

The other major acoustic event for recognizing plosives' place of articulation is the release burst (or 'burst' for short). This can be seen in Figure 2.3:

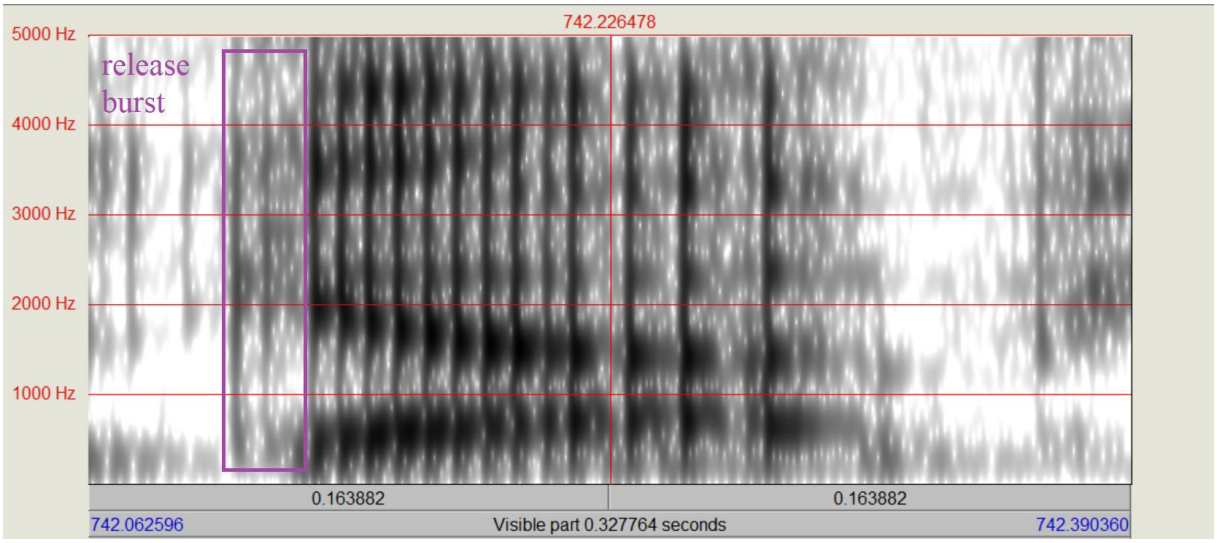


Figure 2.3: Spectrogram of the syllable [gap] from the word *gap*, uttered by a male speaker from Sunderland. The location of the release burst is highlighted in purple.

The burst will be discussed in depth in Section 2.3.

One model for thinking about the acoustics of speech is source-filter theory (Fant, 1960; Stevens, 1998). It posits that the speech signal can be modelled as the product of one or more sources of sound that are subsequently filtered by the shape of the vocal tract (Stevens, 1998: 55). This vocal tract shape is termed the filter and, naturally, its shape differs for different speech sounds, since different speech sounds are articulated at different places of articulation. In Figures 2.1 and 2.2 above, the change in the filter due to the consonant influencing the vowel is reflected in the shifts of the vowel's formant frequencies.

A source can be either quasi-periodic or aperiodic (Clark et al., 2007: 134-140; Harrington and Cassidy, 1999: 30). When the vocal folds are vibrating the result is a quasi-periodic source (often termed a "periodic" source), since there is a pulse of energy emanating from the glottis at close to regular intervals of time. We see this in Figures 2.1 and 2.2 above in the form of dark vertical striations permeating the vowels: these are the periodic glottal pulses produced by voicing. In Figure 2.6 we shall see examples of *aperiodic* sources, i.e. sources that lack this repeating pattern. This periodic-aperiodic duality is important because it defines the difference between the two cues to plosives' place of articulation, the formant-burst distinction.

Once the plosive's closure gesture has progressed sufficiently to stop the intra-oral airflow, the acoustic result is a period of reduced amplitude known as the closure (see Figures 2.4 and 2.5). If the closure is voiceless, as is nearly always the case in /p t k/, then the amplitude is likely to have been reduced fully; if the closure is voiced, as is often the case in English /b d g/, then there can remain some low-amplitude voicing, mostly confined to the first-formant region (Fant, 1973: 24-25).

However in this study we are concerned with the acoustics of place of articulation. Suchato (2004) tested closure duration as an attribute for identifying place of articulation and found it to be the weakest of all the acoustic attributes he examined. Hence closure duration will not be considered further.

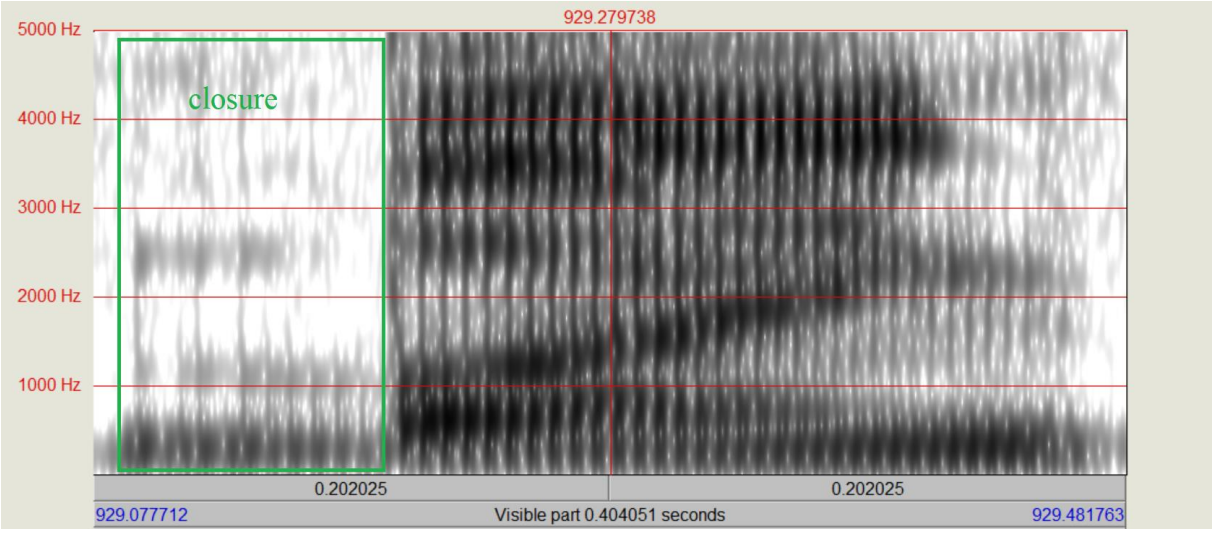


Figure 2.4: Spectrogram of the word *buying* in utterance-initial position, uttered by a male speaker from Sunderland.

Note the near-absence of energy at high frequencies during the closure of the [b].

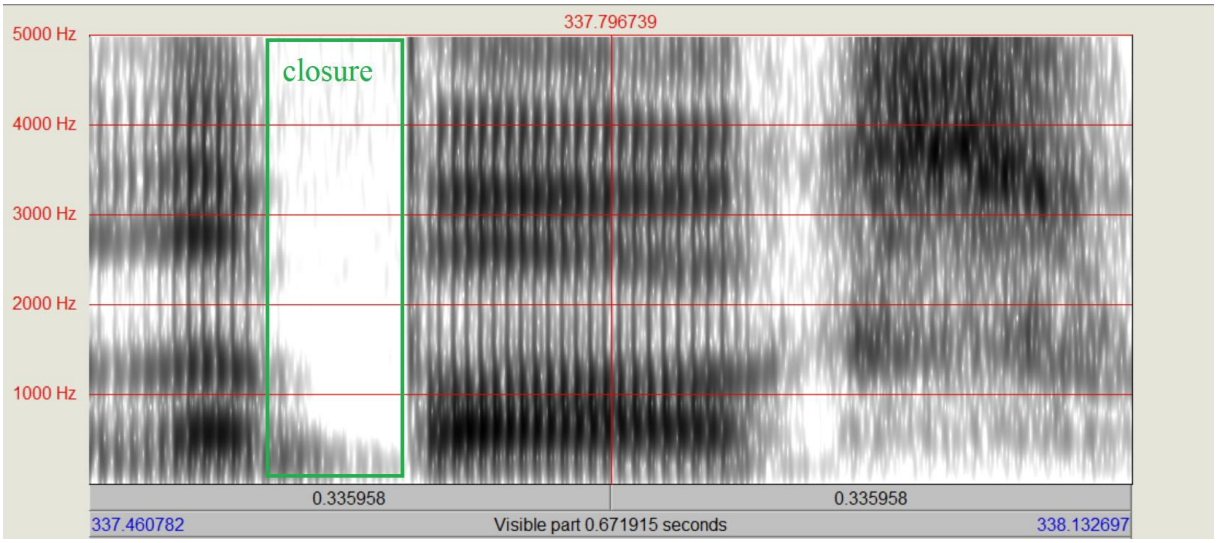


Figure 2.5: Spectrogram of the phrase *the box*, uttered by the same speaker as in Figure 2.4.

Note that only about half of the closure of the [b] is voiced, in contrast to the [b] of *buying* in Figure 2.4. This tendency for the closure to be partly or entirely devoiced is very pervasive in the phonologically 'voiced' plosives /b d g/ of English.

Once the closure gesture is released, the sudden change in air pressure in the mouth generates the release burst, which was introduced briefly in Figure 2.3. The burst can in fact be analysed as consisting of more than one event: a transient and frication (see Figure 2.6). A transient is one kind of aperiodic source (Fant, 1973: 26). It is defined as "the response of the vocal tract to the pressure release, exclusive of any turbulence effects" (Fant, 1973: 111). A transient generally lasts for less than 10 ms (pp. 111-112). Aerodynamically the transient

corresponds to a sudden, step-like (Pickett, 1999: 13) or stepwise (Stevens, 1998: 55) increase in airflow.

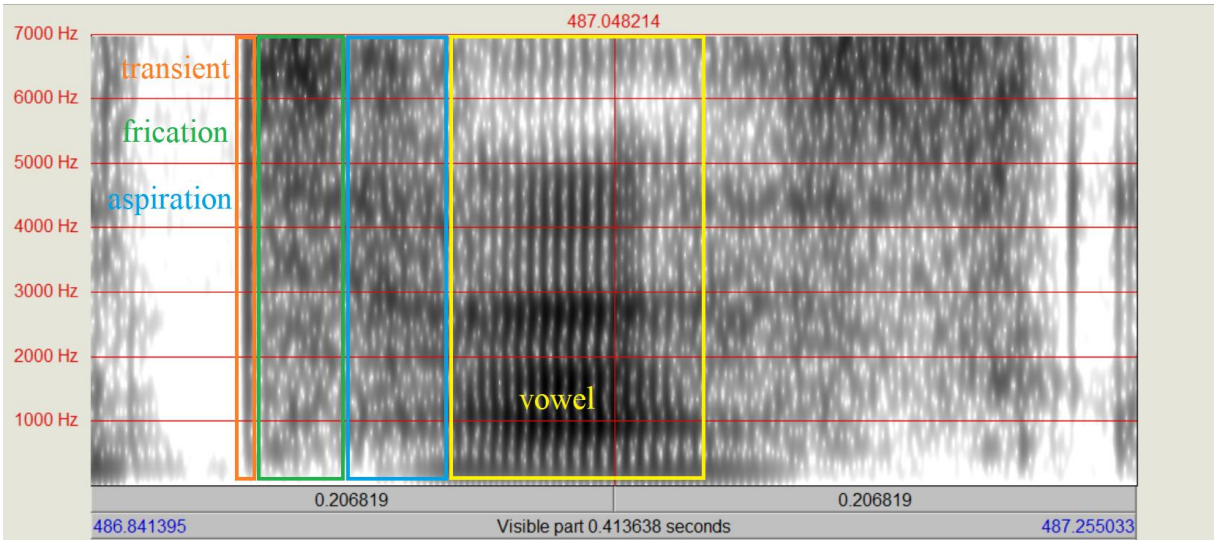


Figure 2.6: Spectrogram of the word *task*, uttered by a female speaker from Tyneside.
Note the position of the transient, frication, and aspiration.

With the closure continuing to open wider, the transient ends. If the closure does not open rapidly, the transient will be followed by a period of frication. Frication is a sound which results from turbulence, i.e. random fluctuations in airflow (Stevens, 1998: 55). Turbulence occurs because the constriction's area is small relative to the air's volume velocity (Clark et al., 2007: 238-239). As in a fricative segment, frication is caused by a noise source or "random source" (Fant, 1973: 112) located at the consonant's place of articulation. In the transient the source of energy was an abrupt step-like increase in air pressure, whereas in the frication the pressure is elevated but fluctuates randomly from moment to moment. Nevertheless, what the two sounds have in common is that they are aperiodic, which means that (in contrast to voicing) they lack a repeating pattern. This can be seen in Figure 2.6 above: the aperiodic transient and frication have much stochastic fluctuation of energy in them, whereas the periodic vowel contains dark vertical striations at regular intervals. Frication is most common with /t d k g/ and relatively uncommon with /p b/, as we shall see.

A note on terminology: because of their similarity in both being aperiodic, the transient and frication are often referred to as a single entity, the "release burst" or "burst" for short (Dorman et al., 1977: 110). Researchers at Haskins Laboratories (Fant, 1973: 113) use the term "burst" to refer to *everything* before the onset of voicing, i.e. not just the transient and frication but also the aspiration. In the present study, we will use "release burst" or "burst" to refer to the transient-plus-frication only.

As the closure opens still wider, there comes a point when the area of the constriction is too large to generate turbulence. And if the vocal folds are not yet in a configuration conducive to voicing, the result will be a period of aspiration. This is an "[h]-like" sound that differs from the transient and frication in having the noise source at the glottis (Fant, 1973: 112). Consequently the aspiration is more likely than the transient and frication to contain some degree of formant structure. This formant structure in the aspiration differs from the formant transitions in vowels in being aperiodic and less intense (Stevens, 1998: 428-441). In English aspiration tends to be considerably longer in /p t k/ than in /b d g/; indeed in the latter series it can be entirely absent. One consequence of this is that in /p t k/ the formant transitions mostly take place in the low-intensity aperiodic aspiration, whereas in /b d g/ they mostly take place in the vowel. Miller and Nicely (1955: 347) argue that this occurrence of the transitions during the aspiration in /p t k/ makes them 'much harder to hear' than in /b d g/.

Nevertheless, the formant transitions of a voiceless plosive can to some degree take place in the vowel if the aspiration is short, which is particularly likely to happen when the plosive is not foot-initial, e.g. the /p t k/ in *happy, city, lucky*.

Having now introduced all the acoustic events possible in a plosive, let us summarize their acoustic similarities and differences:

		source type:	
		periodic	**aperiodic**
source location:	**glottis**	voicing	aspiration/whisper
	place of articulation	-	burst

Table 2.1: The logically possible combinations of source type and source location.
Notice that a periodic source can only be located at the glottis, whereas an aperiodic source can be located at either the glottis or the place of articulation. (Trills do constitute a periodic source at a location other than the glottis but since the present study is concerned with plosives only, I have glossed over this complication.)

To summarize, aperiodic sounds have an unpredictable, moment-to-moment variation in them that can be considered random, whereas periodic sounds do not. The only complications to this dichotomy are: (1) whispery voice and breathy voice, in which there is both a voice source and a noise source at or near the glottis; (2) voiced fricatives, which contain a voice source at the glottis and a noise source at the fricative's place of articulation. A further exception is whispered speech, in which the source at the glottis is a noise source rather than a voice source. Because of these exceptions the dichotomy between periodic and aperiodic sounds discussed below should not be imagined to be a rigid one, but it does help conceptualize the distinction between the burst and the formant transitions in the following discussion.

As we have seen, there are two major regions of the signal where the information for place of articulation is found: (1) after the release of the constriction, which includes the transient and frication, i.e. the burst; and (2) in the formant transitions, i.e. towards the end of the preceding vowel and towards the beginning of the following vowel (and also in the aspiration when this is present). The information in (1) is generated by a source at the place of articulation whereas that in (2) is generated by a source at the glottis.

Let us now explore this dichotomy more deeply.

2.1.4 Supraglottal-Source Acoustics

In Figures 2.1 and 2.2, we noted the presence of formants: frequencies at which the concentration of energy is high relative to surrounding frequencies. Sometimes the acoustic output of speech can be affected by the reverse of this, namely *anti*-formants, which are the mirror image of formants, that is, they are frequencies with particularly low acoustic energy relative to surrounding frequencies. The acoustics of speech can be modelled in the frequency domain by representing the formant frequencies by poles and the anti-formant frequencies by zeros. The poles and zeros of a transfer function are the frequencies for which the value of the denominator and numerator, respectively, of the transfer function are zero (Lyon, 2017: 108).

Sometimes such anti-formants are introduced by a structure that branches off from the vocal tract, e.g. in nasalized vowels the nasal cavity is open, which generates formants and anti-formants of its own in addition to the formants generated by the vocal tract (Fant, 1973: 27; Stevens, 1998: 487-513). For the present study, concerned as it is with plosives, the most noteworthy instance of anti-formants occurs in the frication phase of the burst. Frication (whether it be in a plosive burst or a fricative) is notable for its relative lack of energy at low frequencies. This is because the narrowness of the constriction at the place of articulation separates the acoustics of the spaces in front of and behind the constriction. (The term “back cavity” is used to refer to the space behind the place of articulation, whereas the “front cavity” is used to refer to the space in front of the place of articulation.) In frication, the separation between the front and back cavities results in the formants associated with the back cavity being paired with anti-formants associated with the front cavities. Or to phrase it in conventional acoustic language, “In a frequency region of low coupling between a front cavity and a back cavity […], all the poles of the back cavities […] are bound” (Fant, 1973: 13).

What is the result of this combination of poles and zeros? Fant (1973: 24) writes, “The spectral contribution of a pole and zero of the same complex frequency amounts to nothing, i.e., the pole-zero pair may be removed […] without any effect on the spectrum of the sound

[…]". That is, the front cavity's anti-formants or zeros cancel out the formants or poles of the back cavity, leading to the lack of energy we observe in fricatives at lower frequencies.

Importantly, the front cavity differs depending on the place of articulation. In the velars /k g/, the cavity is relatively long whereas in the alveolars /t d/ it is relatively short. The result of this is that the zeros generated in the alveolar burst range all the way from F1 to F4 (Fant, 1973: 13) whereas in the velar burst all the formants except F1 are clearly visible (Fant, 1973: 28). The bilabial /p b/ bursts, on the other hand, are a special case since there *is* no front cavity beyond the lips: the air, upon release of the closure, rushes out directly to the surrounding atmosphere without having to travel through a filter. The result is that the frication of such consonants has no free poles and no free zeros (Fant, 1973: 14), i.e. zero acoustic output. This explains why such consonants tend not to have frication, containing just a transient on its own. The spectral envelope of the bilabial burst tends to be rather "flat" or "diffuse-falling" (Stevens and Blumstein, 1978).

In summary, the spectrum of the burst differs from that of the aspiration/voicing in containing information about the front cavity only, not the back cavity. When the front cavity becomes longer, the back cavity generally becomes shorter and vice versa (Ladefoged, 1996: 133-134). In this sense the burst does not contain duplicate information about the vocal tract shape in the way that glottal-source speech does. This is an important respect in which the information contained in aspiration or formant transitions differs from that in the burst.

2.1.5 Glottal-Source Acoustics

As mentioned above, glottal-source speech contains information relating to both the front and back cavities. This is because the place of articulation's constriction is generally wider than in noise-source speech, which allows the front and back cavities to be acoustically coupled (Johnson, 2012: 176-178).

Figure 2.7 shows that the formant frequencies associated with alveolar place of articulation are a fairly high F2 frequency (namely ca. 1,850 Hz in a male voice) and a high F3 (in excess of 3,000 Hz). This means that at the onset and offset of vowels on either side of such a consonant, we expect the formant frequencies to approach these values (Ladefoged, 1996: 130).

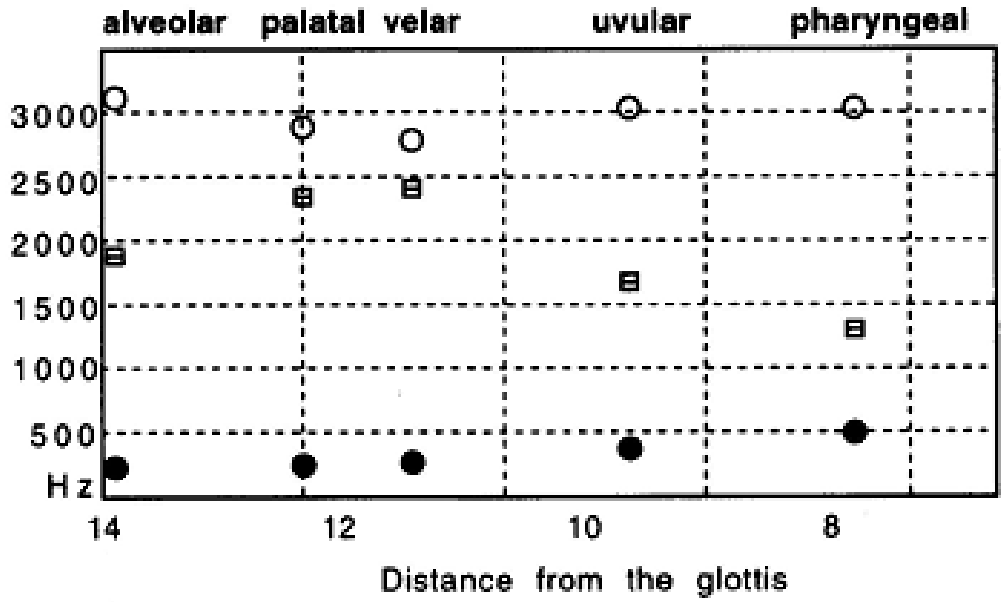


Figure 2.7: The frequencies of F1, F2, and F3 when the constriction in the vocal tract is at various distances from the glottis (in cm).
F1 is represented by black dots, F2 by squares with a line through the middle, and F3 by white dots. Based on a coupled-tube model of the vocal tract developed inter alia by Fant (1960). From Ladefoged (1996: 130).

For velar place of articulation, F2 and F3 are much closer to each other: compared to alveolar, F2 is higher and F3 lower. This proximity of F2 and F3 is termed “velar pinch” (Ladefoged and Johnson, 2011: 311). Figures 2.1 and 2.2 illustrated this: the part of the vowel closest to the velar’s closure showed a smaller distance between F2 and F3 than the middle part of the vowel.

A question arises regarding these velar and alveolar formant frequencies: are they generated by the front cavity or the back cavity? It turns out that some are generated by the front cavity and some by the back cavity, as we can see from Figure 2.8 (Ladefoged, 1996: 133).

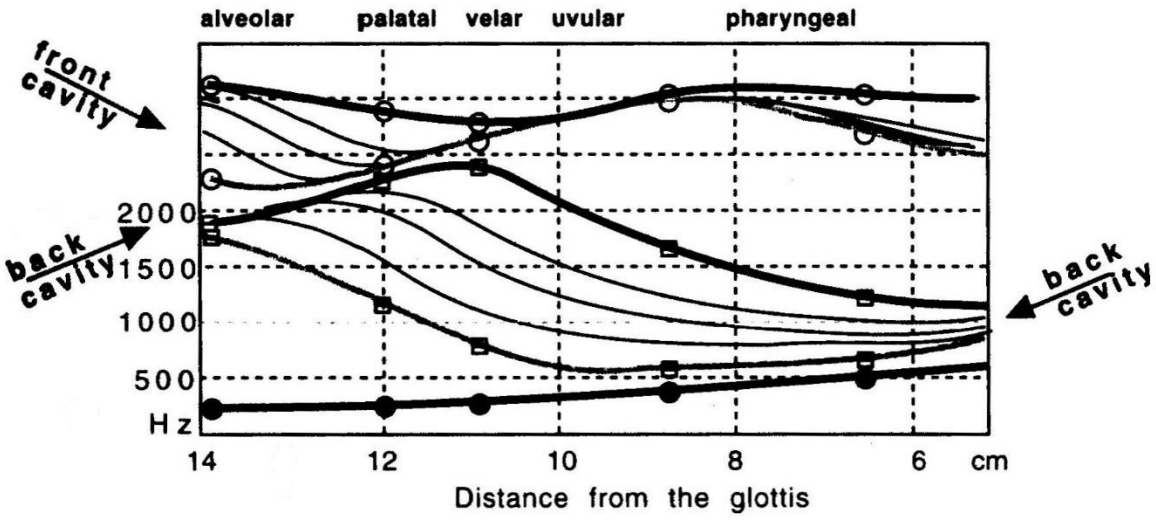


Figure 2.8: The frequencies of F1, F2, and F3 produced as a result of varying the location of a constriction in the vocal tract and also varying the degree of lip rounding.
For each formant, the most unrounded position of the lips is represented by the heavy line, the most rounded position by the thin line that is furthest away from the heavy line. As per Figure 2.7, black dots indicate F1, white dots F3, and squares F2. The important thing to note is the arrows on the left and right, which indicate the cavity that F2 and F3 are affiliated with for a given constriction location. From Ladefoged (1996: 133).

For alveolar place, we see that F2 is the result the back cavity whereas F3 is the result of the front cavity. Holding all else equal, short cavities have higher resonant frequencies than longer cavities (Schnupp et al., 2011: 7-8). Thus the short front cavity of alveolars is associated with the higher-frequency formant (F3) whereas the long back cavity is associated with the lower-frequency formant (F2). For velars, the cavities are close to equal in length, which is reflected in the two formants being relatively close together (Ladefoged, 1996: 133). Note that at some point in the back part of the velar articulatory region F2 and F3 switch their cavity affiliation, that is, F2 comes to be associated with the front cavity rather than the back cavity and F3 becomes a higher resonance of the back cavity (ibid.).

Velar pinch does not occur in all contexts. This is because the precise place of articulation of a velar consonant varies depending on the position the tongue must assume for the adjoining vowels, especially the following vowel (Wada et al., 1970; Ohala, 1971). When the place of articulation is further back, as it is before the back vowels /o u/ (Ohala, 1971), F2 tends to be quite distant from F3, as we see in Figure 2.9:

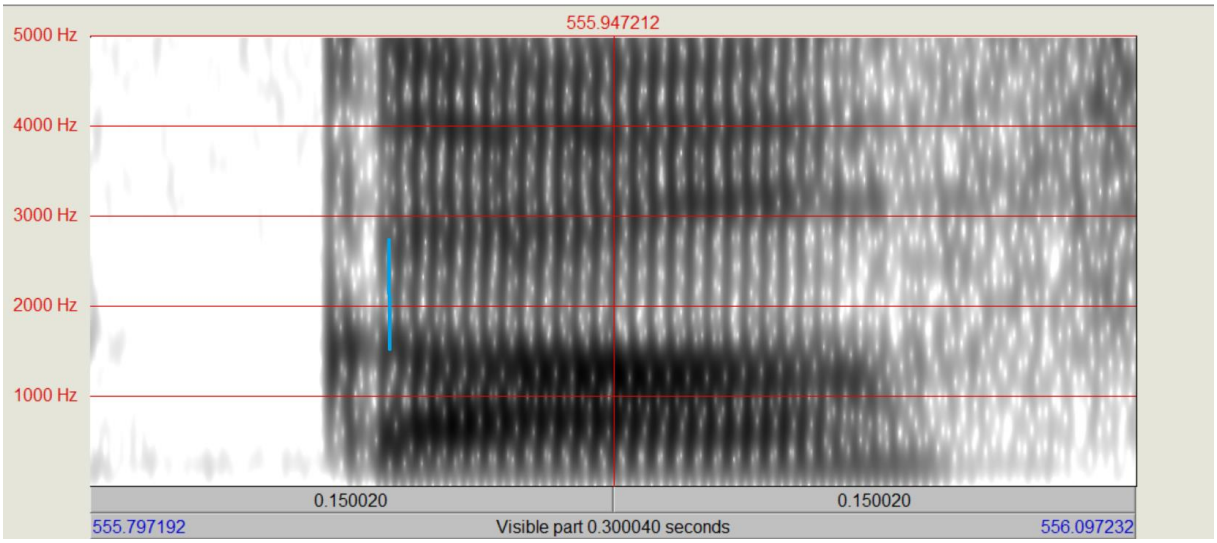


Figure 2.9: Spectrogram of the word *golf* [gɔlf] as uttered by a female speaker from Tyneside.
This example illustrates the lack of velar pinch typical of velar consonants before back vowels, due to the velar place of articulation being further back than the middle of the vocal tract. The vertical blue bar indicates the substantial difference in frequency between F2 and F3 at vowel onset, namely 1,277 Hz (F2 = 1,477 Hz, F3 = 2,754 Hz).

It can be seen that F2 begins relatively low in frequency (1,477 Hz). In back velars such as the one above, of course, F2 is affiliated with the front cavity, given that the front cavity is relatively long due to the place of articulation being further back than it is before front vowels. This is approximately situated in the zone on Figure 2.8 labelled as 'uvular'. In terms of the location of their passive articulator, the velar consonants /k g/ actually have more than one place of articulation.

As for bilabial consonants, Fant (1973: 28) states that a reduced interlabial area results in a lowering of all the formants. Thus when the lip gesture is released, we expect to see all the formant frequencies rise as the area between the lips continues to increase. However, Stevens (1998: 341-343) notes that the precise magnitude of the F2 rise depends on the following vowel, being small for back vowels, moderate for open-mid front vowels and greatest for close front vowels. However, as he also points out, the F2 rise before close front vowels (i.e. [bi]) will tend to be smaller than what he modelled because the first 10 ms of the earliest part of the transition (the most rapidly rising part) will be obscured by the release burst (p. 342). The reason the bilabial formant transitions tend to be rapid in their earliest stage is that: (1) the jaw movement involved in opening the lips is rapid relative to other articulations (Stevens models the rate of increase of the bilabial opening as 100 cm^2/s (p.342), compared to 50 cm^2/s for the tongue tip (p. 355) and 25 cm^2/s for the tongue body (p. 365)); and (2) the tongue body can be largely in place for the vowel at the time of consonant release (p. 342), since bilabials require the lips rather than the tongue, unlike velars and alveolars.

To summarize: before front vowels the F2 frequency at the beginning of the vowel (henceforth $F2_{onset}$) is highest in frequency for velars, lowest for bilabials, with alveolars intermediate; before back vowels $F2_{onset}$ is highest in alveolars, bilabials again lowest, with velars intermediate. Nevertheless, the exact details of the formant transitions vary depending on the particular vowel, and in practice there can be considerable overlap in the formant pattern for two places of articulation in a given vowel context (examples given in 2.2.3).

2.2 Formant Information

In the previous section we overviewed the formant transitions' acoustics. In this section we examine them in greater detail, with the following question in mind: how can the information about plosive place of articulation in the formant transitions be represented? How many features are involved, and to what extent can they be merged to form a smaller number of features?

2.2.1 The Locus Theory

Around 1950 a new instrument for analysing speech was developed, the pattern playback (Cooper et al., 1951). The researcher painted spectrograms by hand and the pattern playback played them aloud. Thus was born synthetic speech. The great benefit of this research was that for the first time the components of speech could be studied individually. For example, if a researcher suspected that the frequency of the second formant at the onset of a vowel (henceforth $F2_{onset}$) played a role in identifying a consonant's place of articulation, they could paint spectrograms in which $F2_{onset}$ varied in equal steps in a set of stimuli, then play them to listeners and ask them which consonant they perceived.

This is precisely how Liberman and his colleagues began their exploration of formants and place of articulation (Liberman et al., 1954). In that study painted two-formant spectrograms of transitions plus vowel were played to listeners. The F2 transition for each vowel varied in increments of 120 Hz from an onset value that was 480 Hz below the vowel midpoint (called 'minus' transitions) to a value 720 Hz above the vowel (called 'plus' transitions):

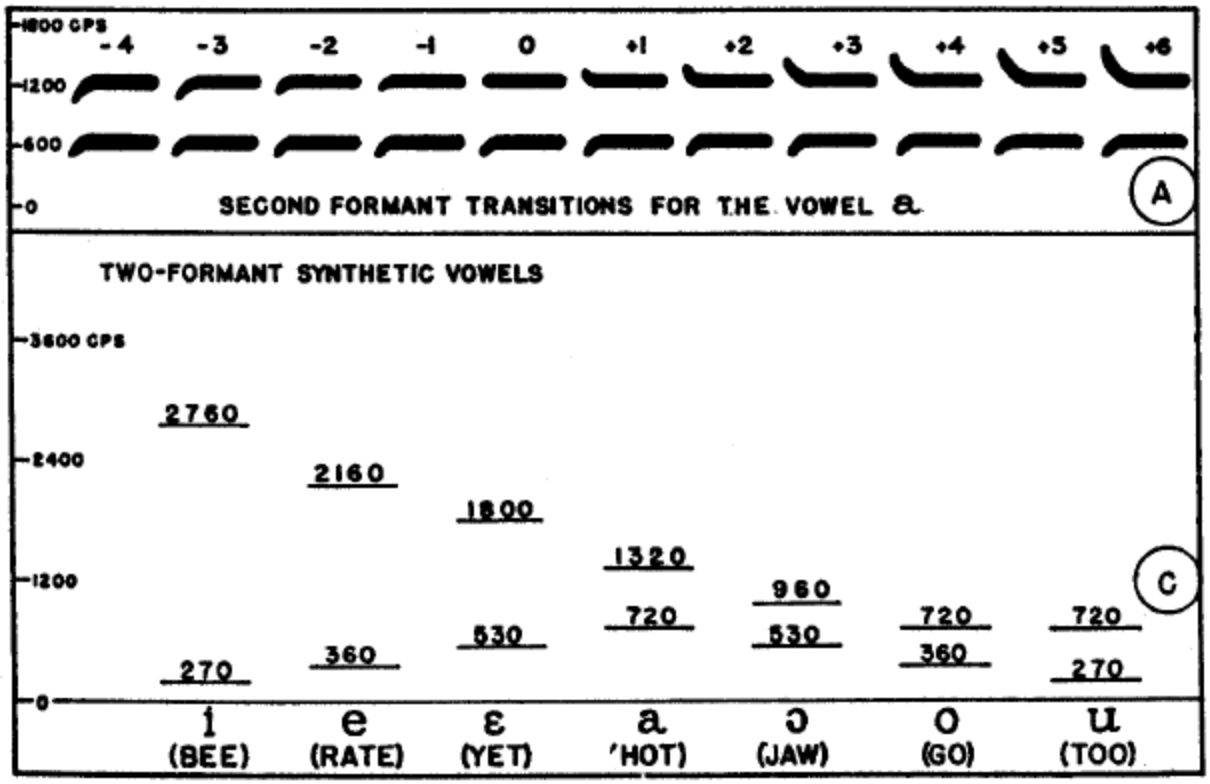


Figure 2.10: Stimuli used by Liberman et al. (1954: 3).
Panel A shows the types of F2 transitions used. Panel C shows the seven vowels with which these transitions were placed.

The minus-transitions were generally perceived as bilabial (/p b/), whereas the plus-transitions were perceived as either alveolar (/t d/) or velar (/k g/), depending on the transition's size and the particular vowel: before *front* vowels the /k g/ percept was elicited by the stimuli with the most sharply-falling F2 transition (i.e. velars had the highest $F2_{onset}$ frequency); whereas before *back* vowels the best velar responses were for transitions that fell only slightly. One interesting observation regarding alveolars was that the best transitions before /e ɛ/ were flat whereas in the series /a ɔ o u/ a progressively larger downward transition was needed, i.e. as the F2 of the vowel became lower, the F2 transitions became increasingly steep in their downward slope.

This mass of curious findings cried out for organization, and it was to this end that subsequent research was aimed. Delattre et al. (1955) again created artificial stimuli for listeners to identify, but this time they began with the assumption that the F2 transition for a particular place of articulation *points* at a frequency, a frequency that is specific to that place of articulation. Here is their line of reasoning (Delattre et al., 1955: 769):

"[I]f we [...] assume that the relation between articulation and sound is not too complex, we should suppose [...] that the second-formant transitions rather directly reflect the articulatory movements *from*

the place of production of the consonant *to* the position for the following vowel. Since the articulatory place of production of each consonant is, for the most part, fixed, we might expect to find that there is correspondingly a fixed frequency position - or "locus" - for its second formant [...]"

The authors posited that this locus frequency was found around 50 ms prior to the observed beginning of the formant transition. Here is an illustration for the alveolar consonant [d]:

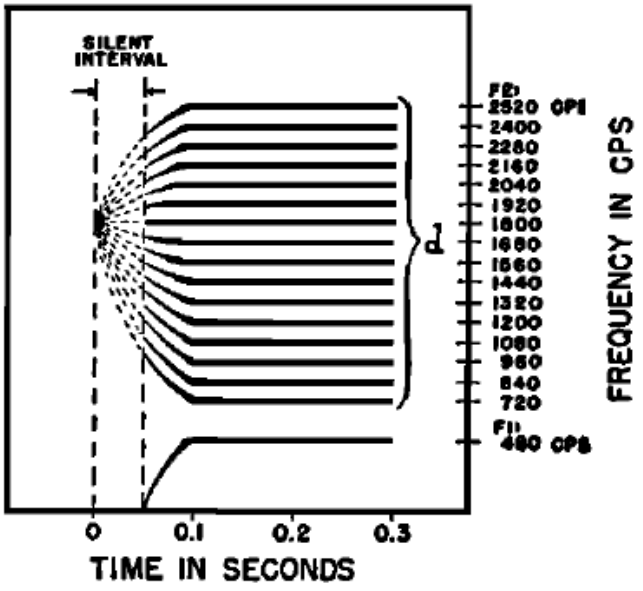


Figure 2.11: Schematic illustration of the locus theory for a /d/ that is paired with a range of vowels.
The vowels vary in backness (i.e. they vary in their F2 frequency but not their F1 frequency). The F2 transitions for all the vowels begin at the same frequency of 1,800 cps (1,800 Hz), at least if one traces their trajectory to an unobserved point in time around 50 ms prior to the vowel onset. From Delattre et al. (1955: 771).

The various vowel F2 trajectories can be seen to point to the same frequency, 1,800 Hz, even though they only reach this frequency if one imagines their trajectory extending backwards in time to around 50 ms prior to their actual beginning.

Note that the only transition in Figure 2.11 in which $F2_{onset}$ and $F2_{locus}$ are the same frequency is the 'flat' transition located at 1,800 Hz. Flat F2 transitions are easy to work with in that the frequency of $F2_{locus}$ can be known with confidence: it is the same frequency as $F2_{onset}$.

Delattre et al. thus used flat F2 transitions in their efforts to locate the locus frequency for each place of articulation. Here is an illustration of the stimuli used to locate the locus:

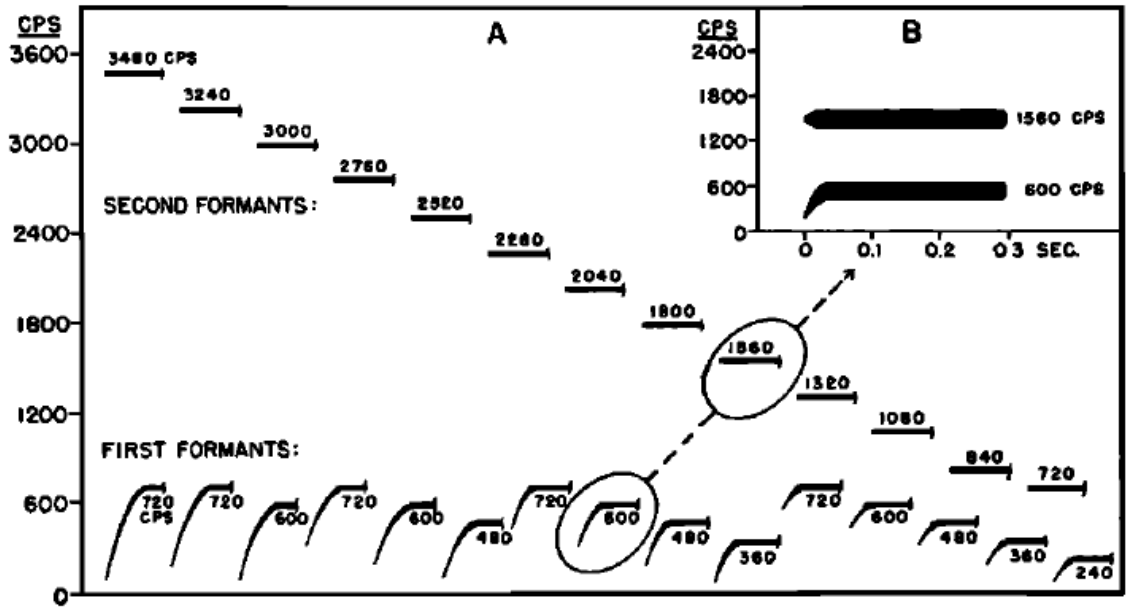


Figure 2.12: Stimuli used by Delattre et al. (1955: 771) to investigate the locus theory.
Panel A shows the range of F2 vowel frequencies involved; Panel B is an example of one of the F1-F2 pairings.

It can be seen that their technique was to place a *flat* F2 transition of varying frequency with different kinds of F1 transition. This resulted in 65 artificial stimuli. The authors of the paper listened to the stimuli and judged whether they sounded like /b/, /d/, or /g/. The fact that the F2 transition was flat meant that by definition the formant transition pointed at the same frequency as itself, which made the task of locating each place of articulation's locus easier. The best /d/ was obtained when the flat transition pointed to 1,800 Hz; for /b/, 720 Hz; and for /g/ (before front vowels) around 3,000 Hz.

At first blush the locus idea looked promising. There were, however, a number of difficulties. Although there was clearly a locus for /d/, the locus for /b/ was less clear-cut: its transition did point downwards in most contexts, but it was difficult to be sure what the precise frequency was to which it pointed, or indeed if all of its transitions pointed to precisely the same low frequency. Thus its 720 Hz 'locus' has to be treated cautiously, as the authors themselves acknowledged. Secondly, the locus for /g/ sounded even less satisfactory to listeners, and only worked for front vowels; before back vowels, no velar locus could be found. However, this could simply be an artefact of the authors' method of identifying the locus, namely to find the vowel frequency at which a *flat* transition creates a velar percept. If the F2 transition of a velar plus back vowel is not flat, then none of the flat-transition stimuli will yield a percept of velar place.

Indeed, the authors' own figure illustrating the best percepts shows that /g/ does not have a flat F2 transition in any vocalic context:

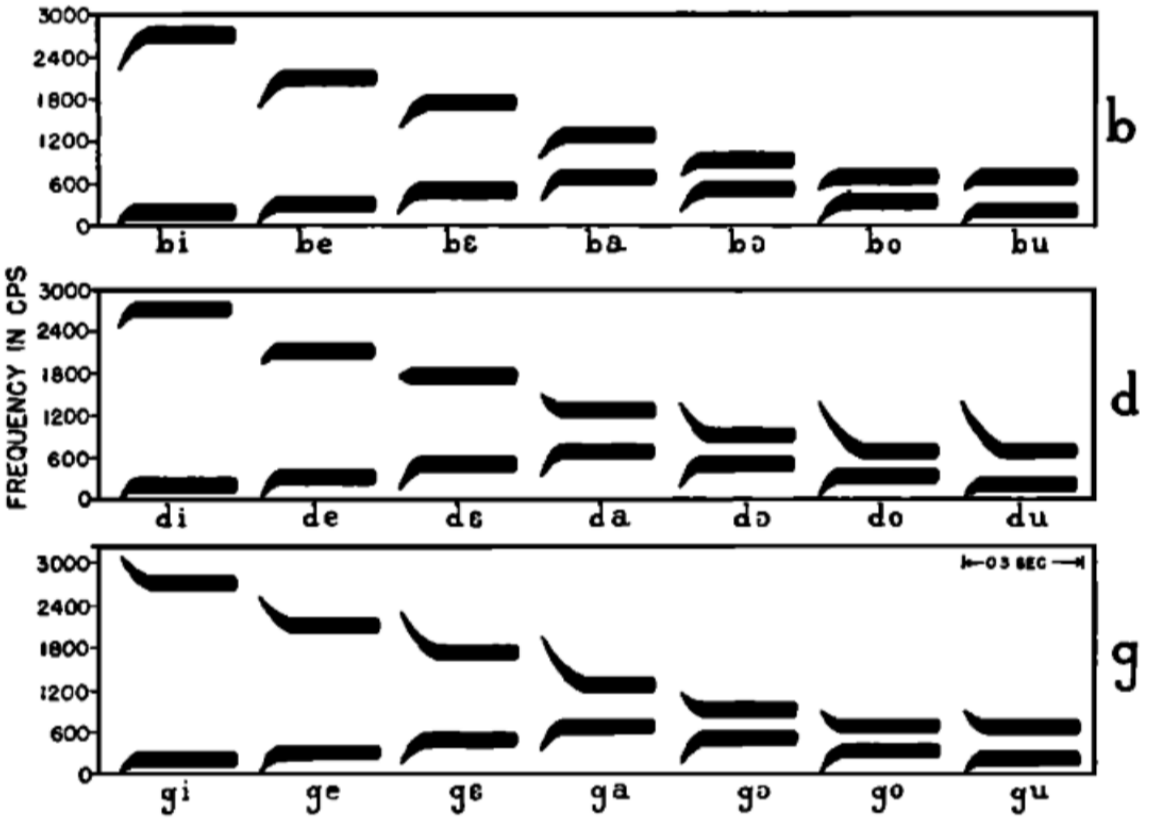


Figure 2.13: Two-formant artificial stimuli showing the best formant transitions for each voiced stop and following vowel.
From Delattre et al. (1955: 770).

There were a number of other studies of the topic during these years, and all found broadly the same picture as Liberman et al. Lehiste and Peterson (1961) measured the formant frequencies occurring in natural speech and failed to find invariant F2 loci for each place. They came closest to succeeding with alveolars (as we have seen, the transitions of this place seem to be the easiest to simulate with the locus concept). In addition, they failed to find the two front-vowel versus back-vowel locus frequencies for velars; instead they found multiple potential loci that varied according to the F2 frequency of the vowel itself.

Despite these difficulties, the locus theory is significant as we can derive from it the following key prediction: if we could somehow replace $F2_{onset}$ with $F2_{locus}$, the accuracy of $F2_{locus}$ at recognizing place of articulation would be greater than $F2_{onset}$. Chapter 5 tests this.

2.2.2 Influence of a Preceding Vowel on Formant Transitions

The stimuli used in the 1950s research consisted of single syllables. Öhman (1966) extended the study of formant transitions to vowel-consonant-vowel (henceforth VCV) context. His data (from Swedish) consisted of all possible VCV combinations of /b d g/ and /y ø ɑ o u/. He hand-traced the first three formants of such sequences and produced diagrams of each, which were

averaged based on three repetitions from a single speaker (himself). This yielded 25 contexts for each place of articulation.

The CV F2 transitions which Öhman found in VCV sequences turned out to be different from the CV transitions found in the 1950s research. Let us take the example of /b/. As we have seen, it was already known that the precise frequency to which this consonant's F2 transition pointed was difficult to locate precisely (indeed some studies such as Lehiste and Peterson (1961) had found more than one locus). However, the possibility of an F2 locus frequency for /b/ still seemed plausible since the F2 transition for this place of articulation pointed downwards for most vowels in CV sequences. But Öhman's data showed that in VCV sequences this was not always true: his diagrams for /ybɑ ybo ybu øbo øbu ɑbu/ show the CV transition pointing *upwards* (and likewise for the VC transitions in /ɑby oby uby ɑbø obø ubø obɑ/ (p. 160)). This means, of course, that the CV transition for a given vowel does not point to a single frequency, but several. For example, in /obo/ (see Figure 2.14) the /bo/ transition appears to point to a frequency of around 500 Hz whereas in /ybo/ it points to a frequency well above 1,000 Hz, perhaps 1,250 Hz. So even when the place of articulation and the second vowel were held constant, the effect of the preceding vowel was to move the locus around considerably.

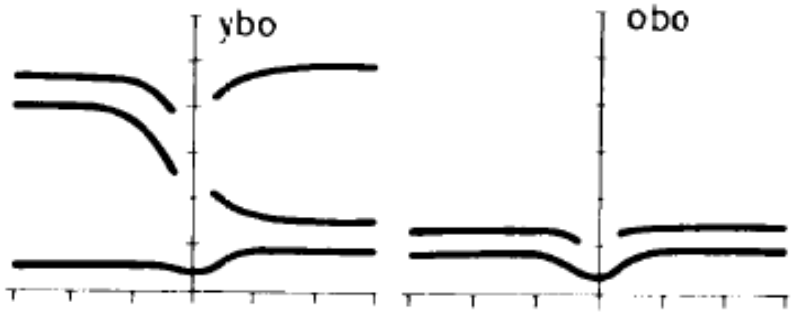


Figure 2.14: Schematic diagrams of /ybo/ and /obo/ based on averaging three repetitions of each from a single speaker.

The notches on the horizontal axis are spaced 100 ms apart; the notches on the vertical axis are spaced every 500 Hz apart. The influence of the first vowel on the onset of the second vowel is apparent: in /obo/ $F2_{onset}$ is ca. 600 Hz whereas in /ybo/ it is ca. 1,000 Hz. From Öhman (1966: 160).

The effect of Öhman's findings was dramatic: it put the nail in the coffin for the debate over an invariant locus frequency and convinced most researchers to abandon the locus theory (Lindblom and Sussman, 2012: 4).

It is worth reflecting on to what extent this move away from the locus theory was warranted. One crucial factor is whether the locus for a given place of articulation is defined as being a frequency *point* or a frequency *zone*. If one considers the locus as being a frequency *point*, then Öhman's finding certainly do seem to rule that possibility out: an $F2_{onset}$ frequency of 600 Hz in /obo/ is quite a different frequency from the $F2_{onset}$ in /ybo/, 1,000 Hz, and the

frequencies to which the transitions *point* are even further apart: roughly 500 Hz and 1,250 Hz respectively. If, however, one construes the locus as being a frequency *zone*, then the variation caused by the preceding vowel, though considerable, is less dramatic: after all, 600 Hz and 1,000 Hz are both low F2 frequencies.

This point is underscored when one compares the F2 frequencies across different places of articulation. For example, the $F2_{onset}$ frequencies in /ydo/ and /odo/ are 1,700 Hz and 1,300 Hz respectively (Öhman, 1966: 161). So even the *lowest* /Vdo/ $F2_{onset}$ frequency, 1,300 Hz, is still *higher* than the highest /Vbo/ onset frequency, 1,000 Hz. This suggests that the coarticulatory effect induced on $F2_{onset}$ by the first vowel, while noticeable, is still modest compared to the effect induced by the plosive's place of articulation (and, of course, by the following vowel).

Given the considerable influence of Öhman's study (Lindblom and Sussman, 2012), it is worth noting five of its features: (1) it consisted of the speech of just a single speaker; (2) that speaker was the author; (3) the material consisted of nonce VCV sequences; (4) the speaker said the syllables in a monotone with equal stress on both syllables; (5) the data consisted of three repetitions of 5 × 5 contexts for three different consonants, i.e. N = 225. It is something of an open question to what extent such material is representative of real-life speech. It seems that, given the paucity of subjects on which Öhman's findings were based and the artificial nature of the utterances, the findings should be treated with caution. It would be interesting to see if material involving a larger number of informants, more natural material, and automatic formant extraction would succeed at replicating Öhman's main finding of $F2_{onset}$ (and hence $F2_{locus}$) being heavily influenced by $F2_{mid}V1$. To this end, in Chapter 5 intervocalic /b d g/ in natural speech in English will be investigated using the present study's dataset with the aim of quantifying the effect of $F2_{mid}V1$ on $F2_{onset}V2$ relative to $F2_{mid}V2$. This will help us to gauge whether the influence of $F2_{mid}V1$ on $F2_{onset}V2$ is as large in natural speech as Öhman found in his nonce sequences.

2.2.3 Locus Equations

The 1950s research on CV sequences identified the importance of the F2 frequencies at the onset of the vowel ($F2_{onset}$) and at the steady part of the vowel ($F2_{mid}$) for recognizing place of articulation.

One way that this observation has been formalized is with the locus equation (not to be confused with the locus *theory* discussed above). In a locus equation $F2_{onset}$ is represented by the y-axis and $F2_{mid}$ by the x-axis. For each place of articulation, the consonant is paired with a variety of following vowels, and the data are fitted using linear regression. This yields a slope

and intercept for each place of articulation. For example, one might plot /di de da do du/ and repeat for /bi be ba bo bu/ and /gi ge ga go gu/. In articulatory terms, the slope of the line for each place of articulation is thought to represent the extent to which the following vowel and consonant overlap (or "coarticulate"). For example, if the line is relatively flat, this indicates that the vowel allows the consonant to coarticulate with itself relatively little. This method quantifies the acoustic coarticulation of a place of articulation with a following vowel. Given that different places of articulation coarticulate with the vowel differently, this means that locus equations are a potential correlate of place of articulation.

Two noteworthy findings have emerged from this research. The first is that the onset and midpoint F2 frequencies for a given place of articulation in different vowel contexts fit the line of best fit strikingly well. In other words, as the frequency of $F2_{mid}$ changes, the frequency of $F2_{onset}$ changes at a rate proportional to $F2_{mid}$. This can be seen in the following example, from Sussman et al. (1991: 1313):

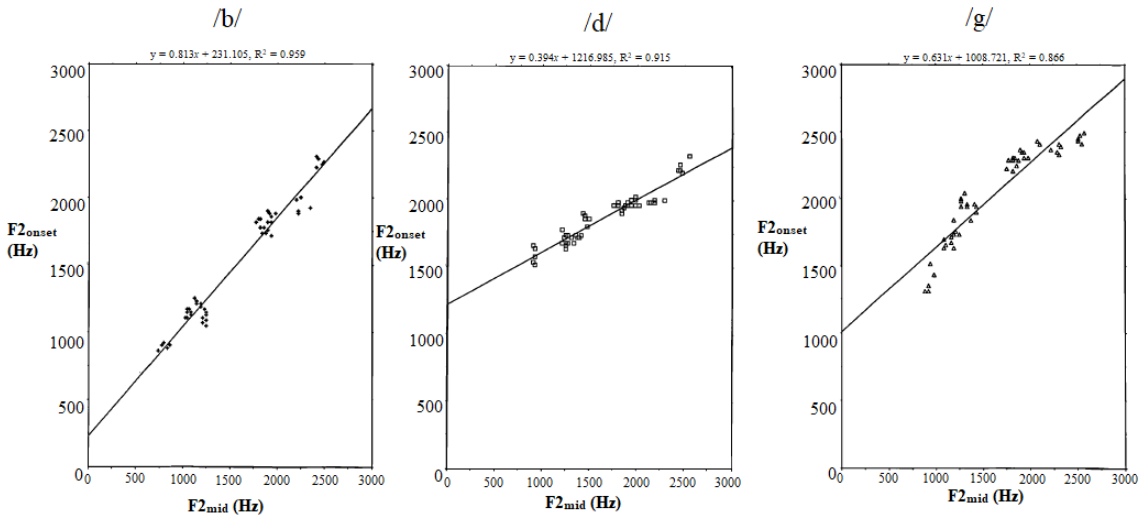


Figure 2.15: F2 locus equations for each of /b d g/, as spoken by a male American English speaker. From Sussman et al. (1991: 1313).

This finding of linearity is significant because when one looks at the F2 transitions of consonant + vowel sequences individually, it is not obvious that they have something in common. In contrast, when they are plotted together in the linear regression, one sees that they have a shared linearity. This is significant because it is a putative sign of invariance: what superficially look like very variable F2 transitions for a given place of articulation do in fact hold something in common: the degree of coarticulation between vowel and consonant.

In Figure 2.15 the slope of the line for /d/ is flatter than the line for /b/. This indicates that $F2_{onset}$ and $F2_{mid}$ tend to be more dissimilar from each other in /d/ than in /b/. In articulatory

terms, this is related to the fact that having the tongue body in a relatively front position facilitates the tongue tip touching the alveolar ridge (Stevens, 1998: 355). In /b/, by contrast, the articulator is the lips, which – unlike the tongue tip – are not joined to the tongue body. Thus the tongue body gesture for the vowel is not delayed by a tongue body requirement for the consonant, unlike in the case of /d/. The steepness of the locus equation slope for /b/ relative to that for /d/ has been repeatedly found by numerous locus equation studies (see the review in Fowler, 1994).

For /g/ the picture is more complicated in that it is typically the case (Fowler, 1994) that its locus equation line fits the data imperfectly; observe this in Figure 2.15. Notice how the data points on the left (representing back-vowel context) are arrayed more steeply than the data points on the right (representing front-vowel context). As a result its front-vowel and back-vowel contexts are often plotted separately (e.g. Sussman et al., 1991: 1316; Sussman et al., 1998: 247). When this is done it is found that the line for front-vowel contexts is flatter than the line for back-vowel contexts. Recall that as long ago as Delattre et al. (1955) it was found that to obtain the best velar percept, the F2 transition had to fall sharply before front vowels but only shallowly before back vowels (as shown in Figure 2.13 above).

In any event, the fact that each place of articulation has its own characteristic coarticulatory pattern might lead one to consider locus equations as a correlate for distinguishing place of articulation. Sussman et al. (1991) recorded ten male and ten female speakers producing five repetitions of /bVt dVt gVt/ in a carrier phrase, with /V/ being one of ten American English vowels. A discriminant analysis in which $F2_{onset}$ and $F2_{mid}$ were the predictors yielded 82% correct classification of labials, 78% of alveolars, and 67% of velars, or 77% overall. Although these classification rates are well above chance, we should remember that the consonants were uttered in a carrier phrase, i.e. the preceding vowel did not vary. In other words, the data were quite heavily controlled, and yet the classification rates were well below 100%. In this sense, then, $F2_{onset}$ and $F2_{mid}$ seem to be insufficient cues on their own. However, when Sussman et al. instead used locus equation slope and intercept as predictors, their place of articulation accuracy jumped to 100%.

However, as Fitch and Hauser (1998: 265) remark, “because the method to create a locus plot involves already-classified data, 100% classification of locus data is unimpressive”. Furthermore this method required having a separate slope and intercept for each of the 20 speakers, since the within-speaker variation would have been too large to yield such high classification accuracies if all-speaker slopes and intercepts had been used (Sussman et al. note (p. 1315) that some speakers’ /g/ slopes overlap with other speakers’ /b/ slopes). Thus high classification accuracy from locus equations appears to be only possible if the slopes and

intercepts are allowed to be different for each individual speaker. In Section 5.4.2 the role of normalizing formant frequencies by individual speaker in improving the classification accuracy of place of articulation will be investigated.

Hasegawa-Johnson (1996: 25) notes that most of the studies of F2 transitions for voiced stops have yielded classification accuracies between 65% and 70%. Thus even the 77% classification accuracy that Sussman et al. achieved is higher than most studies, which is likely due to the fact that Sussman et al.'s data were highly controlled in nature, viz. /CVt/ words read five times in a carrier phrase.

Despite all of this, as Lindblom and Sussman (2012) note, the observation that locus equations yielded high classification accuracies (at least under the special condition of constructing separate locus equations for each speaker separately) changed the emphasis from them being viewed primarily as a means of quantifying a place of articulation's coarticulation (Lindblom, 1963; Krull, 1987) to being regarded as potential cues for place of articulation (Nearey and Shammass, 1987; Sussman, 1989; Sussman et al., 1991; Sussman et al., 1995; Fruchter and Sussman, 1997). This interest in locus equations culminated in the "Orderly Output Constraint" (Sussman et al., 1998), a putative rule that speakers must follow in speech production to make their speech intelligible to the listener. On this view, the straightness of locus equation lines is due to speakers deliberately following a *rule* for coarticulation rather than, say, the straightness of locus equations merely being some by-product of the motor system's inherent design (Lindblom and Sussman, 2012). As Sussman et al. put it, "the articulatory system, across diverse articulators (tongue, lips, jaw, velum), adjusts consonant-vowel coarticulation with respect to the acoustic output in order to fine-tune a feature of that output, the F2 onset/F2 vowel ratio" (Sussman et al., 1998: 255).

As Brancazio (1998: 261) points out, this picture portrays the linearity of locus equations as perceptually (rather than articulatorily) driven. Fowler (1998: 265) sees the linearity of locus equations as being more plausibly explained by the fact that "vowels all use the tongue body, so their interference with a given consonant should be approximately the same". In more recent years Sussman (Lindblom and Sussman, 2012: 3) has moved away from as strong a formulation as the OOP:

"This descriptive characterization of [locus equations] is not meant to imply a specific, perceptually relevant, invariant cue for perception of stop place. Instead, LEs are simply viewed as providing conceptual insights into the acoustic organization, and hence possible neural coding, of stop-vowel sound categories."

The OOP was also criticized by Fitch and Hauser (1998: 265) on the grounds that: "To construct a locus plot for a given consonant, the listener must have already classified a number of syllables correctly, which requires the identification problem to be solved already." Chennoukh et al. (1997: 2380) make a similar point: "[...] being a statistical representation of data obtained from a set of utterances with the same consonant in varying vowel environments [...], the equation does not allow a listener to deduce place of articulation from a single token". This is not necessarily a problem if we assume that the speaker has a stored representation of possible locus-equation spaces from previous speakers and compares the new speaker's vocal characteristics, matching them to the most similar voice's locus equation in memory. On the other hand, it does raise the question of how locus equations are acquired by the learner in the first place; presumably the fact that most $F2_{onset}$-$F2_{mid}$ tokens are in reality paired with a release burst aids the learner in doing so. The release burst, as we shall see in Section 2.3 (and 6.2), seems to show more vowel-independent invariance than the formant transitions (especially for alveolar consonants), and Sussman et al. do acknowledge that the burst is another "crucial" cue (p. 246) to consonant place. If the release burst is also crucial, this makes it difficult to interpret their claim (p. 246) that the F2 transition is "perhaps the single most important" cue in speech perception.

Fowler (1994) has argued that locus equations are best seen as a specifier of place, not a full-blown cue. Firstly, she notes that the precise steepness of slopes depends not just on their place of articulation but also their manner. Thus the slope for [s] is shallower than the slope for [d]. This means that locus equations could not be used for perceiving place without the manner of articulation being known first (which is not an unreasonable stipulation). A corollary of these findings about slopes is that consonants with very different features (different place, different manner) may have similar locus slopes. She found this to be the case with [g] and [v], though the intercept of such consonants are statistically significantly different (Fowler, 1998: 604), which means that, even if they do have very similar slopes, they are not located in quite the same region of the $F2_{onset}$-$F2_{mid}$ space.

A second challenge with locus equations is that the slope for a particular consonant tends to be different in the VC transition than the CV transition. For example, [d] has been found to have a steeper locus slope in VC than CV (Al-Tamimi, 2004). Translating this into the language of articulation, this means that the VC transition of [d] is more affected by the adjoining vowel than the CV transition. Öhman's (1966: 161) diagrams of /ɑdɑ odo/ also illustrate this. From a perceptual point of view, it means that one could not use the same locus equation slope for identifying /d/ from a VC transition as a CV transition. As we shall see later (Malécot, 1958), listeners can indeed have difficulty identifying stimuli using VC transitions

alone, and when the VC transition is paired with a burst of a different place of articulation, listeners perceive place of articulation based on the burst, not the VC transitions. Nevertheless, listeners do seem to be capable of identifying the place of burstless VC stimuli above chance. In any event, the point is that the locus equation slopes and intercepts of VC transitions would need to be different from those of CV transitions. Though this is feasible it means locus equations are less invariant for a single place of articulation than a consideration of CV stimuli only would indicate.

A further challenge with using locus equation slopes and intercepts for identifying place of articulation (rather than for quantifying coarticulation) is that locus equation slopes can vary depending on the speaking style. Krull (1987) found that one of her participants had a slope of 0.47 for /d/ in spontaneous speech but 0.43 in read words (a small difference) but the other participant had a slope of 0.45 in spontaneous speech but 0.25 in read speech. Other studies that have found a change in slope between speech styles include Krull (1989), Duez (1992), Crowther (1994), and Bakran and Mildner (1995). None of these studies found as big a difference between the locus equations of spontaneous and read words as what Krull (1987) found for her second participant.

Nevertheless, the fact that locus-equation slopes can for some speakers vary substantially from one speaking style to another raises the question of how the effect of place of articulation on the slope could be disentangled from the effect of speaking style (i.e. formal versus casual or Lindblom's (1990) hyper versus hypo). Might it be the case, for example, that the slope of an alveolar in hypo-articulated speech is similar to that of a bilabial in hyper-articulated speech? Brancazio and Fowler (1998: 29) report that, for one of their participants, the slope of /b/ in hyper-speech was 0.697, while the slope of /d/ in hypo-speech was 0.637. Nevertheless the average values for the study's three participants were 0.756 and 0.608 respectively. Thus it seems that the change in slope across speaking styles, while occasionally substantial, is not enough to cause full overlap in the slopes /b/ and /d/ at least (though Sussman et al. (1991: 1315) showed there can be overlap between /b/ and /g/, at least if different speakers are compared). But importantly, it should be remembered that even if the slopes for two places *could* under certain conditions overlap, the *intercepts* can nevertheless be different, so similarity of slopes does not on its own necessarily lead to the locus-equation space for two places of articulation colliding.

Even so, how to disentangle variations along the hyper-hypo dimension from those along the place-of-articulation dimension in locus equation slopes and intercepts is difficult. The obvious solution is to measure tempo (say, by measuring syllables per second) and to stipulate that the higher the tempo, the higher the coarticulation (i.e. to expect steeper locus

equation slopes in shorter syllables). However, experimental evidence shows that it *is* possible to hyperarticulate *and* speak rapidly (Lindblom, 1983). Thus to use tempo as a proxy for hypoarticulation is fallible since it would not be able to distinguish rapid hypo speech from rapid hyper speech. In summary, the use of locus equation slopes as an attribute for place of articulation across different speaking styles seems challenging, at least in light of Krull's (1987) finding for her two speakers. However, this challenge is more a product of the nature of hyper versus hypo speech rather than being a fault with the locus equations themselves. Also, Krull's (1987) finding of a large (0.2) difference in the slope between spontaneous-speech /d/ and read-speech /d/ in the speech of one of her speakers does not appear to have been found in other studies. For example the largest difference between spontaneous- and read-speech /d n l/ in Krull's (1989: 92) five speakers is 0.11.

The challenges involved in using locus-equation information that we have discussed up until now, though noteworthy, do not fundamentally question the idea that F2 transition information is sufficient for correctly identifying place of articulation. We now turn to a deeper issue: it turns out that the locus equation lines for two places of articulation can intersect. This means that there are zones in $F2_{onset}$-$F2_{mid}$ space which can be associated with more than one locus-equation line, i.e. more than one place of articulation. These zones of confusion have been identified and discussed by Sussman et al. (1998: 256-257) and are illustrated by the following diagrams:

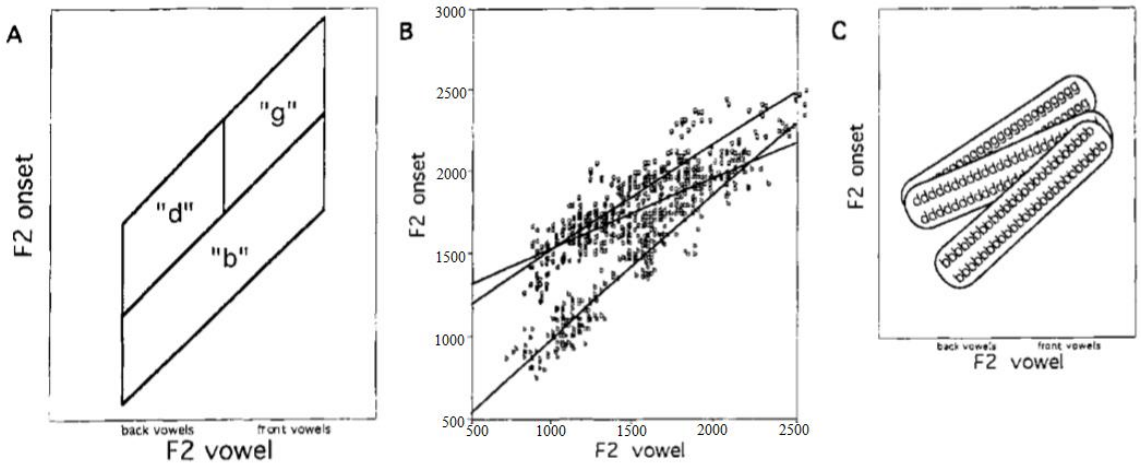


Figure 2.16: Three diagrams that illustrate the areas of overlap in the locus-equation space of /b d g/. Figure A: Schematic /b d g/ identification territory map. Figure B: Raw $F2_{onset}$-$F2_{mid}$ data for 10 vowels of American English, showing that before front vowels /b d/ overlap; before back vowels /d g/ overlap. Figure C: A dominance hierarchy hypothesis for how place of articulation is identified in such areas of ambiguity. From Sussman et al. (1998: 257).

Sussman et al. note that /b d/ overlap before front vowels and /d g/ overlap before back vowels (Figure 2.16 B). They hypothesize that in cases of /b d/ ambiguity listeners will be biased in

favour of perceiving /b/ rather than /d/, whereas in cases of /d g/ ambiguity they will perceive /d/ rather than /g/.[3] Thus $F2_{onset}$ and $F2_{mid}$ are insufficient on their own to reliably identify place of articulation. Sussman et al. concede this: "The cues that allow normal identification of [d] in front vowel contexts and [g] in back vowel contexts are not to be found in this acoustic space" (p. 257). (If $F2_{onset}$ and $F2_{mid}$ are ambiguous in these contexts, what do listeners rely on to identify /d/ before front vowels and /g/ before back vowels? The authors suggest "the release burst, the shape of the onset spectra, and voice onset time" will also contribute during normal speech perception.) We shall see experimental evidence attesting to this ambiguity of /d g/ before back vowels shortly.

Brancazio and Fowler (1998) played to 32 listeners non-synthesized /b d g/ in which the burst had been removed. The listeners were told that the stimuli were either /b/, /d/, or /g/. Even with this prior-knowledge advantage, the listeners still only identified the consonants with 66.2% accuracy (p. 40). Prior to the loss of the burst, their accuracy had been 95.7%.

The research on locus equations has focused much more on F2 than F3, though exceptions include Nearey and Shammass (1987), Fruchter and Sussman (1997), and Sussman et al. (1998). Lindblom (1990, 1996) replotted Öhman's (1966) data and found that $F2_{onset}$ and $F2_{mid}$ on their own do not appear to be sufficient for distinguishing /b d g/. However, when Lindblom added F3 at vowel onset ($F3_{onset}$) to make a *three*-dimensional space, the three places of articulation were in fact separate from each other with no overlap. On a cautionary note, Öhman's data are based on *averages* of three repetitions, not individual tokens, and so hide between-repetition variability. This is important because Sussman et al. (1998: 251) have shown that there is a lot more token-to-token variability in $F3_{onset}$ than there is in $F2_{onset}$, with the result that locus-equation datapoints involving F3 show a poorer fit to the regression line than F2.

Indeed, a close comparison of Sussman et al.'s (1998: 251) F3 plot (which shows individual $F3_{onset}$ datapoints) with Lindblom's (1996: 1685) replotting of Öhman's data (which averages three repetitions into one) reveals that in Öhman's data, $F3_{onset}$ for /g/ is never higher than ca. 2,450 Hz, whereas in Sussman et al.'s data, $F3_{onset}$ for /g/ ranges all the way between 1,900 Hz and 2,900 Hz. Thus in Öhman's data $F3_{onset}$ shows almost no overlap between /b g/ and /d/ whereas in Sussman et al.'s dataset there is considerable $F3_{onset}$ overlap between /g/ and

[3] I think the reason for this hierarchy is related to a fact noted by Fant: when /b d g/ contain voicing, /b/ has the shortest/weakest burst, followed by /d/, followed by /g/ (Fant, 1973: 64). This can be seen in Fant's (1973: 112) spectrograms of Swedish [bi di gi]: [bi] has no burst, [di] has a diffuse burst, and [gi] has a short but non-diffuse burst. Also, Fant notes the mean VOT of Swedish /b d g/ are 8, 12, and 20 ms respectively. Notice how this ordering of burst prominence follows exactly the same hierarchy as that found in Sussman et al.'s (1998: 257) theory.

/d/. In sum, the separability between /b g/ and /d/ that Lindblom achieved using $F3_{onset}$ appears not to generalize to other datasets. In Section 5.4.2 the effect of normalizing $F3_{onset}$ by individual speaker on improving the classification accuracy of place of articulation will be presented.

A perceptual study that suggests the importance of F3 is that of Mann (1980). Mann synthesized a set of CV stimuli whose formant transitions ranged between those appropriate for /dɑ/ and /gɑ/ (shown in Figure 2.17).

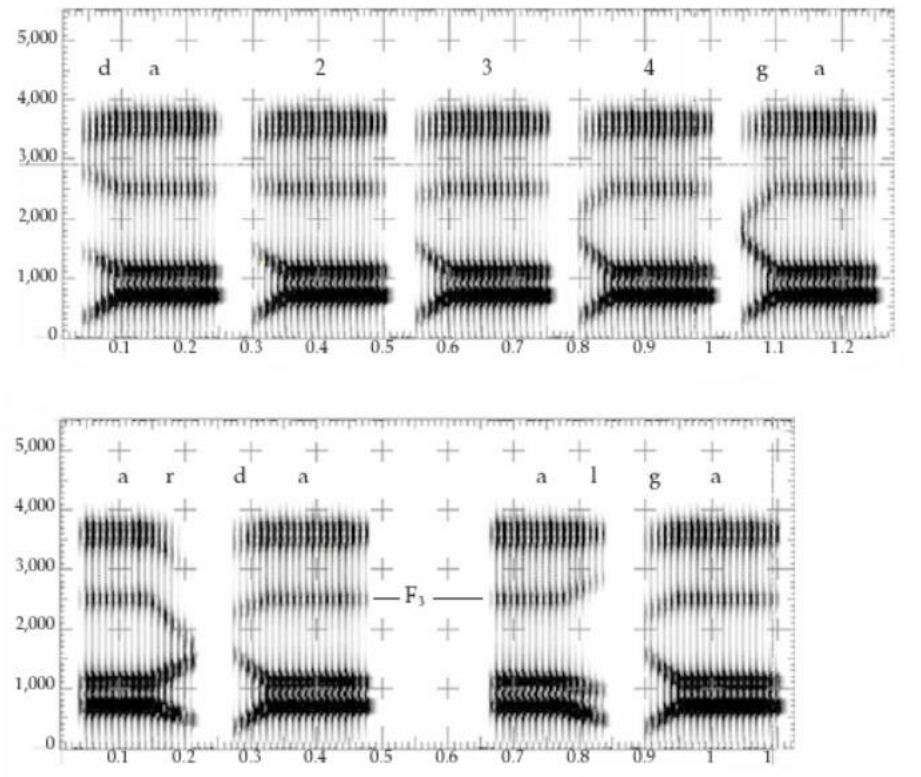


Figure 2.17: Spectrograms of the stimuli used in Mann's (1980) study.
Adapted from Johnson (2012: 103).

Stimuli in the middle of this continuum sounded ambiguous between /dɑ/ and /gɑ/. She then synthesized two preceding syllables, /ɑl/ and /ɑr/, and found that when listeners heard the ambiguous /dɑ ~ gɑ/ paired with /ɑr/, they were more likely to perceive it as /dɑ/, whereas when the ambiguous /dɑ ~ gɑ/ stimulus was paired with /ɑl/, they were more likely to perceive it as /gɑ/. The most plausible explanation for this finding is that because the preceding /l/ raises F3, the F3 at the beginning of the following syllable sounds lower, hence the perception of /ɑlgɑ/ (/g/ has a lower $F3_{onset}$ than /d/). When the preceding segment is /r/, in contrast, this raises the perceived frequency of the $F3_{onset}$ in the following syllable.

Mann's finding is similar to that of Öhman (1966), except that it pertains to F3 rather than F2 and was discovered using synthetic speech rather acoustically analysing human speech. In both cases the formant frequency of the segment *before* the plosive (V1 in the case of Öhman, /l ~ r/ in the case of Mann) appears to have an impact on the perceived formant frequency of

the segment *after* the plosive. Chapter 5 begins with an investigation of this kind, though it is restricted to intervocalic tokens: I will investigate the effect of a preceding vowel on the frequency of $F2_{onset}$.

One cautionary note regarding Mann's findings is necessary. Firstly, Mann used stimuli with no release burst to be sure that listeners' responses were definitely caused by the formant transitions. In real speech, there would very likely be some degree of burst. Also, Miller and Nicely (1955: 347) suggest that the F2 transitions in /dɑ/ and /gɑ/ are in fact too similar to each other to reliably distinguish them, and suggest that the burst is needed to do so. This ambiguity of the transitions in /gɑ/ versus /dɑ/ has been verified by Li et al. (2010), who found that a /dɑ/ in which the earliest 5 ms of the burst was removed was identified as /dɑ/ by only 70% of listeners. These researchers also found that a /gɑ/ in which the burst was removed entirely was perceived to be /dɑ/ by 100% of the listeners. Thus it seems that in non-synthetic speech, the formant transitions of /d g/ before /ɑ/ really can be ambiguous, and that F3 does not resolve this. This perceptual bias in which burstless /gɑ/ is perceived as /dɑ/ confirms Sussman et al.'s (1998: 257) perceptual dominance hypothesis that we saw in Figure 2.16: they theorized that before back vowels (the context in which /d g/ $F2_{onset}$-$F2_{mid}$ overlap) listeners would be biased to hearing burstless stimuli as /d/ rather than /g/.

One final point about most locus-equation studies is that they have tended to examine the formant patterns of CV sequences only. Öhman's (1966) findings suggested that the acoustics of CV in VCV sequences is far more complicated than in CV-only sequences and Mann's findings for synthesized /ɑlgɑ/ and /ɑrdɑ/ add weight to this. Yet there have been relatively few locus equation studies that have attempted to scale up to this three-segment sequence, though one exception is Lindblom and Sussman (2012). The result is that one has to take the striking linearity of locus equation regressions as being well established for CV contexts only. In Section 5.1 the acoustics of VCV sequences will be investigated in order to quantify the effect of V1 on $F2_{onset}$ on a more naturalistic dataset than that collected by Öhman.

To summarize the main findings of locus equation studies:

(1) The slope for velars is different before front vowels and back vowels;

(2) The slope for a given place also varies depending on whether articulation is hyper or hypo, and there is no obvious way of separating this effect from place of articulation itself. Nevertheless the differences between hyper and hypo (or read and spontaneous) speech do not appear to be large enough to make the slopes of two different places of articulation the same.

(3) The locus equation lines of different places of articulation can intersect, and it is not clear how to correctly identify the place of articulation of tokens in the overlapping area

without resorting to some other acoustic correlate, e.g. burstless /g/ before back vowels is known to be identified as /d/.

(4) Some of this ambiguity might be mitigated by including a third dimension, $F3_{onset}$ (Lindblom, 1996). However, Brancazio and Fowler (1998) found that listeners identified non-synthesized burstless /b d g/ (which contains F3) with just 66% accuracy, and Sussman et al.'s (1998) F3 locus equations show considerable overlap in $F3_{onset}$ between /g/ and /d/, unlike Lindblom's (1996) diagrams of Öhman (1966).

The overall conclusion is that formant information appears to be able to identify the place of articulation of /b d g/ well above chance but well below 100% accuracy. This is true whether the study is purely acoustic in nature (e.g. discriminant analysis on measured formant frequencies) or perceptual (the response of listeners to formant-only tokens). Note also that this result pertains to /b d g/; for /p t k/, as will be shown in Section 5.4.1, the role of formant information is expected to be even less.

2.2.4 Further Formant-Based Attributes

Stevens et al. (1999) examined dF2 while Suchato (2004: 49) examined both this and dF3. These attributes subtract $F2_{onset}$ or $F3_{onset}$ from their values 20 ms later. Using F-ratios and estimated Maximum-Likelihood classification error Suchato found dF3 to be a weak attribute but dF2 to be relatively strong. One possible explanation for this difference between F2 and F3 is related to what we noted in the discussion of locus equations above: the relationship between $F3_{onset}$ and $F3_{mid}$ seems to be more haphazard than that between $F2_{onset}$ and $F2_{mid}$, as indicated by the poorer fit to the locus equation line for each place of articulation of $F3_{onset}$ and $F3_{mid}$ compared to $F2_{onset}$ and $F2_{mid}$. Nevertheless, the theoretical motivation behind either attribute is not clear. If anything, theoretical considerations would lead us to expect that dF2 should not be a strong attribute, because it has been known since at least Delattre et al. (1955) that formant transitions for [d] before [i e] point *up*, before back vowels point *down*, and before [ɛ] are flat. Thus the dF2 values for [d] would vary between positive, negative, and zero depending on the vowel. Nevertheless, given its reasonable classification accuracy, dF2 will be one of the formant attributes included in the pilot study (Chapter 3).

Suchato also examined F3o-F2o, which subtracts the frequency of $F3_{onset}$ from $F2_{onset}$. This attribute was not as strong as dF2. The attribute was intended to capture the difference between velars and non-velars, since velars often have $F2_{onset}$ and $F3_{onset}$ closer together than the other places of articulation ("velar pinch"), as was shown in Figures 2.1 and 2.2 at the beginning of this chapter. However, as we saw in Figure 2.9 (which showed the acoustics of /g/ before a back vowel), F2 and F3 are actually quite far apart in frequency in back velars. Another

difficulty with the F3o-F2o attribute is that some *non*-velar contexts, such as [bi], also involve $F3_{onset}$ and $F2_{onset}$ being close in frequency (see, for example, the diagram of /ibi/ in Öhman, 1966: 160).

Suchato examined F2b and F3b, which measure the frequency of F2 and F3 in the burst rather than at the vowel onset; he found them to be weaker at classifying place of articulation than F2 and F3 measured at vowel onset. This result is contra Modarresi et al. (2005), who recommended measuring $F2_{onset}$ in the burst and found it classified place of articulation better than measuring it at vowel onset. Both of these studies, however, relied on measuring the burst's F2 manually (Modarresi et al., 2005: 106; Suchato, 2004: 42). Furthermore, it is not even clear whether F2 appears consistently. Blumstein and Stevens (1979) found that only approximately one third of their /d/ bursts and one sixth of their /t/ bursts contained F2. This inconsistent presence of formants in the burst has also been observed by Nossair and Zahorian (1991: 2981): "For many tokens, particularly labial and alveolar stops, the formants are simply not well defined in the burst and aspiration segments."

A further difficulty with extracting F2 from the burst is that some bursts can contain formants generated by the subglottal space (Blumstein and Stevens, 1979; Fant et al., 1972). Section 4.4 presents the results of inspecting several thousand plosive bursts and plosive aspiration for the presence of F2 and F3, in which the visibility of the formants in each token was rated on a scale from 1 to 5.

2.2.5 Conclusion

In this section we have reviewed a variety of approaches to formant-based information. In Chapter 3 (the pilot study) wide variety of formant-based attributes will be tested, including some of the attributes presented above as well as certain new attributes. The attribute that classifies the best out of this group will be brought forward to the main study (Chapter 5).

Most research, as we have seen, has used $F2_{onset}$ and $F2_{mid}$ as separate attributes (via locus equations). Given the relatively strong correlation between $F2_{onset}$ and $F2_{mid}$ (typically between ca. 0.4 and 0.75), collapsing the two into a single attribute seems justified, and the 1950s locus theory was one method for doing so. As Lindblom and Sussman (2012) note, the abandonment of the locus theory was prompted by Öhman's (1966) findings for VCV sequences of significant coarticulatory influence of V_1 on $F2_{onset}$.

However, in our examination of Öhman's (1966) study we noted a number of its features:

1. The data collected came from the speech of a single speaker.
2. The material consisted of nonce VCV sequences.

3. The speaker was requested to pronounce each of these vowels with the same stress.

It is surprising that the findings of just a single small-scale study were deemed sufficient grounds for abandoning the locus theory. Chapter 5, which examines formant-based attributes, begins by revisiting VCV sequences but using the speech of 20 speakers reading real words in real sentences. The aim is to quantify the influence of V_1 on $F2_{onset}$ in natural speech, to see if it really is as large as Öhman (1966) found in his data. The results of this will be used to undergird the approach to formant information in Chapter 5, in which $F2_{onset}$ and $F2_{mid}$ are collapsed into a single attribute.

2.3 Burst Information

In Section 2.1 we introduced the release burst's acoustics. In this section we revisit this in greater detail with the following question in mind: How can the information about place of articulation in the release burst be best represented?

Answering this question will be done in two stages. The first stage will review the burst attributes used in previous studies. This will show that there has been considerable uncertainty as to how to capture the information in the burst, with a consequent proliferation of acoustic attributes.

Given this uncertainty, the second stage will be to examine the results of several kinds of perceptual studies, with a view to establishing what property of the burst is likely to be the most important. These perceptual studies come in several forms:

(1) The burst and/or the transitions are eliminated and presented to listeners for identification;

(2) Syllables presented with varying degrees of background noise;

(3) Plosives in which certain frequency regions have been amplified or attenuated;

(4) Simulation of forward masking from the burst on the prominence of $F2_{onset}$.

Taken together, these studies will refine our sense of where the most important information in the burst for distinguishing place of articulation is likely to be found.

2.3.1 Acoustic Attributes of the Burst

2.3.1.1 The Burst Peak

As we saw in Section 2.2, the arrival of the pattern playback allowed the development of carefully controlled artificial stimuli to study the formant transitions systematically. It also allowed researchers to do the same for the release burst.

Liberman et al. (1952) studied /p t k/ using this method. The authors note that, although release bursts contain energy at a wide range of frequencies, they believed that a characteristic difference between the three places of articulation lay in the frequency at which the energy is centred (p. 499). This frequency at which the burst's amplitude is greatest is known as the peak or, less often, as the "burst frequency" (Zue, 1976) or "burst" (Allen and Han, 2011). This belief underlay Liberman et al.'s decision to represent the burst using a single 600-Hz-bandwidth blob of energy rather than attempting to paint the entire burst. Figure 2.18 shows these stimuli:

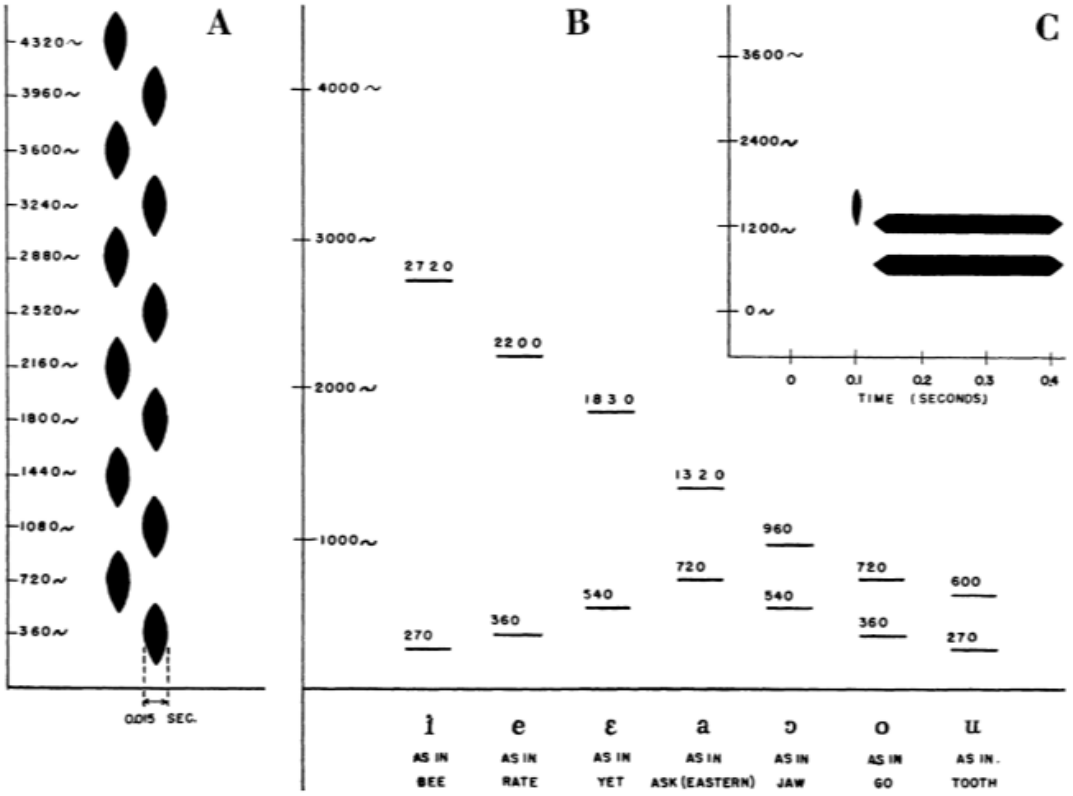


Figure 2.18: Schematic diagrams of the stimuli used by Liberman et al. (1952) for the study of the release burst. Panel A shows the 12 frequencies at which the burst peak was located; Panel B shows the seven following vowels with which these burst stimuli were paired, yielding a total of 84 stimuli; Panel C provides an example of such a pairing. From Liberman et al. (1952: 502).

The 84 stimuli were played to 30 listeners, who were strongly urged to identify all of them as either /p/, /t/, or /k/, "even though in some cases the judgment might represent no more than a guess" (p. 505) and to rate their confidence in their judgment in each case on a four-point scale.

These were their results:

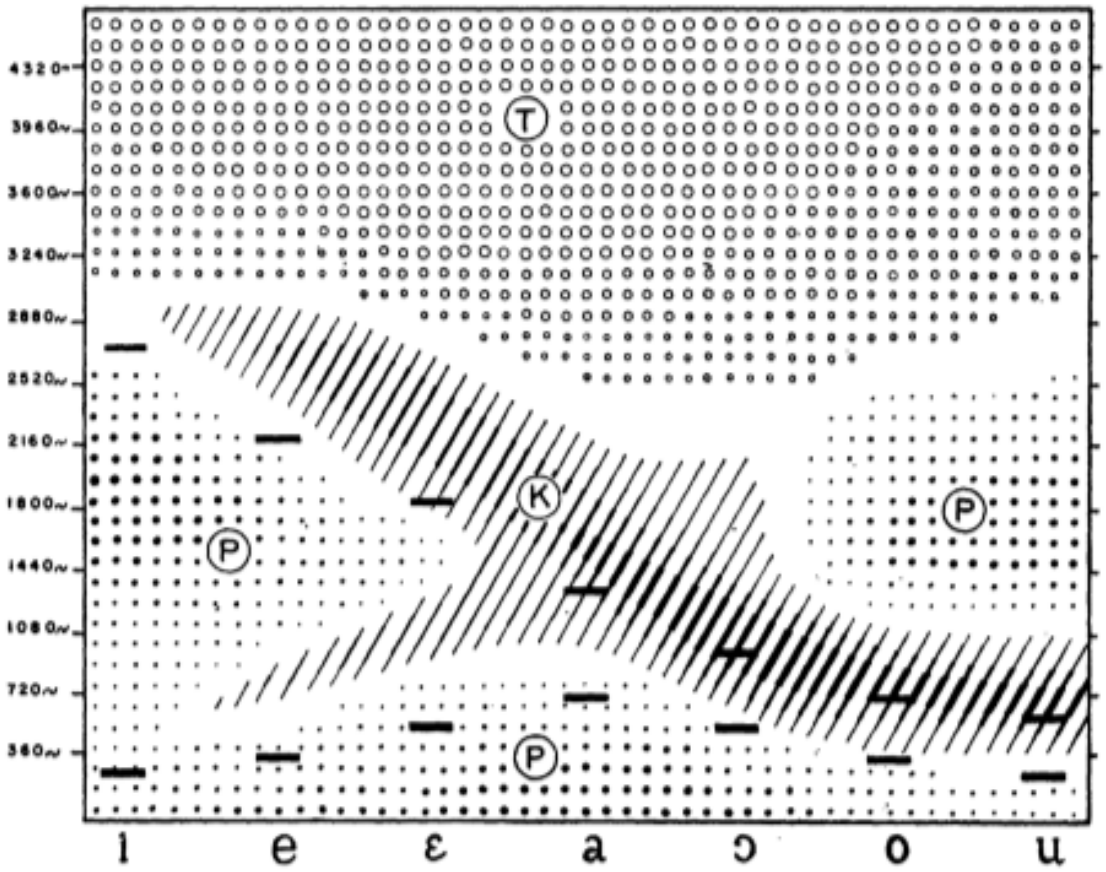


Figure 2.19: Diagram showing the results of Liberman et al.'s study.
The size of the dots/lines indicates the degree of confidence the listeners had in their judgments. From Liberman et al. (1952: 509).

A relatively high-frequency peak yielded a /t/ percept no matter what the following vowel. In contrast, for /k/ the frequency of the peak varied heavily depending on the following vowel: the higher the vowel's F2, the higher the burst peak had to be for listeners to hear it as velar. The most convincing velar stimuli involved the burst peak being fractionally higher than the vowel's F2 for back vowels, and moderately higher than the vowel's F2 for front vowels.

For /p/ the picture is somewhat complicated: before /i/ if the burst peak is somewhat lower than the vowel's F2 the result is a clear percept of bilabial place, and the same can be said of /a ɔ/ context; before /e ɛ/ it was difficult to find any burst peak frequency that sounded convincingly bilabial, though a few of the very low-frequency ones did; before /o u/ a very low-frequency burst could again give a perception of bilabial place but strangely, so could a burst peak that was much higher in frequency. The fact that two very different burst-peak frequencies could both give the bilabial percept before /o u/ suggests that the burst stimuli used in the study may well have been missing something important about what makes a burst sound bilabial. Recall that all the burst stimuli had the narrow bandwidth of 600 Hz (i.e. a relatively compact or 'spiky' frequency spectrum), which makes them different from most real-life bilabial bursts,

which tend to have flatter spectra (as we shall see for the present study's dataset in Section 6.1). Variability in measuring the bilabial burst spectrum is something that has been observed in other studies. Zue (1976: 112) writes:

"In measuring the burst frequency [= burst peak] for the labials, it was found that there is a wide range of variation in the values found. Since the spectra of /p,b/ show no distinct burst frequency and the RMS amplitudes of these stops are weak, we have decided not to present results on the burst spectrum for labials."

Abdelatty Ali et al. (2001) found the burst peak to be the strongest of the burst-based attributes they studied. It correctly classified 74.5% of alveolars, 81.5% of velars, and 58% of bilabials. They noted, however, that the burst peak was highly context-dependent (p. 835). They found that if the $F2_{mid}$ was added to the classification, the accuracy of the burst peak improved to 89% for alveolars, 87.6% for velars, and 66% for bilabials.

Suchato (2004) did not examine the burst peak attribute. This is unfortunate as it would be interesting to compare the attribute's performance to other frequency-domain attributes such as the centre of gravity. This is one gap which the present study seeks to fill.

In the main study of this thesis (Chapters 6 and 7), two aspects of the burst peak will be measured: (1) its frequency, and (2) its amplitude. The first of these will be labelled 'AllPeakHz' or 'AllPeakBark', depending on the spectral representation from which it is derived, while (2) will be labelled 'AllPeakdB', 'AllPeakPhon', or 'AllPeakSone', again depending on the spectral representation from which its derived.

2.3.1.2 Grave and Acute Spectra

Halle et al. (1957) analysed three speakers' productions of monosyllabic words with the six stops in 11 contexts. They devised a two-tier method of classification: the consonant's intensity between 700 and 10,000 Hz was subtracted from its intensity between 2,700 and 10,000 Hz. If these figures differed little from each other, it indicated that the consonant had a large concentration of energy in the high-frequency region: such consonants were classed as 'acute', whereas a big difference in the two values indicated that the consonant had most of its energy at relatively lower frequencies, classed as a 'grave' consonant. The authors found that alveolars and pre-front-vowel velars were classed as 'acute', bilabials and pre-back-vowel velars as 'grave'.

To further divide these two groups to yield the three-way distinction of place of articulation, Halle et al. split the grave consonants into two categories by identifying the two

tallest peaks in the spectrum and subtracting the amplitude of the lower-frequency one from the other. They plotted this peak-intensity difference (in dB) on the vertical axis. On the horizontal axis they plotted the higher-frequency peak's frequency. This classification system might sound convoluted but what Halle et al. were essentially trying to capture was the flatness of the bilabial burst relative to the velar burst:

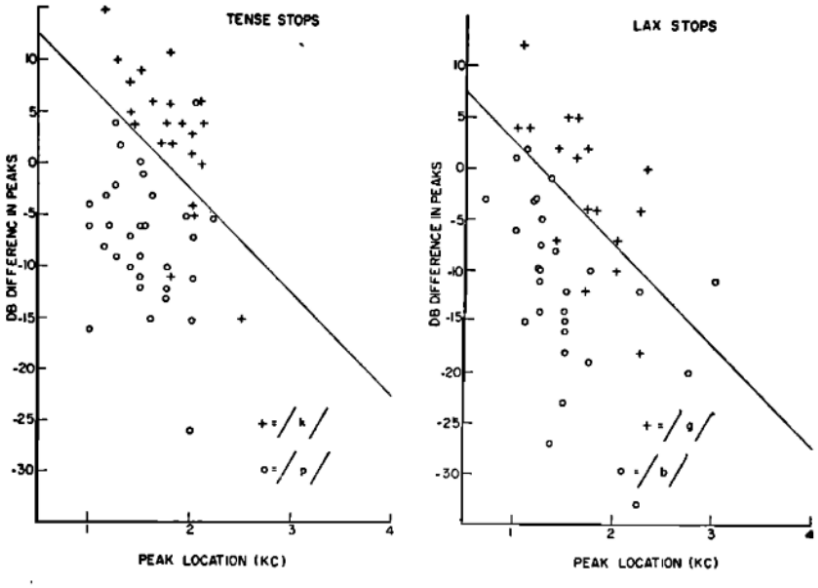


Figure 2.20: Diagrams illustrating how Halle et al. (1957) 'grave' consonants, i.e. pre-back-vowel /k g/ from /p b/.

The left panel displays results for the /k p/ distinction (85% accurate), the right panel displays the /b g/ distinction (78% accurate). The intensity of the two tallest peaks in the spectrum were noted and the intensity (in dB) of the lower-frequency peak was subtracted from that of the higher-frequency peak (vertical axis). The horizontal axis shows the frequency of the higher-frequency burst peak (in kHz). From Halle et al. (1957: 112).

Halle et al. separated /t d/ from pre-front-vowel /k g/ by measuring the mean sound level of the burst spectrum between 300 and 10,000 Hz and subtracting from it the mean sound level of the spectrum between 2,000 and 4,000 Hz. The velars were found to have a lot of energy concentrated in this latter region relative to the alveolars. This acoustic attribute separated /k g/ from /t d/ with 85% accuracy, though as with the attribute for separating /k g/ from /p b/ it was slightly easier to separate place in the voiceless series than in the voiced series. The overall classification accuracy of place in the study was 79%.

Halle et al.'s study was relatively small-scale, involving the speech of just three speakers and a relatively small dataset. Their precise attributes for distinguishing the three places of articulation do not appear to have been used by later works. Their seemingly unusual approach of first splitting the consonants into 'grave' and 'acute' rather than going directly to a three-way place of articulation reflected one of the then current speech-feature theories of their time (Jakobson et al., 1952).

Perhaps the most noteworthy difference between this study and Liberman et al. (1952) is in their recognition of more than one peak in the burst envelope. Their subtraction of the intensity of the lower-frequency peak from the higher-frequency peak appears to be the first attempt to capture the difference between velars and bilabials with reference to the amplitude domain rather than the frequency domain.

The grave-acute classification system will not be tested in the present study due to difficulties in replicating Halle et al.'s notion of a 'peak': some spectra are relatively flat and hence not appear to have much of a peak; it is unclear how such cases would be dealt with within the criteria of the original study.

2.3.1.3 Spectral Templates

Blumstein and Stevens (1979) designed templates for classifying the burst spectra of the three places of articulation: diffuse-rising for alveolars, diffuse-falling for bilabials, and compact for velars. These templates were then fitted to each burst envelope.

The diffuse-rising template is fit by first identifying the tallest peak in the burst spectrum above 2,200 Hz and then superimposing the dotted lines of the template onto the spectrum in such a way that the upper line of the template touches the tallest peak:

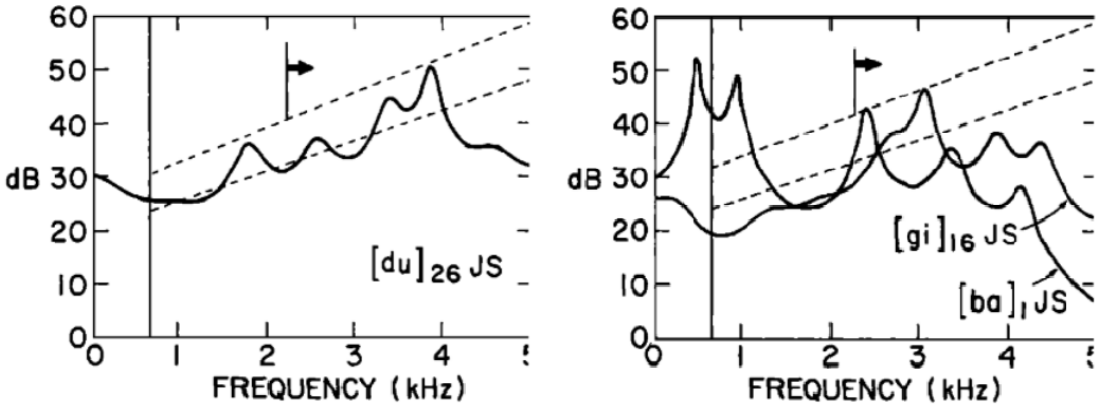


Figure 2.21: Diagrams illustrating how the diffuse-rising template (= the two dotted lines) for identifying alveolars was fit to spectra.

The tallest peak in the spectrum above 2,200 Hz is used to fit the template. For a spectrum to be classed as diffuse-rising, at least one other peak must fall between the two dotted lines of the template, and the higher-frequency peak must be greater in amplitude than the lower-frequency one. In the [du] example (lefthand panel) these criteria are satisfied and so the consonant is correctly identified as alveolar; in the righthand panel the [gi] is not classed as alveolar by the template because only one of the spectral peaks is situated inside the template's dotted lines; likewise for [bɑ]. From Blumstein and Stevens (1979: 1005).

If more than one peak is found inside the template's dotted lines and if the higher-frequency one is greater in amplitude than the lower-frequency one, then the spectrum is classed as diffuse-rising (as shown in Figure 2.21).

These criteria, it need hardly be said, are complicated. Yet it turned out that further criteria had to be appended to make the template work: many alveolar spectra were found in which there was a peak between 800 and 1,600 Hz. These peaks were ignored by the template by stipulating that any peak occurring between 800 and 1,600 Hz whose peak fell 10-12 dB above the upper dotted line of template were to be omitted, as in the following example:

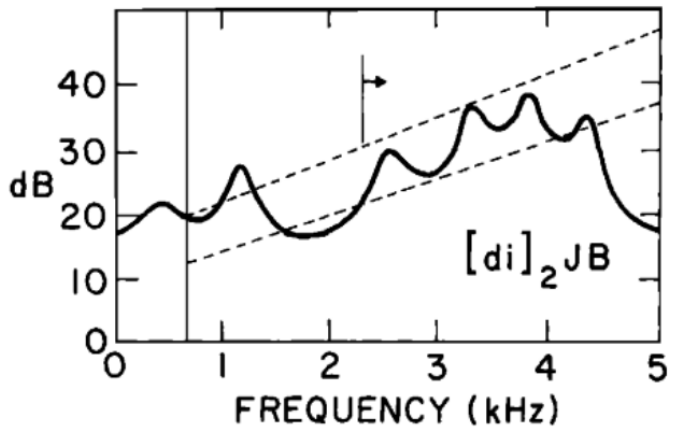


Figure 2.22: Example of an alveolar onset spectrum containing a prominent peak at ca. 1,200 Hz. This peak is ignored by the diffuse-rising template. From Blumstein and Stevens (1979: 1305).

The authors justified this decision by pointing out that even though the burst spectrum contains a prominent 1,200-Hz peak it nevertheless shows the diffuse-rising shape typical of alveolars in the region above 2,200 Hz. They regard the 1,200 Hz peak as being caused by a subglottal resonance, which can occur when there is a glottal opening (Fant et al., 1972) during the production of the consonant. Because this peak is not generated by the oral tract it does not bear information on the consonant's place of articulation.

However, the authors found that it is not just this subglottal resonance that can cause problems for the alveolar template: a peak in the spectrum corresponding to the $F2_{onset}$ frequency can occur as well. Because F2 is located in the mid-frequency region, this peak has the potential to misclassify the spectrum as velar. Therefore the authors made an ad-hoc modification to the diffuse-falling template as follows:

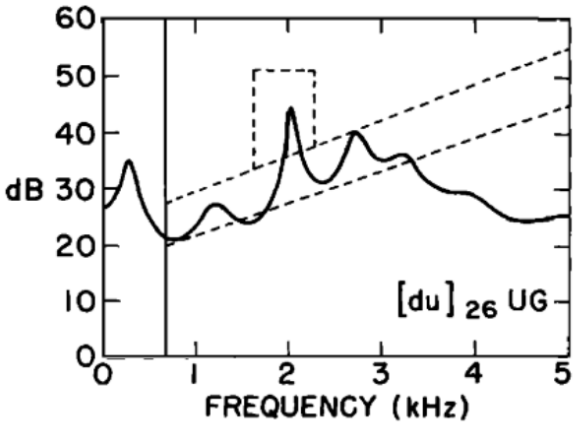


Figure 2.23: Modified diffuse-rising template for classifying alveolar place, showing a box around the $F2_{onset}$ frequency present. From Blumstein and Stevens (1979: 1306).

The authors found this $F2_{onset}$ peak in 33% of /d/ tokens and 17% of /t/ tokens.

Regarding bilabial spectra, the authors note that the spectrum tends to be either flat or somewhat falling (p. 1306). Here is the diffuse-falling template they designed to capture this:

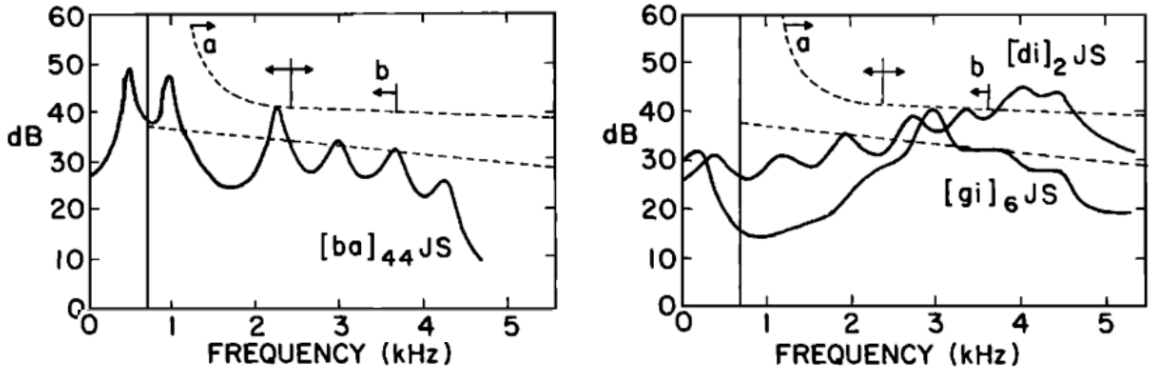


Figure 2.24: The diffuse-falling spectral template.
The largest peak in the spectrum between 1,200 Hz (point a) and 3,500 Hz (point b) is fitted to the template's upper dotted line. Between these lines at least two peaks must occur, one lower than 2,400 Hz, the other between 2,400 and 3,600 Hz. The bilabial example in the lefthand panel satisfies these criteria, whereas the alveolar and velar examples in the righthand panel do not. From Blumstein and Stevens (1979: 1306).

The compact template, recall, is intended for classifying velar place. Blumstein and Stevens were working with spectra whose frequency scale was linear. The auditory system, in contrast, utilizes a frequency scale that is non-linear (approximately logarithmic above ca. 500 Hz, as we shall see later). When Blumstein and Stevens examined the tallest peak in velar spectra, they found that the higher the frequency of the peak, the wider its bandwidth. In the auditory system, of course, these peaks would probably appear to be roughly equally wide, given that filter bandwidth increases with frequency (Moore, 2012: 76).

However, because the authors were using spectra with a linear frequency scale, it meant that their spectral template for velars had to broaden with increasing frequency:

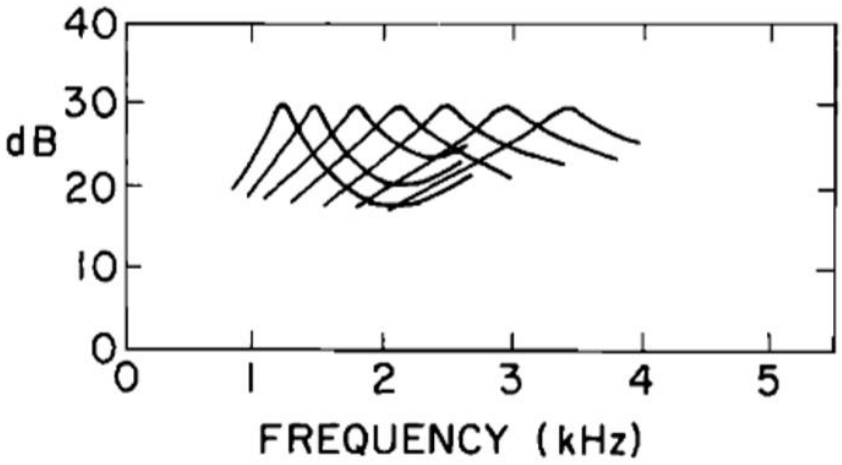


Figure 2.25: Schematic illustration of the compact template for identifying velar place.
Note how the bandwidth of the template's peak has to broaden with increasing frequency, an artefact of using a linear rather than a perceptually-motivated frequency scale. From Blumstein and Stevens (1979: 1007).

Here is an example of how the template discriminated velars from non-velars:

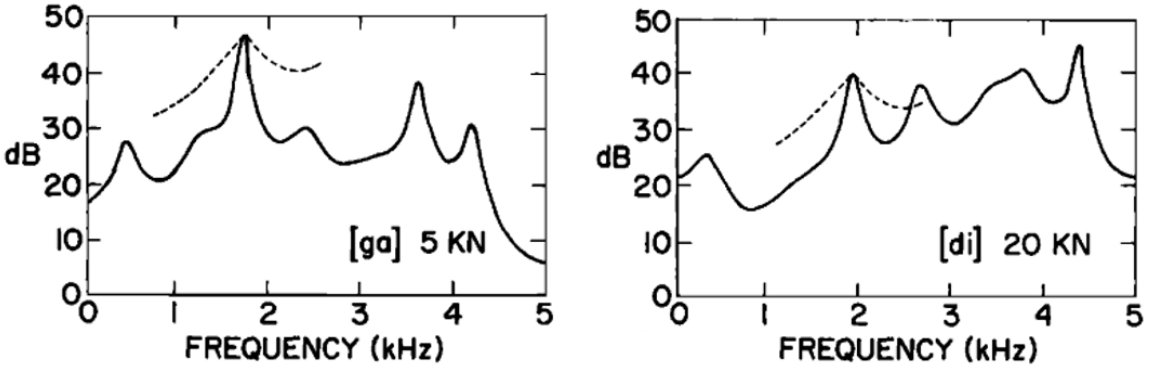


Figure 2.26: Example of a compact template being fitted to two spectra.
From Blumstein and Stevens (1979: 1307).

The lefthand panel above shows the successful fit of the compact template to a velar consonant: it is successful because the peak is the largest one in the spectrum, and the template intersects just a single peak and that peak is the highest one in the spectrum. In the righthand panel, in contrast, the template intersects another peak and the peak in question is not the largest peak, hence the spectrum is not classified as velar.

Blumstein and Stevens found that these three templates classified plosive place of articulation correctly with 84.3% accuracy. (N = 900; material was read by six speakers and involved five vowel environments.)

These results were obtained by applying the templates to the very beginning of the consonant's release. The authors also attempted to apply these templates in VC sequences to the part of the vowel at the very edge of the plosive closure but did not succeed at classifying place of articulation (perhaps because there is no burst just after the end of a vowel whereas there is one just before the beginning of a vowel).

To summarize: although the study obtained a reasonably good classification of plosive place, it did so using many arbitrary stipulations: the number of peaks, the cut-off frequencies, and other stipulations such as the 'box' around $F2_{onset}$ in the alveolar template appear to have been developed on an ad-hoc basis. For velar place there was in fact not one but several templates, a different template having to be chosen depending where on the frequency scale the peak happened to be located; this complication was an artefact of using a linear frequency scale.

These complexities perhaps explain why most subsequent research has not utilized templates. Spectral-template attributes will not be examined in the present study.

2.3.1.4 Dynamic Displays

Subsequent research elaborated on Stevens and Blumstein's approach in different ways. Kewley-Port (1984) for example, substituted the single 25.6-ms window for a 40-ms display consisting of a sample of the spectrum every 5 ms. This method improved classification accuracy from 84.3% to 88%, a small improvement given the nine-fold increase in information entailed by having multiple windows.

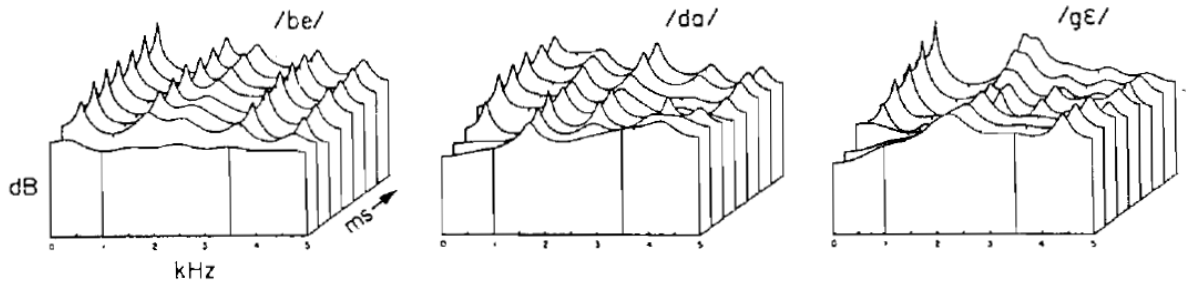


Figure 2.27: Examples of the dynamic displays used by Kewley-Port (1984).
The slices are 5 ms apart, with the closest slice being the earliest one. From Kewley-Port (1984: 323).

Kewley-Port explains her choice of dynamic displays by citing Fant's (1973) observation that the release burst of bilabials is ca. 5-10 ms in duration whereas that of velars is longer, namely 20-30 ms (pp. 324-325). This shows up in her dynamic displays as a rapid change in the spectrum of bilabials after the first 10 ms compared to a more slowly changing spectrum in the velars. However, could this difference between the two places not be more economically captured by measuring the burst duration? Given that the burst tends to be longer

in velars than bilabials, this burst duration measurement would add a second cue, one of a dynamic nature, without having to increase the amount of information on the onset nine-fold. This seems reasonable given that, by her own acknowledgement, the spectra for the velar burst over the 40-ms period show little difference from each other (p. 324).

Given the heavy redundancy and vast increase in information involved in dynamic displays, they will not be examined in the present study. Nevertheless, because the present study segments the burst manually (i.e. with high precision), an acoustic attribute that measures the burst duration will be included to quantify its improvement on the classification accuracy (Section 7.4).

2.3.1.5 Spectral Tilt

Another elaboration of Stevens and Blumstein's approach was that of Lahiri et al. (1984), who compared the amplitude of F4 and F2 by drawing a straight line from the peak of one to the peak of the other. This yielded a measure of spectral tilt. They performed this ratio at two points: (1) the release burst; (2) the onset of the vowel at the first three glottal pulses. The resulting classification rate was 91%, which at first blush appears to be an improvement on Kewley-Port (1984) and was successfully performed on three languages (English, French, Malayalam), which makes it noteworthy. On the other hand, the study only sought to distinguish /b/ from /d/ and did not investigate velars; in that sense, the classification accuracy does not reflect the real-life task of distinguishing three places of articulation.

Suchato (2004) examined spectral tilt in two ways: Ehi-E23 involved subtracting the total energy (in dB) of the high-frequency region (3,500 to 8,000 Hz) from that of the mid-frequency region (1,250 to 3,000 Hz). He found this to be the second strongest attribute in his study (in terms of its ML classification error; p. 95). The other spectral tilt measure in his study, Ahi-A23 (which he adopted from Stevens et al., 1999), involved subtracting the two *peak* amplitudes in each frequency region rather than using the frequency regions' total energies (which is what Ehi-E23 does). This attribute was also relatively strong (fourth strongest), though not as strong as Ehi-E23 (p. 95).

Given their strong performance in previous research, the present study will examine various kinds of spectral-tilt attribute (Section 6.4.5).

2.3.1.6 Difference in the Amplitude of a Frequency Region in the Burst Relative to the Following Vowel

Suchato (2004) investigated attributes in which the maximum amplitude of a frequency region in the burst was subtracted from the maximum amplitude in the same frequency region at the

beginning of the following vowel. He tested three such attributes: Avhi-Ahi, Av3-A3, and Av2-A2. The first of these pertained to the high-frequency region, namely from 3,500 to 8,000 Hz; while the second and third ones pertained to the mid-frequency region (1,500 to 3,000 Hz and 1,250 to 2,500 Hz respectively).

Presumably the intention with such attributes was to normalize for speaker distance. That is, if a speaker is far away from the listener, then the amplitude of everything in their speech signal will be lower. Thus using absolute amplitudes as attributes is vulnerable to differences in acoustic conditions that are irrelevant to place of articulation. If, however, the intensity of a given stretch of speech (e.g. the burst) is instead gauged relative to some other stretch of speech (the vowel), then perhaps it would make the amplitude measured by an attribute more robust to differences in speaker distance and other acoustic changes that are irrelevant to identifying the place of articulation.

Suchato found that none of these attributes was strong at classifying place of articulation (p. 94). This is perhaps unsurprising given that voice quality – which affects the energy in a given frequency region considerably – varies considerably between individuals and, within an individual, varies within the intonation phrase (see Gobl and Ní Chasaide (2010) for further details). Thus gauging the amplitude of a region in the burst relative to the amplitude of the same region in the speaker's voiced speech does not appear to be a promising approach (unless perhaps one first estimated the speaker's phonatory quality and used this to correct the F1 amplitude prior to using it to normalize the burst amplitude).

Given the relatively weak performance of these attributes, they will not be examined in the present study.

2.3.1.7 Difference in the Peak Amplitude of a Frequency Region in the Burst Relative to the Following Vowel's F1 Amplitude

The logic of this type of attribute is similar to the one described in the previous section except that instead of subtracting the peak amplitude in a particular frequency region of the burst from the peak amplitude in the *same* frequency region of the vowel, it is instead subtracted from the amplitude of the vowel's *F1*. Suchato (2004) employed two such attributes, Av-Ahi and Av-Amax23 (which will be termed "HiPeak-F1(dB)" and "MidPeak-F1(dB)" in the present study). The former subtracts the amplitude of F1 at the onset of the following vowel from the peak amplitude in the burst's high-frequency region, while the latter does the same for the mid-frequency region. Suchato found these attributes to be stronger than the ones in the previous section, though not as strong as his spectral tilt attributes (Ehi-E23 and Ahi-A23) nor as strong as centre of gravity.

It would be helpful to know to what extent the F1 amplitude actually improved these attributes; unfortunately Suchato (2004) does not contain results for how Ahi and Amax23 classify without the inclusion of the F1 amplitude. Thus one of the aims of the present study is to fill this gap (Section 6.4.7).

2.3.1.8 Spectral Moments

Mathematically, a moment is a method of quantifying the shape of a set of points. Given that a spectral envelope consists of a set of points in a frequency-by-amplitude space, moments can be used to describe the envelope's shape.

The basic core from which spectral moments are computed is the following formula (adapted from Harrington, 2010: 298):

(1)

$$\frac{\sum a(f-k)^m}{\sum a}$$

in which *a* represents the amplitude of a given spectral component, *f* represents its frequency, *k* is a constant and *m* varies depending on the moment being calculated.

For the first moment, the centre of gravity (CoG), $k = 0$ and $m = 1$, whereas for the other three spectral moments k = CoG and m = the spectral moment number (viz. 2 for standard deviation, 3 for skewness, 4 for kurtosis; there are also additional differences between the formulas for these three – more on this shortly). This results in the following variant of Formula (1):

(2)

$$CoG = \frac{\sum af}{\sum a}$$

That is, the amplitude of each spectral component is multiplied by its frequency, and these products are added. This sum product is then divided by the sum of the said components' amplitudes.

What is the logic of CoG? Essentially it indicates the location (in the frequency domain) of the most intense components of the burst. Thus it tends to have high values for alveolars and lower values for velars and bilabials. This means that the attribute is similar to Burst Peak. Although both attributes involve the entire spectrum as input, they differ in that centre of gravity is calculated using every single component in the envelope, whereas the Burst Peak simply reports the frequency of the most intense spectral component.

Suchato (2004) moved a 6.4-ms Hanning window over the burst at 1-ms intervals beginning at 7.5 ms before the burst and continuing until a specified point thereafter. He tested three variants of CoG: cgF10a, cgF20a and cgFa. In cgF10a the windows end 10 ms after the onset of the burst; in cgF20a the windows end at 20 ms after the onset; in cgFa they continue all the way up to the onset of voicing, i.e. this variant of CoG blends the burst acoustics with the acoustics of any following aspiration. Suchato found cgF10a to be the strongest of the three variants. This is not surprising given that the source in the aspiration is located at the glottis rather than at the place of articulation, as is the case for the transient and frication. Thus the inferior performance of cgF20a and cgFa is presumably due to the fact that they were more likely than cgF10a to mix the aspiration acoustics in with the burst. This is likely to have distorted the acoustics of voiceless stops in particular, for the obvious reason that they have far longer aspiration than the voiced series (and hence the aspiration would have constituted a larger proportion of the averaged CoG). Out of the voiceless stops, /t/ is particularly likely to have been adversely affected because its concentration of energy predominates in the high-frequency region in its burst but not in the aspiration. Suchato's mean alveolar value for cgF10a was 2.7 kHz, which rises to 3.0 kHz for cgF20a and falls to 2.4 kHz for cgFa. These figures seem to indicate that the most characteristically alveolar (i.e. high-frequency) values reside 10 to 20 ms after the beginning of the burst. Nevertheless the fact that cgF10a was the strongest of the three attributes suggests that, on average, it is the earliest part of the burst that contains the best information for place of articulation (though Section 8.3 will discuss some occasional exceptions to this).

In any event cgF10a turned out to be the strongest burst-based attribute of all the ones Suchato tested (though it should be noted that he did not test the Burst Peak).

The second spectral moment, standard deviation, is given by the following formula:

(3)

$$SD = \sqrt{\frac{\sum a[(f - CoG)^2]}{\sum a}}$$

The formula quantifies (in frequency) how far each spectral component is from the centre of gravity, CoG. The amplitude of each spectral component (represented by a) is multiplied by the distance (in frequency) of that component from the centre of gravity (this distance is squared to prevent negative values) and then these frequency × amplitude products are summed. This sum product is divided by the sum of the envelope's component amplitudes, $\sum a$. The resulting figure is known as the variance; the square root of this figure is the standard deviation, SD.

The function of standard deviation in effect is to quantify the variability in the spectral components' amplitudes.

In Section 2.1 bilabials were hypothesized to have the flattest spectra on average, with velars and alveolars expected to be less flat due to the presence of mid- and high-frequency burst peaks respectively. In terms of the above formula, we would thus expect the standard deviation to be largest for bilabials, smallest for velars and alveolars.

The formula for standard deviation given above computes standard deviation in the frequency domain. However, given that standard deviation effectively measures variation in *amplitude*, a different version of the above formula can be made in which the frequency domain is bypassed in favour of the amplitude domain alone, as follows:

(4)

$$SD = \sqrt{\frac{\sum[(a - \mu)^2]}{n}}$$

In the above formula μ represents the mean amplitude of the spectrum, *a* represents the amplitude of each spectral component, Σ represents the sum of these (squared) amplitude deviations, and *n* is the number of components in the said spectrum. The aim of this attribute is to capture the difference between bilabials (which have a relatively flat spectrum) and velars and alveolars (which have a relatively peaky spectrum; see Section 6.1 for diagrams). In Section 6.4.1, its performance will be compared with the more complicated frequency-domain version of SD in Formula (3) to see which has the higher classification accuracy (Section 6.4.1). For clarity, in Chapter 6 the SD in Formula (3) will be referred to as SDFreq and that in Formula (4) will be referred to as SDAmp. In Chapter 3 (the pilot study), only one of the standard deviation formulas is examined, namely SDFreq, which will be referred to as simply 'SD'.

The third spectral moment, skewness, is given by the following formula:

(5)

$$Skew = \frac{\sum a[(f - CoG)^3]}{[\sum a].[SD^3]}$$

Skewness, as its name indicates, measures how skewed or lopsided a probability distribution is. Its function thus overlaps considerably with that of centre of gravity, since both capture an aspect of the points' distribution in the frequency domain. Alveolars are expected to have the lowest (i.e. negative) values, since negative skew involves a long lefthand tail on the spectrum; velars and bilabials are expected to have higher values.

Kurtosis measures the weight of the tails relative to the rest of the distribution. It is given by the following formula:

(6)

$$Kurt = \frac{\sum a[(f - CoG)^4]}{[\sum a].[SD^4]} - 3$$

The reliability of skewness and kurtosis has been called into question by McNeese (2016), who has shown that the accuracy of these two statistics is heavily dependent on sample size. Even with a large sample size of several hundred datapoints, he found that these statistics did not yield very good estimates of the true skewness and kurtosis. McNeese quotes the statistician Wheeler (2004: 56), who writes: "[…] skewness and kurtosis are practically worthless. […] The statistics for skewness and kurtosis simply do not provide any useful information beyond that already given by the measures of location and dispersion." A further difficulty with kurtosis in the context of phonetics has been articulated by Harrington (2010: 300): "Kurtosis is often described as a measure of how "peaked" a distribution is. [… I]f the distribution is flat […] then kurtosis is negative, whereas if the distribution is peaked, then kurtosis is typically positive. However this general assumption only applies if the distributions are not skewed". Given that spectra are not normally free of skew, this assumption almost never holds in practice.

Thus centre of gravity and standard deviation are expected to provide information on place of articulation that will not be enhanced by the addition of skewness and kurtosis. This can be seen in the results of Forrest et al. (1988: 119) for /p t k/: whenever skewness is low, mean (i.e. centre of gravity) tends to be high (in the diagrams this can be observed in the fact that the scatter of the datapoints in the skewness-CoG space runs along a relatively straight diagonal path). The spectral moments have been used on fricatives (e.g. Koenig et al., 2013) and, less frequently, on plosives (e.g. Forrest et al., 1988). Suchato (2004) only examined centre of gravity. Forrest et al. (1988: 118) intended to examine all four moments but ended up excluding standard deviation as it yielded very similar results to kurtosis, such that it did not add to the discriminability of the plosives.

2.3.1.9 Front Ends in Automatic Speech Recognition

The acoustic attributes that we have looked at up until now have been used in phonetic science but little used in automatic speech recognition (ASR).

The two most widely-used ASR acoustic models are Mel-frequency cepstral coefficients (MFCC, Davis and Mermelstein, 1980) and perceptual linear prediction (PLP,

Hermansky, 1990). Both of these begin with an FFT followed by a warping of the frequency axis in a manner that is somewhat similar to how the frequency axis is warped by the cochlea (though see Lyon (2017: 80-81) for a more detailed critique): MFCC use the Mel scale (Stevens et al., 1937), while PLP uses the Bark scale (Hermansky, 1990: 1739; Schroeder, 1977). Both the Mel and Bark scales are approximately logarithmic above 1,000 Hz and more linear below this frequency.

Both MFCC and PLP reduce the number of data points in the input spectral envelope but in different ways. MFCC achieves this by use of a Discrete Cosine Transform (DCT) whereas PLP uses linear prediction.

1. Discrete Cosine Transform (DCT)

A DCT is similar to standard Fourier analysis, the difference being that the analysis is performed on the frequency spectrum rather than on the waveform. This means that the features yielded by the DCT, unlike those of Fourier analysis, do not correspond to frequency; rather, they correspond to quefrency. Different quefrencies represent differing degrees of spectral detail: a low quefrency represents coarse spectral details such as overall amplitude and spectral tilt whereas higher quefrencies represent progressively finer spectral details.

The DCT consists of a set of basis functions. Each basis function is a sinusoid in cosine phase whose number of half-cycles across the spectrum is a whole number (Lyon, 2017: 76). The lowest 12 coefficients of the DCT are widely used in ASR as they are sufficient for representing the acoustic features of speech (excepting pitch) in sufficient detail. (Features describing the difference between the features of neighbouring frames, as well as the rate of change between two pairs of neighbouring frames, are also often used. These time derivatives are termed delta and delta-delta features respectively.)

The logic of the DCT is perhaps best illustrated by comparing the shape of its basis functions to a principal component analysis (PCA) of the spectrum.

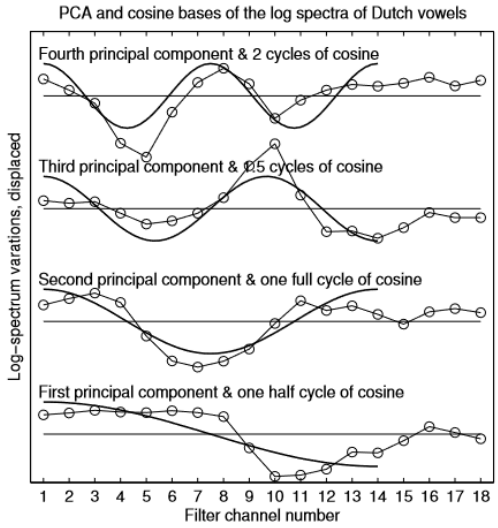


Figure 2.28: Comparison of PCA and DCT.

The diagram shows the first four principal components yielded by a principal component analysis (PCA, shown by the lines with circles along them) and the lowest four discrete cosine transform (DCT) basis functions (thicker lines). Note the similarity of their shapes. PCA from Plomp et al. (1967), diagram from Lyon (2017: 77).

A PCA analysis of a spectrum consists of projecting the spectrum onto a set of basis functions, each of which captures as much of the variance in the data as possible (Lyon 2017: 76) within the constraint that each component must be orthogonal to the components that preceded it. A dot product of the spectrum with one of the above DCT basis functions or PCA components yields a coefficient that quantifies the extent to which that component is present in the spectrum's shape (Lyon, 2017: 77).

PCA is a powerful tool for yielding features for maximizing the amount of variance captured from a spectral envelope. Given that the DCT consists of basis functions that are similar in shape to PCA, we should expect it to perform a similar variance-capturing role to the PCA. Indeed, Shanmugam (1975) (cited by Lyon, 2017) outlined theoretical reasons why the DCT should perform similarly to a PCA. The disadvantage of PCA is that the precise shape of the components that it generates will depend on the particular dataset. In contrast, the shape of the DCT coefficients is set independently of the data and thus can be compared across datasets.

Because of this difference in replicability and interpretability, PCA will not be employed in the present study but the similar approach of using DCT coefficients will. The performance of these coefficients will be compared to that of the traditional acoustic-phonetic features reviewed above. This comparison constitutes Aim 4 of the present thesis. The comparison seems necessary given that previous studies of the burst in phonetic science have not used DCT coefficients as a benchmark against which to evaluate the performance of the various traditional burst attributes we have seen. Without this comparison, it is difficult to know

how the performance of these traditional burst attributes measures up to the alternatives that have been widely used in ASR for decades. The results of this comparison will be presented in Section 7.3.

2. Perceptual Linear Prediction (PLP)

Perceptual linear prediction (PLP) is similar to linear prediction except that prior to the linear prediction the spectrum is warped by a series of procedures intended to roughly mimic the representation of frequency and amplitude in the auditory periphery. Before this, the first step, as in the MFCC approach, is to perform a fast Fourier Transform on the waveform. A window length of 20 ms plus 5.6 ms of zero-padding is typical (Hermansky, 1990: 1739). Following this, the frequency scale is warped using Schroeder's (1977) formula for converting frequency from Hertz to Bark. After this the spectral components are boosted in amplitude in a manner similar to the transfer function of sound through the outer and middle ear; this is termed 'equal-loudness pre-emphasis' (Robinson and Dadson, 1956), which boosts the amplitudes of components around 3 kHz relative to those above or below this frequency. The next stage is to scale amplitude in a manner that approximates what has been found in psychophysical experiments regarding the amount of amplitude increase necessary to produce a doubling in perceived loudness, and was first formalized by Stevens (1957) as the sone scale. (PLP uses a mathematically simplified approximation to Stevens' scale.) For components above 40 dB SPL, every 10-dB increase corresponds to a doubling of loudness. Thus 40 dB corresponds to 1 sone, 50 dB to 2 sones, and 60 dB to 4 sones. For components below 40 dB, the relationship between loudness and decibels is less linear: for example, 30 dB corresponds to 0.44 sones, 20 dB to 0.14 sones, and 10 dB to 0.02 sones (see Section 4.5.2 and Fastl and Zwicker (2007: 205-215) for a more detailed discussion).

After the application of these steps, the spectrum is approximated by the envelope of an all-pole model using the autocorrelation method of all-pole spectral modelling (Hermansky, 1990: 1740). The choice of order in the model determines how much spectral detail is to be preserved. Conventional linear prediction typically uses 12^{th} or 14^{th} order, which yields 12 to 14 coefficients per frame. Hermansky found this worked well in speaker-dependent speech recognition but gave inferior results on speaker-independent recognition. In the latter task, a lower order (5^{th} or 6^{th} order) was found to give the best results. He interpreted this as showing that gross spectral shape is sufficient for speech recognition.

3. Summary

In both PLP and MFCCs the output is a relatively small number of coefficients that describe the spectral shape of a speech sound concisely with minimal loss of speech information. There have been many studies comparing the performance of the two approaches, e.g. Psutka et al. (2001) for continuous-speech recognition. These studies have generally found that the performance of the systems is similar provided that both have first had their settings optimized (e.g. number of model coefficients).

Jannedy and Weirich (2017), using the contrast between German /ʃ/ and /ç/ as their testing ground, compared the performance of the four spectral moments to the Discrete Cosine Transform (DCT) and found that the moments failed to reveal the difference between the two fricatives that was apparent from visually inspecting spectrograms, whereas the DCT coefficients did quantify the difference.

For analysing the burst spectrum, the present study utilizes the Bark-phon (and Bark-sone) spectral representations that are available in the Praat software package (Boersma and Weenink, 2014), which warps the burst spectrum in a manner broadly similar to PLP. The accuracy of the acoustic attributes extracted from this spectrum will be compared with the accuracy of the same acoustic attributes extracted from two other spectra: (1) a linear 'Hz-dB' spectrum (widely used in phonetics, e.g. Blumstein and Stevens (1979); Suchato (2004)) and (2) a 'Bark-phon' spectrum, which is the same as the Bark-sone spectrum except that it does not contain the conversion from phon units to sone units that results in high-amplitude spectral components being made more prominent relative to low-amplitude components, as shown in the following figure:

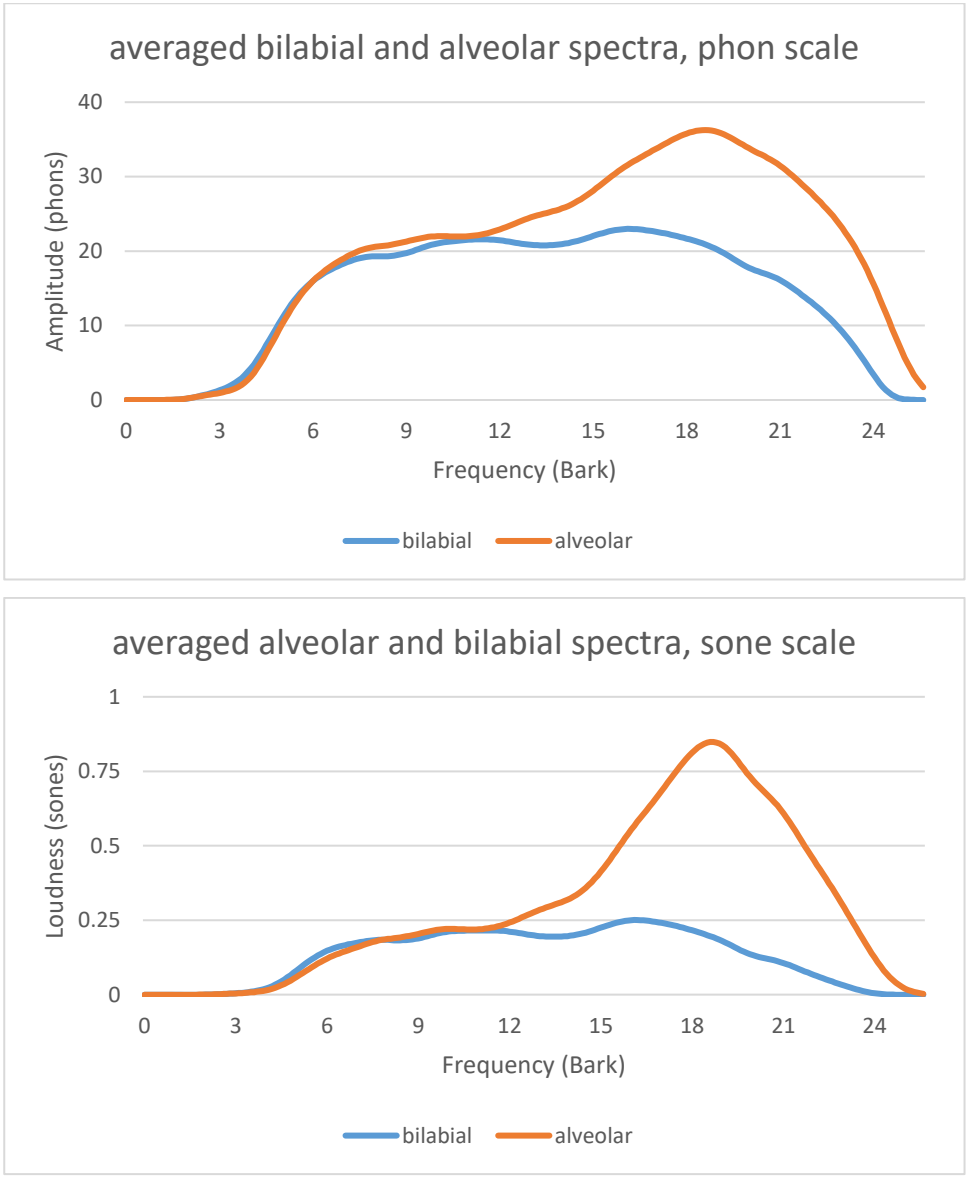


Figure 2.29: Comparison of the mean bilabial and alveolar burst spectra on the Bark-phon and Bark-sone scales in the present study's dataset.

The sone scale makes the alveolar's high-frequency peak more prominent and makes the bilabial's entire spectral envelope less prominent relative the alveolar one, due to the sone spectrum shrinking the amplitude of low-amplitude components relative to high-amplitude components (see Section 4.5.2 for details). Bilabial N = 1,490; alveolar N = 2,429.

The overall aim of this three-way spectral comparison, which constitutes Aim 3 of the present thesis, is to investigate whether different spectral representations affect the accuracy of burst attributes at distinguishing plosive place of articulation. For further details on the three spectral representations, see Section 4.5.2 of the methodology chapter and for the results of the comparison itself, see Sections 6.4.1 and 6.5.

2.3.2 Plosives in Syllable-Final Position

The plosives focused on thus far have primarily come from syllable-initial position. Syllable-final plosives can involve a number of complications that are not found in syllable-initial position. One of these was encountered in Section 2.2: syllable-final formant transitions tend to have steeper locus-equation slopes than syllable-initial transitions (Al-Tamimi, 2004), which theoretically could result in their place of articulation being more difficult to identify (a hypothesis which is investigated in 5.4.5). The other complication with syllable-final plosives is that, depending on the following context, the release burst can be absent. The /g/ in the word *bog*, for example, may have an audible burst if uttered in isolation or before a vowel (or before certain consonants such as /h l r j w/) but may not if uttered before another plosive, such as in the phrase *bog down*. An important question regarding unreleased plosives is to what extent listeners can correctly identify their place of articulation. This question is one way of establishing the perceptual importance of VC formant transitions. Another way of establishing it is this: if a final plosive had the burst of, say, a bilabial but the transitions of a velar, which place of articulation would listeners hear?

Malécot (1958) set out to answer these questions. He played to 50 listeners natural CV and VC syllables consisting of /b d g/ and /ɛ ɔ/, which yielded six syllables. He verified that for transition + burst, listeners' accuracy was very close to 100%. In another set of stimuli the same VC syllables were used but this time the speaker had not released the final plosive, i.e. the listener had no burst to rely on to identify place, only the transitions. In the syllables /ɛb ɛg/ identification remained 100% correct (in /ɛd/ identification dipped to 96%). In the syllables /ɔb ɔd ɔg/ identification dropped to 92, 64, and 88% respectively. Identification of place, then, was patchier than when the burst was available but nevertheless relatively high. The fourth set of stimuli were similar except that the final burst was deleted from the recording (rather than the recording being of a genuinely unreleased plosive). The identification rates were nearly identical to the genuinely unreleased stops.

It should be borne in mind, however, that the listeners were only allowed to respond with 'b', 'd', or 'g'. That is, they could not respond with, say, 'unsure' or 'no consonant' or 'v'. This stipulation likely aided their perception in ways that a genuinely open task (such as shall be seen in Miller and Nicely, 1955; Lobdell and Allen, 2006; Li et al., 2010 to be presented shortly) would not. If the listeners *had* been given a free choice, there might have been cases where they did not detect the unreleased consonant at all.

In the second part of the study the participants listened to stimuli in which the transitions of one place of articulation were paired with the burst of another. When listeners heard [ɛb] with the transitions of [b] but the burst of [d], 80% of them identified it as /d/, and when they

heard the [ɛb] transition paired with the burst of [g], most listeners were again swayed by the burst, 84% of them (of the remaining 16%, just 12% reported hearing /b/). When the syllable [ɔb] was paired with a burst of another place, listeners were again swayed by the burst, but this time even more so: for the [d] burst, 92% reported /d/, and for the [g] burst the percentage reporting /g/ was 100%.

When a burst from [b] or [g] was paired with [ɛd ɔd] transitions, the trend was again similar: the burst overruled the transitions for most listeners (on average, 96% of them). Indeed, if the burst was [g] this was true of all the listeners. When the burst was [b], the percentage of listeners reporting /b/ was a shade less (92%). But of the 8% of listeners *not* reporting /b/, all of them reported hearing /g/, not /d/. Thus the [d] transitions caused none of the listeners to perceive /d/ when the burst was of a different place of articulation. These results strongly suggest that the perceptual importance of plosives' VC transitions is at best weak relative to the burst.

The third batch of hybrid stimuli involved [ɛg ɔg] transitions paired with the burst of [b] or [d]. Although the velars' transitions were able to sway more listeners than the transitions of the other two places of articulation, the overall pattern repeated itself: most listeners perceived place of articulation based on the burst, not the transitions: [ɛg] + [b], 72% of listeners; [ɔg] + [b], 70%; [ɛg] + [d], 76%; [ɔg] + [d], 98%.

Malécot repeated the exercise with the voiceless series /p t k/. The findings were in broad agreement with those from the voiced series.

The overarching message of the study is that in the great majority of cases, listeners' perception of place of articulation is based on the burst whenever there is a contradiction between the place of articulation of the burst and the VC transitions. This indicates that the burst is the dominant cue. Nevertheless, the transitions do seem to matter in those cases where the listener has no burst to rely on: although listeners did not identify the place of articulation consistently from burstless [ɛb ɛd ɔb ɔd], their performance was nevertheless well above chance.

One acoustic prediction that can be derived from this perceptual study is this: the classification accuracy of place of articulation using VC formant transitions will probably be less than that obtained using CV formant transitions. Section 5.4.5 tests this prediction.

2.3.3 Burstless Transitions and Transitionless Bursts

Much research in the 1970s (Cole and Scott, 1974a; b; Dorman et al., 1977; Stevens and Blumstein, 1978) tried to investigate to what extent listeners rely on the burst as a cue relative to the formant transitions in identifying plosives' place of articulation. They did this by taking

natural CV syllables, removing their bursts or transitions, and playing the resulting syllables to listeners and asking them to identify the consonant.

Cole and Scott (1974 b) suggested that the burst is sufficient to cue place of articulation in /b d g/ on its own. They played to listeners the CV syllables /bi bu di du gi gu/ but with the formant transitions artificially removed (leaving the burst and the vowel steady state). They found that listeners were able to correctly identify such syllables. However, as Dorman et al. (1977) point out, the experiment involved a forced choice, i.e. the listeners could only give 'b', 'd', or 'g' as an answer. Dorman et al. also note that a stimulus in which a [b]-burst is paired with a steady-state [ɪ] sounds like a click followed by the vowel [ɪ]. They write (1977: 122):

"[…] if only /b,d,g/ are permitted as responses, then a subject may well feel that, since the click does not sound like a high-frequency alveolar burst and is not affricated like a velar burst, (s)he should respond /b/. A correct /b/ response would then be made to a signal that does not sound like /b/ […]"

Thus Cole and Scott's conclusions regarding their findings for transitionless stimuli are open to debate.

It is important to think deeply about why transitionless stimuli are troublesome for listeners. When the transitions are removed, the steady state of the vowel becomes the vowel onset. F1 normally rises to some degree from the syllable onset to the syllable steady point. But when the transitions are removed, this F1 rise is lost.

What might this mean for the listener? Take the syllable [ga]. When its transitions are removed, F1 begins at a high frequency. But in real life F1 cannot begin at this high frequency, because that would involve the tongue instantly moving from being in contact with the roof of the mouth to being in position for the vowel. The listener, when faced with such a strange stimulus, might interpret the [g] burst as being an extraneous sound that does not belong to the syllable.

This is why listeners' perception of transitionless stimuli is so difficult to interpret. At first blush the poor identification of consonants in transitionless stimuli might seem to show that the formant transitions are the main cue for plosive place, given that their absence results in listeners failing to identify the consonants consistently. But because transitionless stimuli not only entail removing the transitions but also changing the timing in the syllable, this could lead listeners to misinterpret the burst as a non-speech sound.

A solution to this problem would be to remove the F2 and/or F3 transitions using bandstop-filtering while preserving the F1 transition. This way the temporal structure of the

syllable would not have been changed and so the perceptual contribution of F2 could have been isolated.

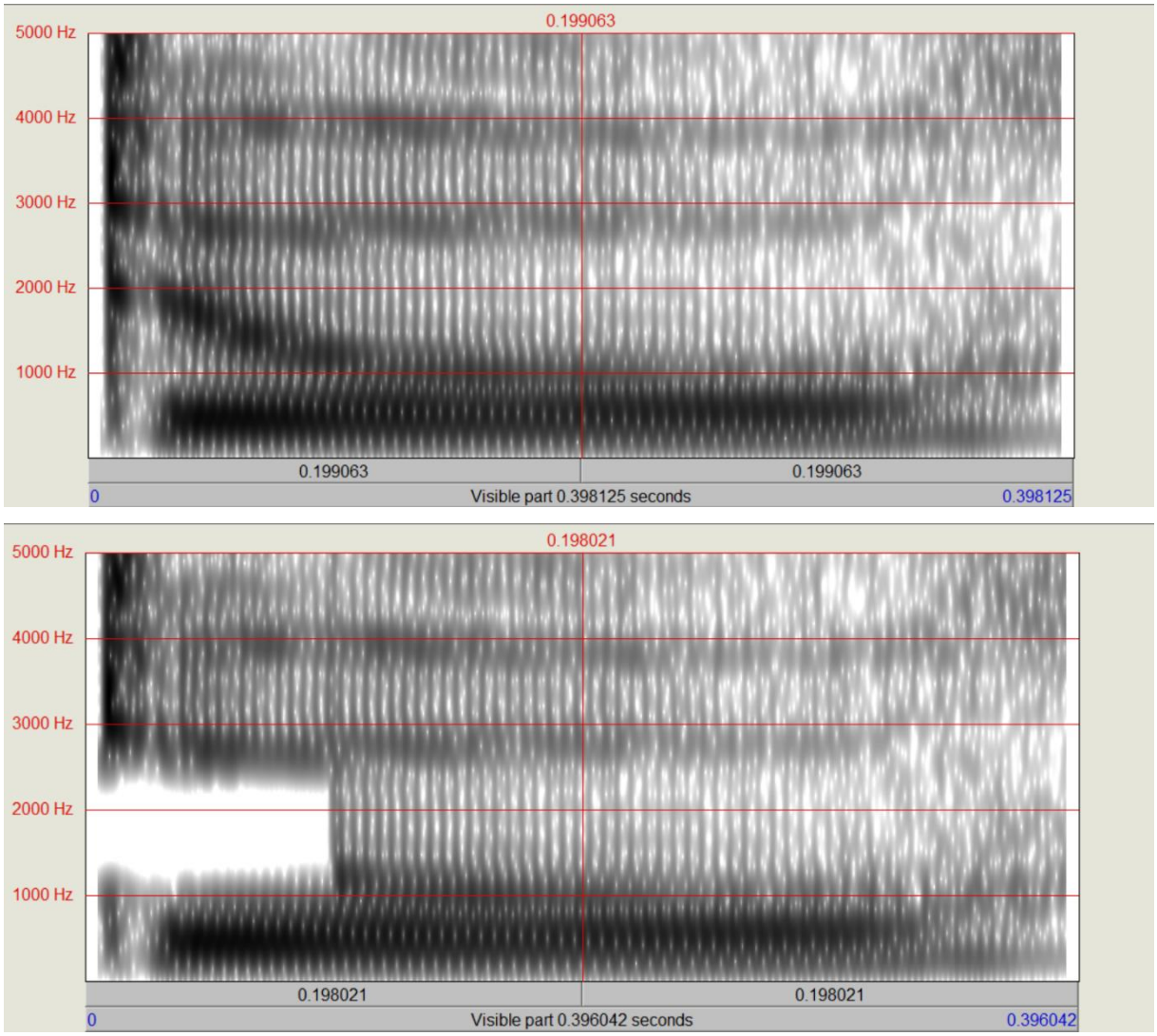


Figure 2.30: Example of the syllable [do] before and after having its F2 transition bandstop-filtered. The filtered region extends from 900 to 2,500 Hz (with Praat's default 100-Hz smoothing). This stimulus preserves the temporal structure of the syllable while allowing the evaluation of the F2 cue. It was not used in 1970s studies, which instead truncated the transitions, violating the temporal structure of the syllable. The two stimuli sound almost identical.

In 2.3.7 a recent study (Cvengros, 2011) will be reviewed that has used stimuli similar to that shown in Figure 2.30. (Another solution is to fill the F2 region with noise, which has been done by Alwan (1992) who found that listeners could correctly distinguish /ba da/ stimuli in which this had been done.) Unfortunately this type of stimulus did not occur to the 1970s researchers. The result is that interpretation of the transitionless stimuli is inconclusive, both for Cole and Scott (1974b) and for Dorman et al. (1977) and Stevens and Blumstein (1978). Fortunately the interpretation of the burstless syllables used in these studies is not as difficult because such stimuli do not consist of changing the timing of acoustic events in the syllable.

Another limitation of Cole and Scott (1974b) is that they tested their hypothesis using only two contexts, /i/ and /u/. Fischer-Jørgensen (1954) examined /Ca/ context and found that listeners were unable to distinguish /b d g/ using transitions alone (which as we saw in 2.2.3 was suggested by Miller and Nicely (1955) for /dɑ gɑ/ and verified perceptually by Li et al. (2010)). As for /Cu/ context Fischer-Jørgensen found that alveolars could not be identified reliably without transitions. This was in spite of the fact that her listeners were only allowed to choose between /b d g/. Thus Cole and Scott's claim that transitionless /b d g/ stimuli could be readily identified was not found by Fischer-Jørgensen.

In a related experiment, Cole and Scott (1974a) removed the bursts from one vowel context and spliced them onto another vowel context. They found that removing the [d] burst in [du] and splicing it onto [di] did not reduce the ability of listeners to correctly identify place of articulation, and likewise with splicing the [d]-burst of [di] onto [du]. This indicates that the coarticulatory influence of the vowel on the burst is not large enough in the case of /d/ to obscure the identity of the phoneme across contexts. They repeated this procedure for [bi] and [bu] and obtained the same result.

For /g/, however, the finding was different: when the burst of [gi] was spliced into [gu], the identification rate plummeted to 21% (of the errors, 90% involved mishearing the place as bilabial – somewhat reminiscent of what Liberman et al. (1952: 509) found for /p k/; see Figure 2.19). When the burst of [gu] was spliced into [gi], perception was again impaired though not to the same extent: 82% accurate. The failure of the burst of front velars to sound velar when spliced onto back velars is presumably related to the greater vowel coarticulation on velar bursts, noted as long ago as Potter et al. (1947) and examined in Section 6.1 of this study.

Dorman et al. (1977) used natural CV stimuli consisting of /b d g/ followed by one of nine vowels. They split the cues for place of articulation into three parts: the burst, the aspiration (which in a voiced stop usually lasts less than 10 ms), and the formant transitions. Various aspects of the original syllable were played to listeners: (a) burst + aspiration + transitions + vowel, i.e. the whole of the original syllable; (b) burst + vowel; (c) burst + transitions + vowel; (d) transitions + vowel; (e) aspiration + transitions + vowel. They found that aspiration aided perception only marginally, which is unsurprising since the aspiration is low in amplitude (Miller and Nicely, 1955: 347) and is short in /b d g/.

In the case of /b/, if listeners were not given the transitions (option (b)) they were normally only able to correctly identify the consonant at the level of chance or worse (though with three exceptions). In contrast, when listeners were not given the burst but *were* given the transitions and vowel (option (d)) they were able to identify /b/ almost as frequently as when

they had the entire syllable: in both cases identification rates were at or above 90%. There were, however, three exceptions: when the following vowel was [ɔ u ɝ], not having the burst reduced correct identification to 84% for [ɔ], 63% for [u], and 61% for [ɝ]. Interestingly it was in exactly these three environments that the performance of the burst-only stimuli was at its best.

What might be the cause of this curious pattern? Dorman et al.'s interpretation is that there is a trading relation between the burst and the transitions: when the transitions give unreliable information, the burst is relied on; when the burst gives unreliable information, the transitions are relied on. A more precise way of putting it is that when a devoiced stop is presented in utterance-initial position without its release burst, there is nothing except the formant transitions left to signal to the listener the existence of the consonant. Consequently the formant transitions must change in frequency for the /b/ to be perceived. The fact that the vowels [ɔ u ɝ] are contexts in which the burstless /b/ transitions are identified less well might be because these are all contexts in which the formant transitions happen not to change in frequency that much. Figure 2.14 (from Öhman (1966: 160)) illustrated this for [bo], and his diagram of [bu] likewise shows the F1, F2, and F3 transitions changing relatively little in frequency. These are context which – like [ɔ u ɝ] – involve vowels with a low or relatively low F2 frequency. All but one of them also involve a close-mid or close vowel, in which the F1 movement appears to be modest or flat.

In contrast, before *front* vowels Öhman's data show much more F2 change. What this means is that even when one can hear no burst, one can still figure out that there must have been a preceding bilabial consonant from the fact that the F2 is moving. But before back vowels, this is harder to do since F2 happens to be steady or close to steady. The perceptual utility of the /b/ burst, then, would be that it helps the listener recognize that a transition pertains to a CV syllable (rather than a V syllable) in cases where the F1 and F2 transitions happen to be flat or close to flat.

In cases where the transitions are *not* flat, removing the transitions leaves the listener with just the burst to rely on. Given that Dorman et al.'s burst + vowel stimuli for such contexts were identified correctly by listeners in less than 25% of cases, this appears to suggest that the burst cannot cue place of articulation (in voiced stops) on its own. As has already noted, however, this is not the only way to interpret this finding. The fact that listeners mostly failed to identify transitionless stimuli does not necessarily show that the F2 transition is what is being used to recognize place of articulation. This is because the transitionless stimuli, unlike the burstless stimuli, involve changing the temporal structure of the syllable: the steady-state part of the vowel becomes the syllable onset. This could confuse the listener into thinking that the burst does not belong to the syllable (a point Dorman et al. themselves made when criticizing

Cole and Scott (1974b), as noted earlier). This is plausible because in real speech F1 has not normally reached its value for the vowel at vowel onset, especially if the vowel is open. Thus an artificial syllable in which the F1 has reached its vowel midpoint frequency at vowel onset might confuse the listener into thinking the burst is an extraneous noise, since in real speech bursts cannot occur next to the vowel's steady part. Thus it is difficult to interpret Dorman et al.'s findings as showing that the /b/ burst is unsuccessful at cuing place in transitionless stimuli.

Broadly similar comments can be made about Dorman et al's findings for /d/ and, to a lesser extent, /g/. The overall finding for the burstless stimuli is that listeners were best able to identify place of articulation in those syllables where the transitions happened not to be flat, especially for /b/ and, to a lesser extent, /d/ (though not for /g/). If the transitions are flat, the burst is also needed, presumably to signal to the listener that the flat transitions pertain to a CV syllable rather than a V syllable. As has been discussed, the interpretation of the findings for the transitionless stimuli is problematic.

Stevens and Blumstein (1978) conducted a perceptual study that was broadly similar to Dorman et al. (1977) except that they used synthesized speech instead of cutting and pasting natural speech. They played to listeners three kinds of artificial /b d g/ stimuli: (1) syllables with burst, transitions, and vowel; (2) syllables with transitions and vowel; (3) syllables with burst and vowel. The identification rate for (1) was 90%; for (2), 81%; and for (3), 18%. These results do not, however, necessarily show that the burst is a less important cue than the transitions, because a burst that is paired with a formant transition whose F1 does not rise may well sound like an extraneous sound that does not belong to the syllable. This point has been made already in the review of Dorman et al. and in Dorman et al.'s criticism of Cole and Scott's stimuli.

In conclusion, although the results of experiments using burstless transitions and transitionless bursts are interesting, their interpretation is difficult. The artificial stimuli used in these 1970s studies were highly unnatural, especially the transitionless bursts since they involved violating the temporal structure of the original syllable by putting the vowel's steady part next to the burst. As a result, it appears that stimuli in which the formant transitions have been removed cannot be interpreted as demonstrating that the burst is an unimportant cue, since putting the burst next to a steady F1 could lead listeners to perceive the burst as not belonging to the syllable.

Future research should instead examine the perception of stimuli in which F2 (and/or F3 and F4) have been filtered but F1 has been preserved (as shown in Figure 2.30). This would remove the F2 cue to place of articulation while preserving the temporal structure of the original

syllable. In 2.3.7 the results of a study (Cvengros, 2011) will be reviewed that used stimuli of this ilk.

2.3.4 Filtered Speech and Speech in Noise

The present study is acoustic in nature. Nevertheless the accuracy of human perception can be regarded as a benchmark against which to judge the performance of acoustic classification studies. The study of filtered speech and speech in noise offers a benchmark on the accuracy of human listeners. It also provides a clearer sense of what information about plosives' place of articulation is most robust to noise. Given the wide variety of burst attributes developed in phonetics (2.3.1), such studies of speech in noise are one kind of knowledge that can be utilized by acoustic studies in deciding what kind of acoustic attributes to develop.

Section 2.3.1 presented Liberman et al.'s pioneering experiments using artificial stimuli. Miller and Nicely (1955), on the other hand, pioneered experiments using natural stimuli. They played nonce CV syllables to four listeners under three conditions: (1) six degrees of background noise, namely signal-to-noise ratios of -18, -12, -6, 0, 6, and 12 dB; (2) filtered speech in which the low-frequency cutoff was held constant at 200 Hz and the upper cutoff was varied from 300, 400, 600, 1,200, 2,500, to 5,000 Hz; and (3) filtered speech in which the upper cutoff was fixed at 5,000 Hz and the lower cutoff was varied from 200, 1,000, 2,000, 2,500, 3,000, to 4,500 Hz. The consonant was one of /p t k b d g f θ s ʃ v ð z ʒ m n/; the vowel was always /ɑ/. This experimental paradigm was better than that of Liberman et al. in that the listeners were not forced to report the stimuli as voiceless plosives. All the consonants had an equal probability of occurrence, which meant that listeners could not use their linguistic knowledge of English to improve their recognition. The study was thus a study of raw acoustic recognition devoid of lexical factors.

The data collected by Miller and Nicely are enormous in scale and have been statistically reanalysed by several subsequent studies (e.g. Johnson, 1967; Shepard, 1972; Carroll and Wish, 1974; Soli et al., 1986; Allen, 2005). Out of all the acoustic conditions examined in the study, the one that comes closest to ideal acoustic conditions is the +12 dB signal-to-noise ratio with the headphones' entire spectrum (200 to 6,500 Hz). The listeners were native speakers of the speech variety and had no hearing impairments. Even under these conditions the correct recognition of place of articulation of consonants was noticeably below 100%, namely 92.3%. In contrast the correct recognition of voicing was much higher, 99.6%, as was the perception of the nasal/non-nasal contrast, 99.9%. Voicing and nasality involve a binary contrast (voiced versus voiceless, nasal versus non-nasal) whereas place of articulation (except in the case of /m n/) involves at minimum a ternary contrast. Thus some of the difference in the accuracy of

listeners at identifying place of articulation can probably be attributed to the larger number of values this feature can assume relative to the other features of voicing and nasality. Nonetheless, the fact that the identification of place under near-optimal listening conditions by native listeners with no known hearing impediments was on average 7.7% in error suggests that the information for place of articulation is not always present in the signal itself. The study used nonce syllables, which removed the listeners' ability to use the linguistic context and forced them to rely on the signal alone. Perhaps in real-life situations, where syllables occur as part of a linguistic context, the 7.7% error from the signal is made up by the listener's linguistic knowledge, such as knowledge of the lexicon syntactic constraints, and semantic plausibility.

We turn now to the more specific matter of how well the listeners identified the place of articulation of plosives. Here are the results in the highest speech-to-noise ratio, 12 dB:

	p	*t*	*k*	*f*	*θ*	*s*	*ʃ*	*b*	*d*	*g*	*v*	*ð*	*z*	*ʒ*
p	240		41	2	1									
t	1	252	1	1						1				
k	18	3	219											
b					1			242			24	12	1	
d									213	22			1	
g					1				33	203		3		

Table 2.2: Confusion matrices of /p t k/ and /b d g/ with a speech-to-noise ratio of +12 dB and all of the headphones' spectrum (200 to 6,500 Hz) played to listeners.
The letters along the vertical axis indicate what the speaker produced, while the letters along the horizontal axis indicate what listeners reported hearing. Adapted from Miller and Nicely (1955: 342).

We see that, out of the 281 instances of /p/ that were correctly recognized as being a plosive, 15% were misperceived as /k/, with none being misperceived as /p/. For /k/, 7% of tokens were misperceived as /p/. In contrast /t/ was very seldom misclassified as /p/ or /k/. Thus among the voiceless plosives the misclassifications mostly involve bilabials being confused with velars and vice versa. Of the voiceless plosives that were correctly identified as plosives, 91.7% were correctly identified in terms of their place of articulation. This figure should be borne in mind when examining the results of the present study in Chapter 7.

For the voiced series, we see that there is a considerable trend for /b/ to be misperceived as a fricative (14%). This is perhaps due to the fact that the burst of /b/ can sometimes be so low in amplitude that it seems to be inaudible; Li et al. (2010) found that listeners were unanimous in correctly identifying /bɑ/ as /bɑ/ (and not /vɑ/) when the token contained a definite burst, which was often not the case for their tokens. Given that release bursts are presumably peculiar to plosives, if one is not detected it might increase the chance that the listener perceives a fricative (especially since labial and dental friction is low-intensity and hence vulnerable to being obscured in real-life conditions). This suggestion is lent plausibility

by the fact that /d/ and /g/ (whose bursts tend to have greater intensity than that of /b/) are misperceived as fricatives in just 0.4% and 1.7% of cases respectively.

Instead, the misperceptions of /d/ and /g/ involve each other: 9% of /d/ are misperceived as /g/ and 14% of /g/ are misperceived as /d/. In Section 2.2.3 we noted that the formant transitions of /d/ and /g/ before /ɑ/ are indeed very similar and have been shown to be confusable if the burst is partly or totally removed (Li et al., 2010). Such misperceptions between alveolar and velar place occur far less frequently in /t/ and /k/ (Figure 2.30). Given that the transitions occur during the aspiration in /t k/ and are consequently much harder to hear (Miller and Nicely, 1955: 347) than in /d g/, it might seem plausible that the ambiguous formant transitions are why /d g/ are more likely to be confused with each other than /t k/.

However, there is a second factor that could also be playing a role. Zue (1976: 112, 115) found that the mean frequency of the burst peak in /t/ was around 3,660 Hz, whereas for /d/ it was around 3,300 Hz. He found this difference in all the phonetic contexts he investigated. In contrast, he found the burst peaks for /k/ and /g/ to be almost identical in frequency, averaging 1,970 and 1,940 Hz respectively (pp. 119-123). Thus the difference in frequency between the burst peaks of /d/ and /g/ tends to be smaller than that between /t/ and /k/. This would lead one to expect /d/ and /g/ to be more frequently confused than /t/ and /k/, which is the case.

Of the voiced plosives that were correctly identified as plosives, 92.3% were correctly identified by the listeners. The combined figure for the voiced and voiceless series was 92.0%. This gives us a helpful baseline for classification by native listeners without hearing impairments in near-optimal acoustic conditions in which the linguistic context has been removed. One might have assumed that the figure would be 100%, but it appears that the acoustic information for identifying phonemes is not always encoded unambiguously, even when those phonemes are uttered in controlled conditions in which the surrounding phonetic context is held constant.

The study tested consonants in CV context only. This context is universal among the world's languages (Greenberg, 1978). If the consonants had been tested in other phonetic contexts, such as VCC or VCP contexts (P = pause), the error rate at recognizing place of articulation would likely have been far larger than the 8.0% found for CV context, since such contexts often involve the loss of many of the features for identifying place of articulation, such as the loss of the formant transitions of /k/ in *task* due to the preceding /s/ and the loss of the /k/ release burst in *act* due to the presence of the following plosive (the /t/), as well as the loss of the burst in connected phrases, e.g. the /g/-burst in *bog down*.

Even when a plosive is in stressed syllable-initial position, if it is followed by another consonant the acoustic change to the plosive can be quite substantial: Zue (1976: 126) found

that the amplitude of the burst was reduced in /k g/ in plosive + sonorant sequences relative to plosive + vowel sequences by 8.2 and 8.7 dB respectively. Reductions in amplitude were also found for /t d/ but were smaller (2.3 and 4.0 dB respectively). Presumably such reductions have the effect of shrinking the distinction in acoustic space between these consonants and /p b/ (i.e. when the amplitude of /k g t d/ is reduced, they are closer in acoustic space to the low-amplitude /p b/).

Although the present study is concerned with acoustics rather than linguistic knowledge, the underdetermining of speech features by the signal when the linguistic context is missing is a noteworthy finding and should be borne in mind. Miller and Nicely's study, in its removal of the linguistic context, highlights this underdetermination by the signal, which is apparent for place of articulation even under near-optimal listening conditions. The idea that speech sounds can be recognized with 100% accuracy without having a linguistic context turns out to be naïve.

A similar point has been made by Huckvale (1996: 5) in the context of ASR:

"A phone recognition rate of 70% (without higher-level constraints) is not a disaster because a 70% correct transcription is not generated and then filtered by higher levels. Rather the evidence about possible transcriptions is utilised directly in the recognition of [the] utterance. By postponing phonetic transcription until after the word sequence is identified, poor segment-level performance can be converted to good word-level performance. […] But notice that this postponement of decisions means that the transcription of the utterance comes *after* the words have been recognised, and hence looks rather good." [Emphasis in original.]

That is, the key to good speech recognition in real life (in which, unlike in experimental conditions, there is a linguistic context) is to delay the decision of identifying the phonemic units in the signal until a longer (word-scale) context is processed. This allows the incorporation of higher-level information (such as lexical and syntactic knowledge) into the identification of the phoneme stream rather than using the acoustic information on its own.

The take-home message from all of this for acoustic classification studies such as the present one is that we should not expect real-life speech to acoustically encode phonemes with 100% fidelity, nor is this necessary if the acoustic model is paired with a good language model.

2.3.5 The Three-Dimensional Deep Search

Although the results of Miller and Nicely's study have been widely analysed and re-analysed, one significant drawback is that the stimuli used are not available to researchers (Alwan, 1992: 25). Lovitt and Allen (2006) replicated Miller and Nicely with a larger sample of listeners and with a tweaked methodology, e.g. they gave listeners the option of reporting 'no consonant

heard', which was warranted for the stimuli in the lowest speech-to-noise ratios in which the noise can be sufficiently powerful to mask the consonant cues entirely. Most of Lovitt and Allen's findings are in accord with the original study, the main exception being that they found place of articulation to be somewhat *more* robust to increasing noise and voicing somewhat *less* robust to increasing noise than what Miller and Nicely report.

Several research groups have continued the study of plosive place of articulation in noise in recent years. Alwan et al. (2011) investigated the /pa ta/ and /ba da/ distinctions with increasing levels of speech-weighted noise. They found that the greater the noise, the more listeners seemed to rely on formant frequencies rather than the release burst. However it is important to note that their methodology for ascertaining this was a correlational analysis in which they took acoustic measurements from a (non-auditorily-oriented) spectrogram and sought to correlate them with the patterns of listener responses and errors. Section 2.3.7 will present the results of a different method of examining perceptual prominence that manipulates the burst and formant correlates directly.

Hedrick and Younger (2007) examined the perception of /pa ta/ in speech-weighted noise versus reverberation. Their results suggested that listeners utilized formant transition information less in noise relative to reverberation. In the present review there will be no further discussion of reverberation; the focus will be on speech-weighted noise.

Allen and co-workers (Allen, 2005; Lobdell and Allen, 2006; Régnier and Allen, 2008; Phatak, et al., 2008; Li et al., 2010; Cvengros, 2011; Kapoor and Allen, 2012) have extended this investigation of speech features with an approach called the three-dimensional deep search (3DDS). This approach was developed as an alternative to the synthesized speech studies which we have reviewed (e.g. Liberman et al., 1952; Stevens and Blumstein, 1978). The authors (Li et al. 2010: 2599-2600) criticize the use of artificial stimuli on the following grounds:

(1) To generate convincing artificial stimuli, prior knowledge of the cues being sought is required. Given that knowledge of the cues is limited, the result has been synthesized speech that is often of poor intelligibility, which itself suggests that the synthesized speech encodes the speech cues unreliably. As Liberman (1996: 12) remarked of the stimuli he and his coworkers used in Cooper et al. (1952):

> "Of all the synthetic patterns ever used, these burst-plus-steady-state-vowel 'syllables' were undoubtedly the farthest from readily recognizable speech. Indeed, they so grossly offended Pierre [Delattre]'s phonetic sensibilities that he agreed only reluctantly to join Frank [Cooper] and me as a subject in the experiment."

(2) Natural speech varies from one individual and instance to another, e.g. one speaker's [b] may be confusable with [v] whereas another speaker's [b] may be confusable with [g]. The best current artificial speech synthesis is not advanced enough to replicate such everyday, token-by-token acoustic differences.

They propose (2010: 2600) that speech cues be instead studied by starting with natural (i.e. human-generated) speech and performing various modifications to it:

(1) Adding noise of varying type and degree (to determine which frequency regions are the hardest to "flood" with background noise, i.e. the most salient under real-life listening conditions);

(2) Truncating speech from the onset onwards (to see where a cue is located in the time domain); the truncated portions are 5, 10, or 20 ms long depending on the phoneme;

(3) Highpass- and lowpass-filtering the speech with various cut-off frequencies (to see whether the loss of energy from a certain frequency region leads to a different place of articulation being perceived).

Parts (1) and (3) were also undertaken by Miller and Nicely (1955). Figure 2.31 illustrates the technique:

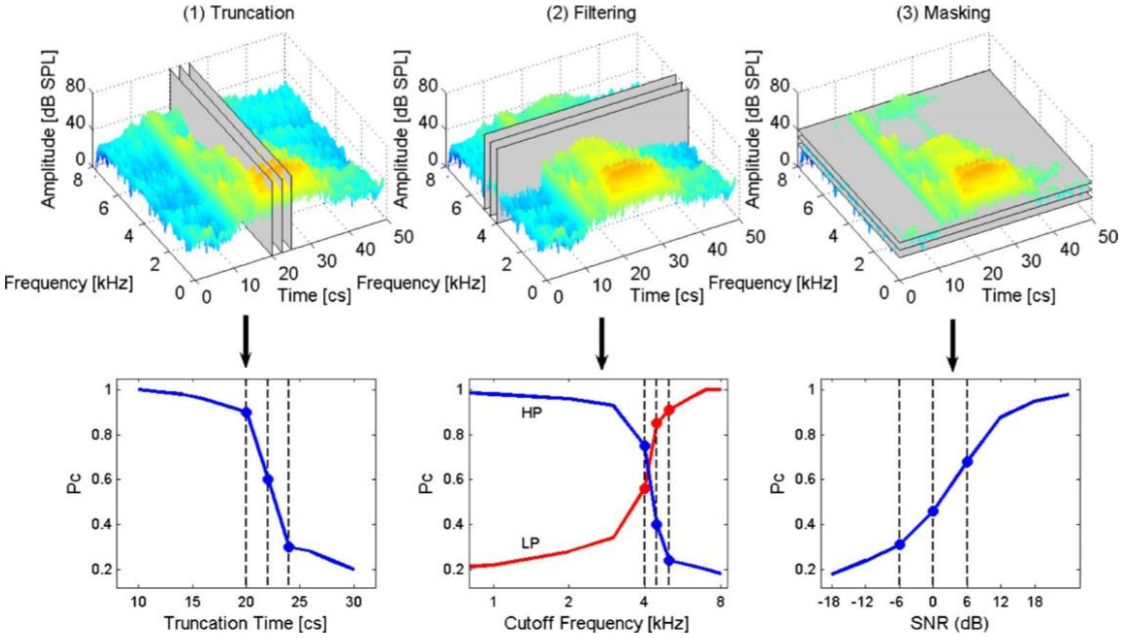


Figure 2.31: Diagrams illustrating the three-dimensional deep search (3DDS).

Truncation involves removing a gradually increasing amount of the syllable onset and asking listeners to report what consonant they hear. In the bottom row is a typical example of how listeners' accuracy decreases as this truncation is increased. Filtering involves removing spectral components from a specified frequency region, whether it be highpass filtering (HP) or lowpass (LP). The plots in the bottom row show how listeners' accuracy at identifying the consonant changes as a result of this, which for /t/ indicate that the region above 4,000 Hz seems to be particularly important to listeners. Masking involves playing the same syllable in increasing degrees of speech-weighted noise. From Li et al. (2010: 2601).

The research team has developed a novel tool, called the AI-gram (Lobdell and Allen, 2006) for displaying and modifying speech. This is a kind of auditorily-inspired spectrogram with an added trick: when speech is placed in noise, the AI-gram displays only those spectral components of the speech that are greater in intensity than the noise, displaying all other spectral components (including the noise) as white, as in the following diagram:

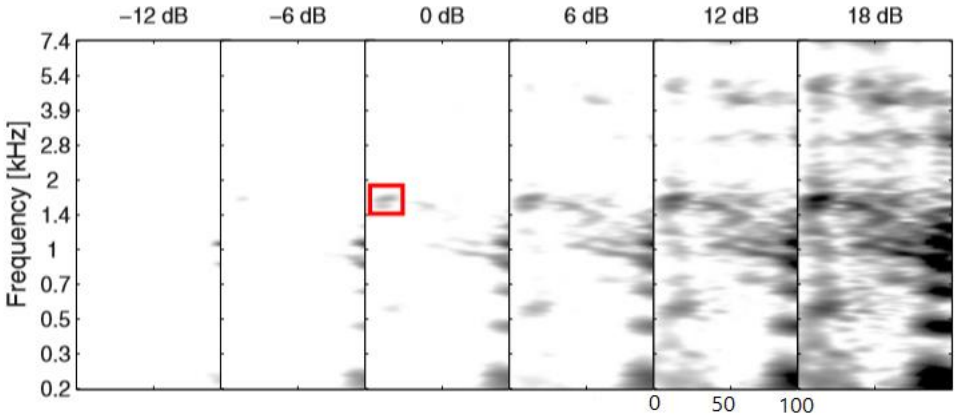


Figure 2.32: Diagrams illustrating the AI-gram display.
Each image shows the first 100 ms of the syllable /kɑ/ in decreasing speech-to-noise ratio (SNR). The AI-gram is designed to only display those spectral components in the speech which are greater in intensity than the noise. This is why in the left-most image, the AI-gram is almost white: the -12 dB SNR means that the noise is so intense that it has almost totally drowned out the speech components, whereas in the rightmost image, showing a +18 dB SNR, the speech is clearly visible. Adapted from Li et al. (2010: 2606).

The team have used this tool to investigate the perceptual cues to plosive place of articulation (among other features) in detail. The overall aim is to identify the cues that signal a particular consonant contrast; such cues they term acoustic 'events' (Régnier and Allen, 2008; Li et al., 2010). An event is defined according to its location in time (via truncation), frequency (via filtering), and intensity (via masking).

Li et al. found that when the /tɑ/ burst was entirely truncated, listeners reported hearing /pɑ/ for five of the six /tɑ/s tested (the one /t/ that was perceived as /k/ had an unusually short VOT of ca. 20 ms rather than the 50-70 ms typical of the other /t/ tokens). Similarly for /kɑ/, when the burst is truncated, the result is again that most listeners report hearing /pɑ/. (This switch to hearing /pɑ/ occurs for four out of the six tokens played to the listeners; the two remaining bursts are mostly heard as a faint /kɑ/.)

The fact that nine out of twelve /tɑ kɑ/ tokens flip from being perceived as /tɑ kɑ/ to /pɑ/ once the burst is removed is striking. Listeners have the formant frequencies appropriate to /tɑ kɑ/ in the aspiration, and yet they still in most stimuli perceive /pɑ/ instead. This seems to challenge Miller and Nicely's (1955: 347) suggestion that the formant transitions in the

aspiration are necessary to perceive the difference between /kɑ/ and /pɑ/: if this were true, one would not expect the correct identification of /kɑ/ to evaporate with the removal of the release burst, as occurred in four of the six tokens examined.

It is not entirely clear why the loss of the /t ~ k/ burst should result in a perception of /p/, but one hypothesis is that aspiration shares in common with a /p/ burst the abrupt appearance of low-amplitude aperiodic components at a wide range of frequencies. Another possibility is that the /p/ burst is so non-prominent that it is easily lost in noise and hence listeners rely on it less than /t k/ bursts. This latter hypothesis has some evidence behind it, as we shall see shortly.

A noteworthy detail in one of the truncation experiments is that correct identification of the syllable /kɑ/ remained at 100% even when the first 20 ms of the release burst were truncated. But when the entire 30 ms of the burst were truncated, there was a sudden plummet of correct identification from 100% to 0%. This seems to undermine the idea that the burst's duration is particularly important for identifying the plosive's place: if this were true, one would expect the correct identification of /kɑ/ to have decreased as soon as the burst was shortened. In fact, even when the burst had been shortened by two thirds from 30 to 10 ms the identification of /kɑ/ remained at 100%. However, this does not show that *subtler* differences in burst duration (say, 5 versus 10 ms) might be playing a role in the distinction. That is, it is possible that if a /k/-burst were shortened to, say, 5 ms that it would sound too click-like to be perceived as a /k/ (being presumably perceived as /p/ instead).

As regards the highpass and lowpass filtering, for /kɑ/ it was found that if everything above 1,400 Hz was removed (the frequency region that contains the burst peak), the burst no longer sounded like /kɑ/, but some other consonant (usually /pɑ/). A similar effect occurred when everything *below* 1,400 Hz was removed (usually /tɑ/).

The results for /pɑ/ were different from /tɑ kɑ/. First of all the six /pɑ/ tokens studied differed somewhat from each other: one had a burst with a shape like a broadband click (i.e. energy dispersed at roughly equal amplitude all the way from 300 to 7000 Hz) and this shape remained the same as more and more (speech-weighted) noise was added. In contrast the other five /pɑ/ tokens only appeared to be shaped like a broadband click when the noise was low-level; once the noise was increased only the components at low frequencies (below 1 kHz) remained visible on the AI-gram.

In the truncation experiment, a second difference between /pɑ/ and /tɑ kɑ/ emerged: the perception of correct place of articulation was not lost as soon as the burst was truncated; instead most listeners continued to report hearing /pɑ/. This may be due to the fact that the release burst

of /pɑ/ was lower in intensity than the aspiration that followed it (see Figure 2.33), whereas the bursts of /kɑ/ and /tɑ/ were more intense than the aspiration:

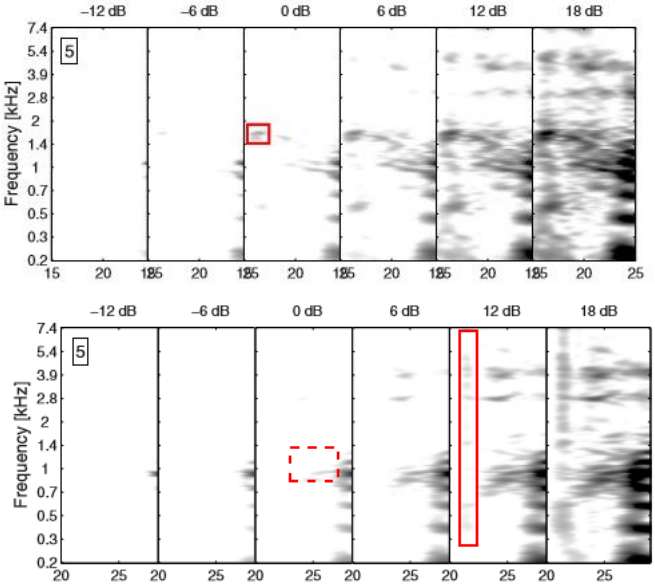


Figure 2.33: AI-grams illustrating the change in the spectrum of /kɑ/ and /pɑ/ with different levels of noise. The /p/ burst (bottom row) disappears into the noise floor at a speech-to-noise ratio of 6 dB, whereas the aspiration does not disappear until 0 dB SNR. In contrast, in the /k/ burst (top row) it is the aspiration rather than the burst peak that is the first to be swallowed by the noise. From Li et al. (2010: 2606).

In the case of /pɑ/, the fact that the aspiration required louder noise than the burst to be masked means that it is the cue that is most robust to noise, whereas in /kɑ/ and /tɑ/ it is the burst that is more resistant to noise than the aspiration. This may explain why listeners continue to hear /pɑ/ after the burst has been lost, both when the experimental technique was noise-masking and truncation. The generalization seems to be this: listeners depend on whatever is most robust to noise.

When the noise eventually became powerful enough to submerge the /pɑ/'s aspiration (or if both the burst and the aspiration are truncated), listeners reported hearing /ɑ/. This is another respect in which /pɑ/ differs from /tɑ kɑ/: when the burst is lost from the latter – whether by noise masking or truncation – listened often report hearing /pɑ/, whereas when the burst of /pɑ/ is lost, listeners do not report hearing /tɑ kɑ/.

Regarding /dɑ/, Li et al. found that if just 5 ms of the burst is truncated – even though the formant transitions are still present – a sizable minority of listeners (around 30%) fail to hear it as /dɑ/, instead reporting /ɑ/. This finding is interesting since we have seen that earlier studies such as Liberman et al. (1954), Malécot (1958), and Cole and Scott (1974b) forced listeners to choose from /b d g/ as responses to their artificial stimuli. This raises questions about the validity of such studies' findings, since Li et al.'s results show that when listeners are allowed to report perceiving no consonant, a sizable minority will do so.

As more and more of the /dɑ/ syllable onset is removed, including $F2_{onset}$, the number of listeners reporting /ɑ/ rather than /dɑ/ continues to grow, and reaches 100% once the entire F1/F2 transitions are removed. Although truncating /dɑ/ did not lead to /gɑ/ responses, lowpass-filtering the syllable at 2,000 Hz did: 30% of listeners reported /gɑ/ (p. 2604). This 2,000-Hz cutoff means that both the burst peak and the F3 zone are lost, which underscores the importance of one or both of these cues for identifying the consonant as alveolar: the F2 region on its own was not enough. When the burst is bandpass-filtered at 2,000 Hz, the 'new' burst peak is the peak located around 2,000 Hz (this particular /dɑ/ token has $F2_{onset}$ in its burst). Having the burst peak in the mid-frequency region is typical of a velar burst; it is no surprise, then, that 30% of listeners perceived this stimulus to be velar.

The unimportance of $F2_{onset}$ relative to the burst peak for identifying the place of articulation in /dɑ/ is further highlighted by the noise masking experiment: at 0 dB SNR, the F2 transition has disappeared beneath the noise floor and yet listeners' reports of hearing /dɑ/ are still at 100% (p. 2604). This is presumably because the burst peak is still above the noise floor. It is not until the noise is turned up to -6 dB SNR, at which point the burst peak has also been submerged by the noise, that the correct identification of /dɑ/ plummets.

The results for /gɑ/ also point in the direction of the burst peak being crucial. When the burst is truncated, approximately 40% of listeners report perceiving /dɑ/. This indicates that the F2 transition on its own seems to be ambiguous for distinguishing /g/ from /d/, at least in this pre-/ɑ/ context. But it also indicates an interesting asymmetry in the ambiguity: listeners hear /dɑ/ for /gɑ/ when the /gɑ/ burst is truncated whereas when the /dɑ/ burst is truncated there are no reports of perceiving /gɑ/. This asymmetry is in accord with Sussman et al.'s (1998: 257) prediction for burstless transitions that was shown in Figure 2.16 C whereby velars are more likely to be identified as alveolars than vice versa.

The results for /bɑ/ were the most difficult to interpret because only one out of the six tokens examined was identified as /bɑ/ in quiet by 100% of listeners. Instead there were some responses of /vɑ fɑ/ in the sample. Figure 2.30 above from Miller and Nicely showed a similar tendency for /bɑ/ to be confused with these fricatives. This potential for confusion seems to have played a role in the history of certain languages, e.g. /v/ merged with /b/ in medieval Spanish (Mackenzie, 2001: 96).

Why is /b/ liable to be misperceived as a fricative? To answer this question, Li et al. used the one /bɑ/ in the study that was correctly identified as /bɑ/ in quiet by 100% of listeners. When the burst of this /bɑ/ was truncated, the number of listeners who reported /fɑ vɑ/ jumped dramatically. When the speech-to-noise ratio was low enough to mask the release burst, the

number of /fɑ vɑ/ responses again jumped. These results together underscore the importance of the /b/ burst for signalling the consonant's manner of articulation.

Both /bɑ/ and /pɑ/ were rarely misidentified as alveolar or velar. There was only one very specific acoustic situation in which this was not the case: if /bɑ/ was lowpass-filtered at either 700 or 1,000 Hz, most listeners reported /dɑ/. However, this is probably an unimportant result because when everything above 700 or 1,000 Hz is lost, the listener has neither the F2 transition nor (most of) the release burst available to them. It is easy in such an extreme situation for different listeners to fill in the missing information differently, which is what is observed since some listeners still reported perceiving /bɑ/.

Overall, the results for /b/ suggest that the burst is important for correctly perceiving these consonants' manner of articulation but not its place of articulation.

Li et al. have concluded that the cues for identifying plosive place of articulation lie in the following burst frequencies before the vowel /ɑ/:

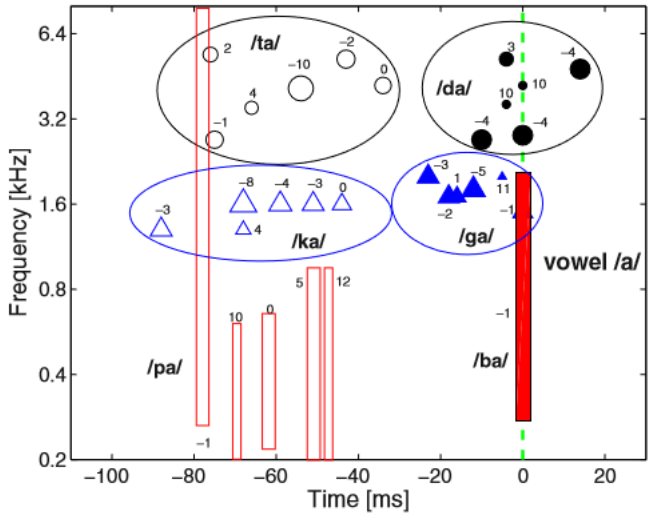


Figure 2.34: Schematic illustration of the key burst frequency regions postulated by Li et al. (2010) as defining the contrast between plosives' place of articulation before /ɑ/.
The x-axis indicates the time from the onset of voicing. If the energy in the above time-frequency regions is lost, then listeners are much more likely to misidentify the consonant than if energy is lost at other burst frequencies. The numbers next to each token indicate the speech-to-noise ratio at which the key burst information becomes masked; the lower the value the more robust the token is to noise. From Li et al. (2010: 2607).

The results are consistent with much previous research, especially for /t d k g/. The bilabials remain somewhat mysterious: one of the /p/ bursts is considerably different from the others in having its key burst information spread over a wide range of frequencies rather than being concentrated below 1,000 Hz. The results for /b/ are even more mysterious, since only one token out of the six was analysed due to it being the only token that was correctly identified as /b/ by 100% of listeners.

Nevertheless the three-dimensional deep search taps into variation between individual speech tokens with a level of detail and rigour that has not been matched by any study reviewed up until now. In particular, the approach has unearthed considerable differences between individual speech tokens in the speech-to-noise ratio at which the key place-of-articulation information becomes masked by noise. Such variation in natural speech would not have been evident if the results for all tokens of a phoneme had been averaged together, nor would they have been uncovered with the same confidence using a single methodology: the three-dimensional deep search, by triangulating on the problem using three methods, allows the researcher to converge on the most important information in the signal with a level of precision that would not be possible otherwise.

The authors acknowledge that in future the research will need to be applied to running speech. Another gap that will need to be filled is the range of phonetic contexts, which has been confined to CV syllables: in English, plosives contrast for place before sonorants (e.g. *brand* versus *grand*, *play* versus *clay*) and in syllable-final position (e.g. *wisp* versus *whisk*). Another challenge with the 3DDS is that because it is such a painstaking, rigorous methodology, the number of tokens that can be analysed is modest. Thus the results for the tokens explored above, while insightful and powerful, by no means exhaust the natural variability of speech, and it is likely that with a larger number of tokens and phonetic contexts further patterns will be unearthed.

2.3.6 Enhancing the Audibility of Plosives

Kapoor (2010) conducted an experiment on the perception of /tɑ kɑ dɑ gɑ/ that sheds further light on the most important information in the burst using a different technique. He modified the release burst to produce four kinds of stimuli: (1) the original stimuli; (2) stimuli in which the burst peak was amplified by 6 dB; (3) stimuli in which the burst peak was attenuated by 6 dB; (4) stimuli in which the burst peak was removed. The burst peak lies at a higher frequency in /tɑ dɑ/ than in /kɑ gɑ/; consequently the frequency region modified in the former was ca. 1,700 to 7,400 Hz and ca. 700 to ca. 2,400 Hz for the latter (p. 13). Once the stimuli were modified, white noise of five levels of intensity was added, yielding the following speech-to-noise ratios: 12 dB, 6 dB, 0 dB, –6 dB, –12 dB.

Here are the results:

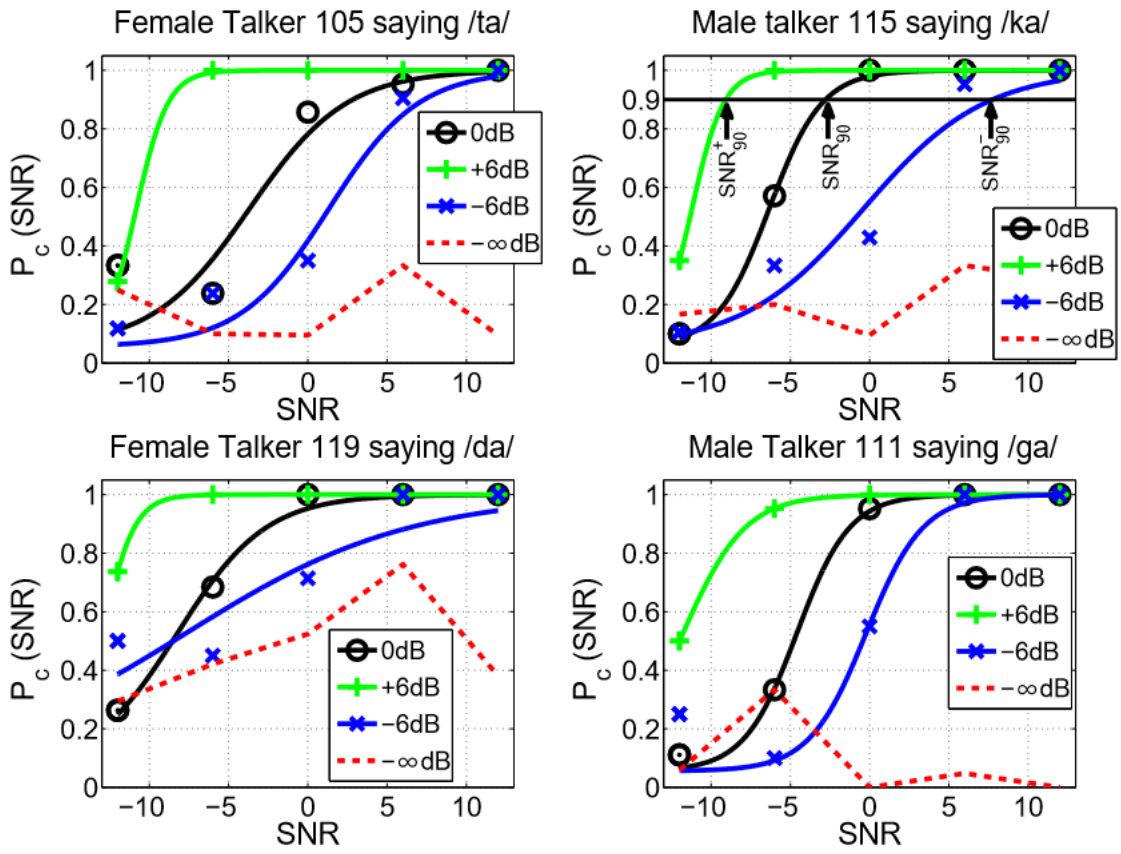


Figure 2.35: Some results of Kapoor's study (2010: 23).

The 21 listeners' rates of correct identification of four consonants are fitted with sigmoid functions. Scores were measured at speech-to-noise ratios of –12, –6, 0, 6 and 12 dB. Each curve is the sigmoid fitted to the measured data points. Curves labelled 'o' show the estimated curve fitted for the unmodified sound, the '+' marker is for the feature-amplified sounds, 'x' is for the feature-attenuated sounds while the dashed curve represents the recognition scores of the feature-removed sounds (which was not fitted with a sigmoid). The top right panel (m115ka) indicates the SNR90, SNR+ 90 and SNR− 90 points, defined as the speech-to-noise ratio at which listeners' accuracy at identifying the consonant goes below 90%.

The overall finding from the above diagrams is that there is a strikingly close relationship between the intensity of the burst peak and the ability of the consonant to be correctly identified in noise: the louder the burst peak, the harder it is to drown out the consonant with noise. This can be seen especially with the green line, which shows the identification of consonants whose burst peaks were amplified by 6 dB. Under this condition the consonant is correctly identified by 100% of the listeners even when the speech-to-noise ratio is –6 dB. In contrast, the original burst (the one not amplified by noise, represented by the black line) is only identified correctly by approximately 30% to 70% of listeners in the same speech-to-noise ratio. The bursts with the peak attenuated by 6 dB (blue lines) show even more sensitivity to noise, as represented by the fall in the blue curves beginning to the right of the black curves in all four diagrams. Nevertheless such attenuated bursts are correctly identified as alveolar or velar by all listeners

when there is no noise (the 12-dB speech-to-noise ratio in the above diagrams), which indicates that such bursts, though modified artificially, nevertheless sound like they belong to the same place of articulation as they did prior to being modified.

Kapoor reports (p. 39) that the speech-to-noise ratio at which the correct identification of /tɑ/ begins to decline is 6.1 dB lower than normal for the burst peak amplified by 6 dB, and 5.6 dB higher than normal for the burst peak attenuated by 6 dB. This is a remarkably close correlation; such results are further evidence pointing to the crucial role of the burst peak in identifying plosives.

The stimuli in which the burst peak has been removed altogether (represented by the dotted red line in the above diagrams) show poor identification by listeners even in the best noise condition of 12-dB SNR. Perhaps this is because (as was defined in Section 2.1) a release burst is a sound with frequency components at a wide range of frequencies. If energy is removed from one of this sound's frequency regions in its entirety, then it ceases to sound like a release burst, since it lacks the burst's characteristic spread of energy across frequencies.

Another noteworthy thing about these stimuli that lack a burst peak is that listeners frequently misclassified the place of articulation of such plosives in the highest speech-to-noise ratio (12 dB), but identified them *better* at a lower speech-to-noise ratio (6 dB). This can be seen in Figure 2.35 above by examining the trajectory of the dotted red lines. Kapoor's explanation of this observation runs as follows: if an alveolar burst has its high-frequency energy removed, burst energy remains in the mid and low frequencies. The energy at these mid- and low-frequency regions consequently becomes the new burst peak, since the burst peak is defined as the frequency in the burst spectrum that contains the most intense component. These new peaks are of course pseudo-peaks. They have the potential to fool the listener into perceiving the wrong place of articulation, hence they have been termed "conflicting cues" (pp. 37-39). Given that a mid-frequency peak cues velar place (and a low-frequency peak can cue bilabial place), the result is that listeners frequently report hearing /tɑ/ as /pɑ/ or /kɑ/ once the high-frequency peak is removed. However, once the noise is turned up (the 6-dB SNR condition), these conflicting cues are masked, having fallen beneath the noise floor. Consequently listeners can no longer be misled by them, and must instead identify place of articulation based on whatever residual cues might exist subsequent to the burst. This is why the listeners are better at identifying such stimuli when the noise is increased from 12 to 6 dB SNR.

Kapoor notes that the prominence of these conflicting cues varies from one token to another, and gives the example (p. 39) of one particular /tɑ/ burst that happened to have an

especially large amount of energy in the low (below 1000 Hz) frequency region: once the high-frequency peak was removed it was readily misheard as /pɑ/.

The results of this study underscore the importance of the burst for identifying plosives, but more specifically it is a powerful demonstration of the critical importance of the burst's peak. If the peak is attenuated it is harder to identify the plosive in noise; if the peak is amplified it is easier to identify the plosive in noise; and the degree (in dB) to which this is true is tightly correlated with the amount of amplification and attenuation applied to the peak.

2.3.7 Masking

The acoustic modelling to be employed in the present thesis does not incorporate masking. Nevertheless, the effects of this phenomenon are important to present because they affect the relative prominence of the two acoustic events that we are concerned with, the burst and the formant transitions.

Masking is a widespread phenomenon in auditory perception in which the presence of one sound reduces or even prevents our ability to hear another sound. For example if one turns up the volume of the radio in a car sufficiently, the engine may no longer be audible (Schnupp et al., 2011). This is an example of *simultaneous* masking: the engine and the radio emit sound at the same time. Masking can also occur *non*-simultaneously: one sound becomes more difficult (even impossible) to hear because of a sound that occurred shortly before or after it. This non-simultaneous masking can be divided into backward and forward masking. Forward masking is a situation in which the presence of a preceding sound reduces or prevents the hearing of a following sound. In backward masking, in contrast, the masked sound (called the probe (Sen and Allen, 2006)) comes *before* the masker. Backward masking is smaller in magnitude than forward masking and is often entirely absent if the experimental subjects are trained thoroughly (Miyazaki and Sasaki, 1984; Oxenham and Moore, 1994; cited by Moore, 2012: 110). Forward masking, in contrast, endures for a much longer time than backward masking, with the potential to last for up to 200 ms (which is approximately the length of a syllable), and is present even if the experimental subjects are highly trained (Moore, 2012: 110). As such it is more important than backward masking. If masking is sufficiently large the perception of one sound may be lost entirely, due to the presence of the other sound. More often the masked sound may be present but be of lower intensity than it would have been if the masker were absent.

The mechanism underlying simultaneous masking is believed to be the nonlinear dynamics of the outer hair cell. The mechanism underlying forward masking is less clear and is still a topic of controversy (Xie, 2013: 21). Moore (2012: 112) lists no less than five factors

that may contribute. For our present purposes the precise mechanisms underlying forward masking are less important than what the consequences of the phenomenon are for the perception of speech. Before examining some speech examples, let us examine the time course of forward masking as found in previous studies:

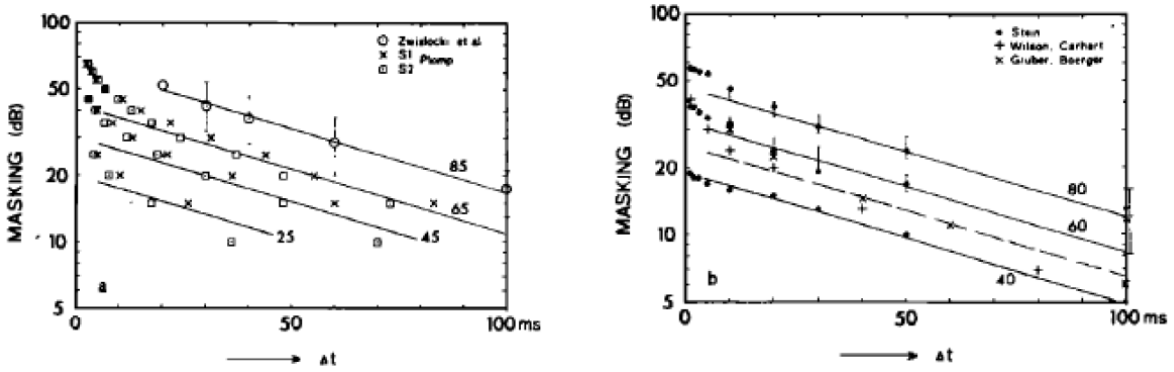


Figure 2.36: The magnitude (in dB) of forward masking as a function of time, as produced by a broadband masker.

(That is, a masker that is relatively acoustically similar to a release burst.) Each diagram is produced by aggregating the results of several previous psychophysical studies. When the masker is of greater intensity (as indicated by the number next to each sloping line) masking tends to last longer. Adapted from Duifhuis (1973: 1483).

More recent research has indicated that forward masking is approximately linear on a logarithmic time scale, and that the rate of its decay is more rapid for high-intensity maskers:

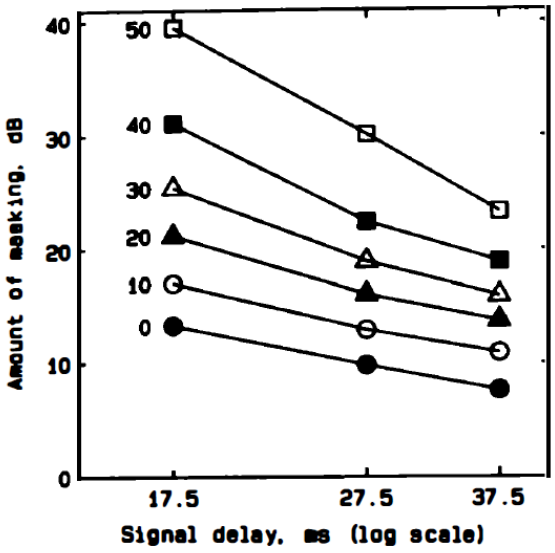


Figure 2.37: The magnitude of forward masking (in dB) as a function of logarithmic time, as produced by a broadband masker of varying intensity.

(Note that the intensity of the masker is next to the square, triangle, or circle which it represents.) When the masker is of greater intensity (as indicated by the number next to each sloping line) masking tends to decay somewhat more rapidly but nevertheless be greater in magnitude. Adapted from Moore (2012: 111), itself adapted from Moore and Glasberg (1983a).

The above diagram shows the masking of the probe for just the first 37.5 ms after the end of the masker. After this time period masking continues to decay and reaches zero approximately 100-200 ms after the end of the masker (Moore, 2012: 111).

We have already encountered the AI-gram, a kind of auditorily-inspired spectrogram used by Allen's research group to study speech in noise. Although the AI-gram is auditorily inspired, there are certain aspects of auditory processing that it does not incorporate. Among these is forward masking. Xie (2013) sought to rectify this by applying to the AI-gram a model of forward masking based on previous psychophysical research.

Xie (2013: 52) and others listened to the speech before and after applying the forward-masking model and reported "there is no change in the speech perception". The results of his model, then, can give us a sense of how the spectral envelope is modified by forward masking, and hence how the acoustics of speech look different on a display that incorporates forward masking relative to conventional displays widely used in speech science that do not incorporate knowledge of forward masking.

Xie modelled the effect of forward masking on a wide variety of speech sounds; this review discusses the results for plosives only. Here are the results for /tɑ dɑ/ before and after Xie's masking model, as displayed on the AI-gram with modest noise (SNR = 12 dB):

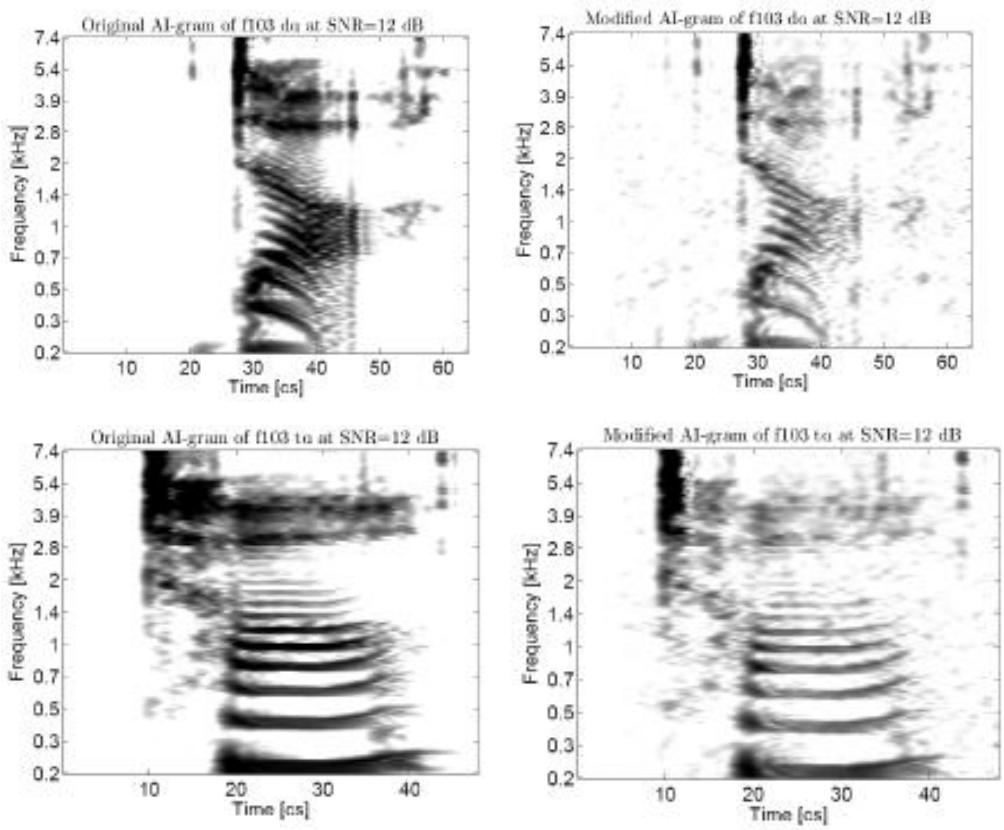


Figure 2.38: AI-grams for /dɑ tɑ/ before and after the application of Xie's model of forward masking. The syllables are played in a speech-to-noise ratio of 12 dB, which is a relatively low level of noise. Because the AI-gram is specially designed to display noise as white, the noise not visible in the above displays; only the speech components are. From Xie (2013: 39).

Figure 2.38 shows that the release burst remains intense, that is, it remains unaffected by the masking. This is unsurprising given that forward masking preserves onset sounds relative to following sounds. The formant transitions, in contrast, are much less intense and hence much less salient than they were prior to the application of forward masking. This should come as no surprise given that the experimental results in Figure 2.37 show that masking is greatest immediately after the masker. Xie (p. 40) reports that the maximum attenuation in the above /dɑ/ and /tɑ/ syllables was 20 dB, which occurs at the beginning of the formant transition.

Here are the results for /gɑ kɑ/:

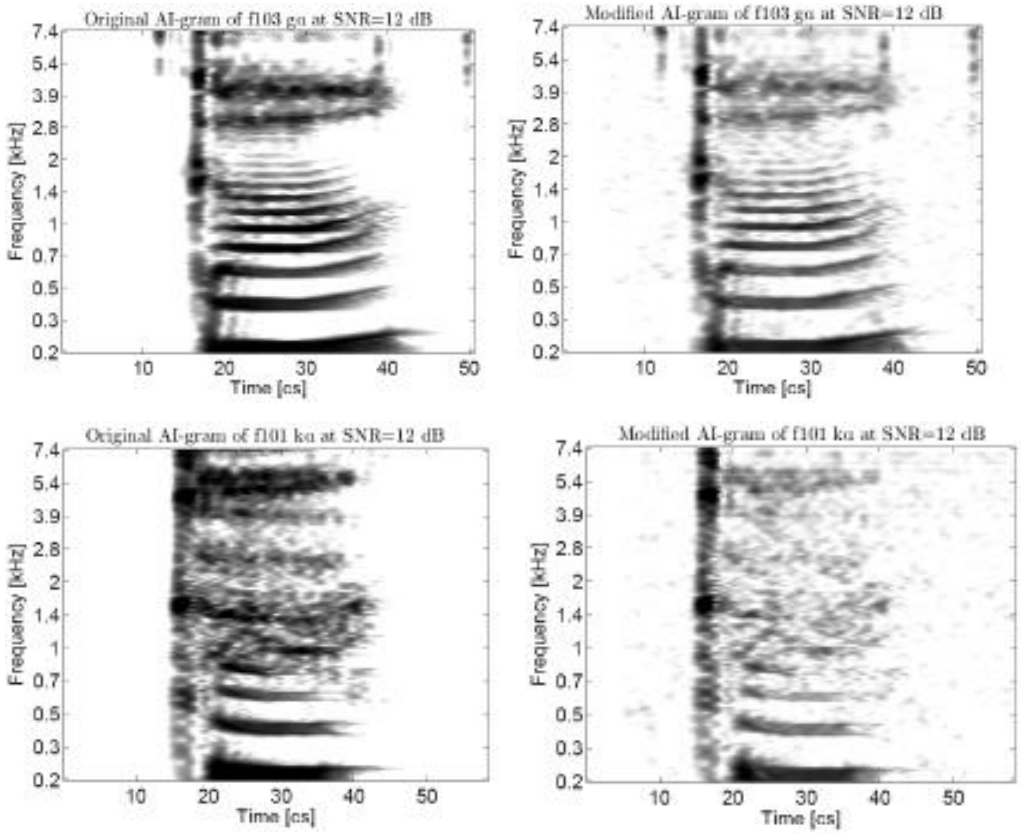


Figure 2.39: AI-grams for /gɑ kɑ/ before and after the application of Xie's model of forward masking. The acoustic conditions are the same as those described under Figure 2.38. From Xie (2013: 38).

We again see a marked attenuation of the formant transitions relative to the release burst. This attenuation is again at its maximum during the formant transition, and is 10 dB for /gɑ/, 18 dB for /kɑ/.

And here are the results for /bɑ pɑ/:

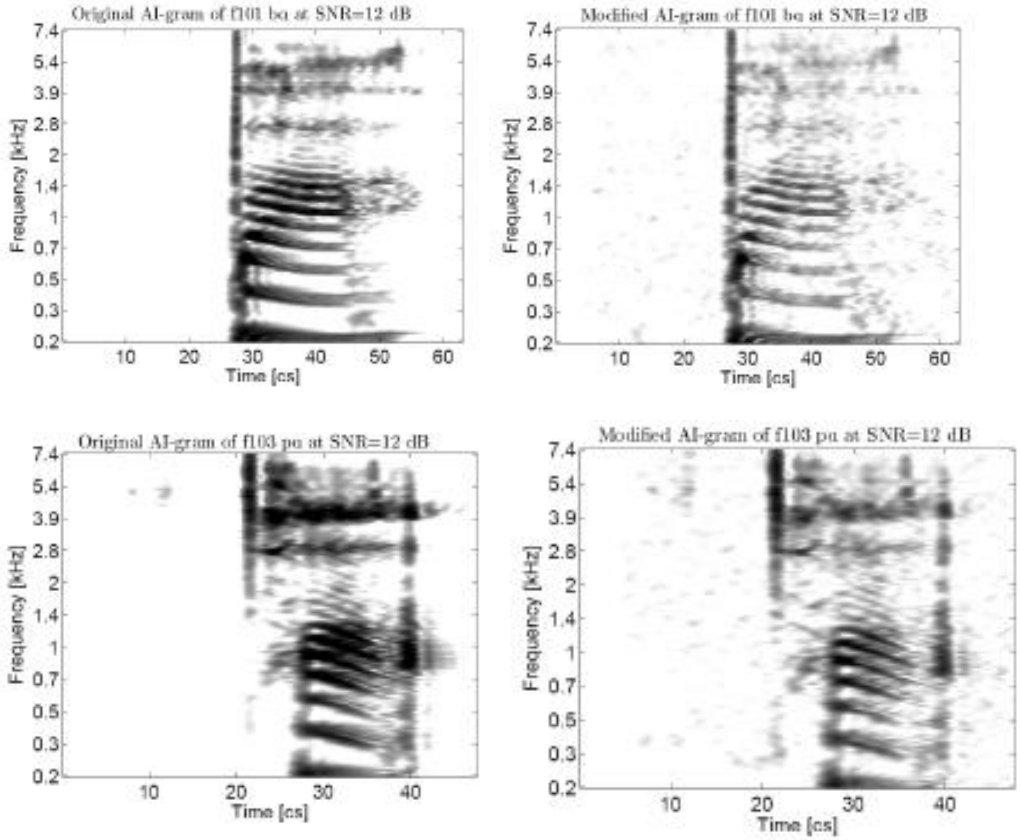


Figure 2.40: AI-grams for /bɑ pɑ/ before and after the application of Xie's model of forward masking. The acoustic conditions are the same as those specified in Figure 2.38. From Xie (2013: 36).

Again, a marked decrease in the magnitude of the spectral components following the release burst can be seen. This attenuation is 14 dB at its maximum in both syllables, which is somewhat less than in the case of /tɑ dɑ kɑ/, which were 20, 20, and 18 dB respectively. This is because the greater the intensity of the onset stimulus, the greater the masking, and /bɑ pɑ/ have long been known to have lower-intensity release bursts, as we saw earlier in Li et al.'s (2010) noise-masking results.

Xie continued his study of masking by examining the syllable /tɑ/ in increasing levels of noise, comparing SNRs of 15, 6, 0, and -3 dB. Here are the results for the two lowest of these SNRs:

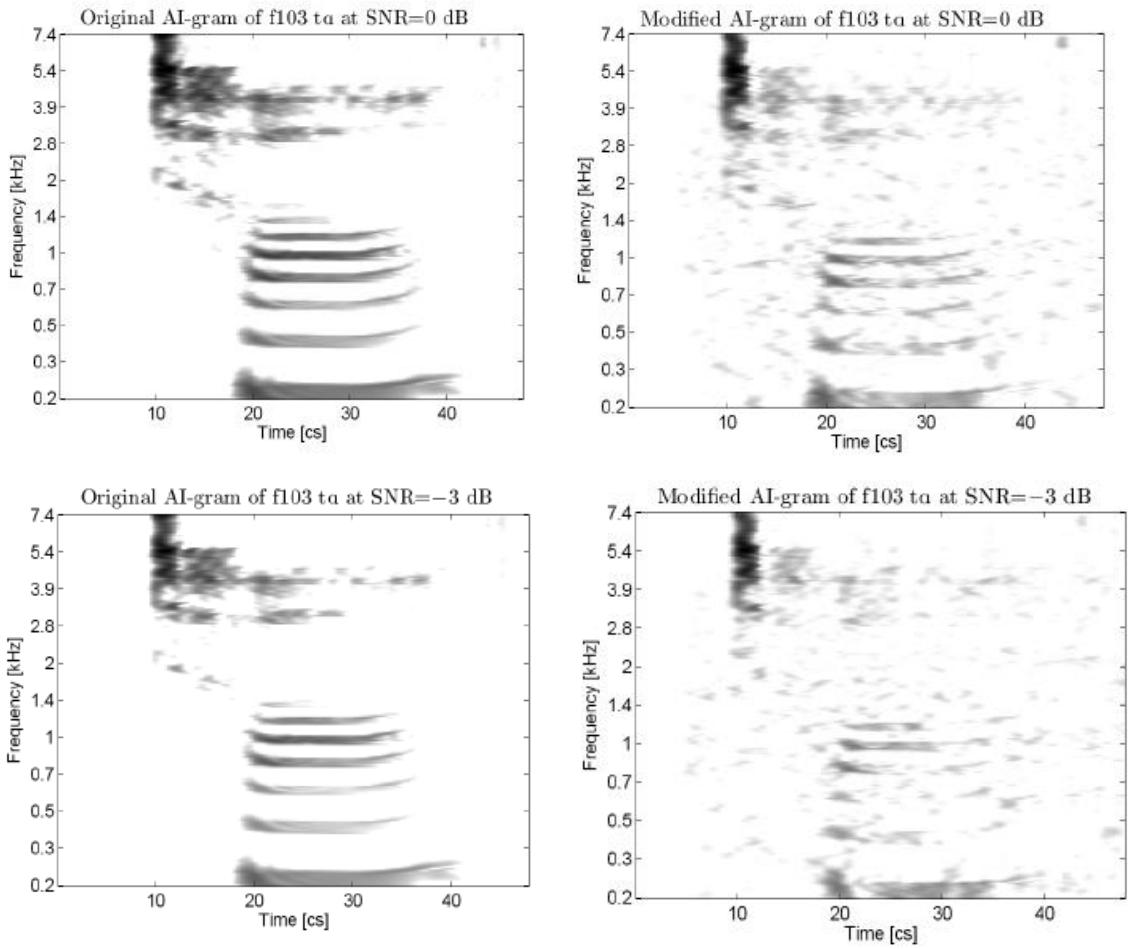


Figure 2.41: AI-grams before and after modification by Xie's forward-masking model.
These are for the two lowest speech-to-noise ratios he examined, namely 0 dB and -3 dB. From Xie (2013: 51).

Recall that the AI-gram displays in white all spectral components that have disappeared beneath the noise floor. Thus the above images indicate how robust particular speech components are to noise when knowledge of forward masking is taken into account. We see that the /tɑ/ burst remains robust to noise whereas the following aspiration is greatly attenuated. The onset of the vowel can also be seen to be more robust to noise than the middle part of the vowel, which is particularly evident when comparing the two AI-grams for the lowest (-3 dB) speech-to-noise ratio. As we saw earlier, Li et al. (2010) found the high-frequency part of the burst to be crucial for the perception of /t/, and this component remains prominent under all noise conditions.

Xie's study is an illuminating glimpse into the effect of forward masking on the perception of speech. Forward masking maintains the prominence of the onset of sounds relative to subsequent sounds. Xie interprets this (p. 52) as undermining the importance of the F2 transition as a cue, since the intensity of the F2 transition is reduced by the preceding burst, which makes it less audible and less robust to noise. Alwan (1992) masked the F2 region of /ba da/ with bandpass noise and found that listeners could still successfully distinguish the place of articulation in noise. Alwan et al. (2011: 197) interpreted this as meaning that there are cues in

the high-frequency region that are sufficient for distinguishing place. However it could also be true that given that the F2 transition is masked by the burst, then perhaps the F2 transition would not be particularly prominent even if there were no bandpass noise.

This claim of Xie's is supported by a study by Cvengros (2011), who designed a perceptual study in which the following seven kinds of stimulus were played to 20 listeners:

(1) Unmodified /$C_{plosive}$V/ syllables in which /$C_{plosive}$/ was one of /d g k p/;

(2) The said syllables but with the burst peak deleted;

(3) The said syllables but with the non-peak part of the burst deleted;

(4) The said syllables but with $F2_{onset}$ deleted;

(5) The said syllables but with the F2 transition deleted;

(6) The said syllables but with both $F2_{onset}$ (= 4) and the F2 (= 5) transition deleted;

(7) The said syllables but with (2), (3), (4) and (5) deleted.

Cvengros found that when either $F2_{onset}$ or the rest of the F2 transition was deleted from the syllable, listeners' accuracy at identifying the consonant was not statistically significantly different from when the entire stimulus was present. It was only when both $F2_{onset}$ *and* the F2 transition were deleted that listeners' accuracy was different enough from those for the full stimulus to scrape over the $p < 0.05$ threshold ($p = 0.04$). Nevertheless, even in this condition the accuracy of listeners was just 3.4% lower than for the stimuli in which the formant information was not deleted. In contrast, when the burst peak was removed the listeners' accuracy plummeted by 39.4%. This is over 10 times larger than the loss in accuracy when the formant information was removed. Furthermore, the loss in accuracy when the burst peak was lost was far more statistically significant than when the formant information was lost ($p = 0.000023$ as against 0.04). Note also that the disimprovement in listeners' accuracy when the non-burst-peak part of the burst was removed was 1.6%.

Cvengros's results deserve to be taken more seriously than the results of the 1970s research that was reviewed in 2.3.3 (Cole and Scott, 1974 a, b; Dorman et al., 1977; Stevens and Blumstein, 1978). This is because his stimuli do not involve violating the temporal structure of the syllable. Furthermore, he only deleted the F2 transition and kept the F1 transition, as we suggested in Figure 2.30 in our discussion of the limitations to the 1970s research. Also, Cvengros did not force listeners to respond with only /p t k/ or /b d g/; instead, listeners could respond with any consonant they liked. This is in contrast to much of the early research (e.g. Liberman et al., 1952; Malécot, 1958) that we reviewed earlier.

The overall conclusion to be taken from Cvengros's work and that of the other researchers we have reviewed in the last six sections is that the burst peak appears to be by far the most important piece of information for perceiving place of articulation, with the non-peak

part of the burst and the F2 transition playing a modest secondary role. Nevertheless because the present study is acoustic rather than perceptual in nature, all potential acoustic attributes will be examined in detail, i.e. both formant attributes (Chapters 5 and 7) and burst attributes (Chapters 6 and 7).

2.3.8 Normalization

As stated at the beginning of the present chapter, one of the aims of the present study (Aim 2) is to compare the classification of burst-based attributes with and without speaker normalization. Speaker normalization can be defined as any normalization of an acoustic attribute in which each token is expressed relative to a speaker-specific value (e.g. the speaker's mean value for that attribute) rather than as an absolute, speaker-independent value.

The motivation for testing normalization on the burst is to fill a gap in the literature: many studies have employed normalization on periodic acoustic phenomena such as vowels, whereas – to the best of my knowledge – no previous phonetic research has attempted to normalize aperiodic phenomena such as the burst. Normalization can mitigate differences in acoustics caused by individual differences in anatomy, speaking style, or articulation. Such differences can arise from a variety of sources: one speaker (e.g. f02 in the present study's dataset) may talk more loudly into the microphone than another (e.g. f03), or one speaker (e.g. f03) may have a tendency to realize /t d/ as apical postalveolar, or one speaker (e.g. m06) may have a greater tendency to prevoice /b d g/ than other speakers, with potential reduction of the amplitude of the burst.

Forrest et al. (1988) found that they could correctly classify the place of articulation of over 96% of the female-produced voiceless obstruents in their data using a discriminant-analysis model trained on their male data. These results suggest that normalization of burst attributes might not be necessary, since the results were achieved without normalization. However, their dataset was relatively small (five male and five female speakers) and consisted of single words uttered in isolation. Thus testing of normalization on burst attributes still seems warranted.

In terms of comparing the performance of normalization techniques for vowel formant frequencies, Flynn (2012) and Adank et al. (2004) are the most recent widescale studies. The latter study utilized three criteria for evaluating normalization methods: (1) preserve phonemic information; (2) preserve information about the speaker's background (i.e. sociolinguistic information); and (3) minimize anatomical/physiological information in the acoustic representation of vowels (p. 3099).

The present study is not sociolinguistic in nature and so criterion (2) will not play a role. Nor is the present study articulatory in nature, which means that criterion (3) above can also be struck from consideration. Instead, the sole criterion that will be adopted for evaluating the quality of normalization in the present study is criterion (1), namely the extent to which the normalization enhances the strength of a given acoustic attribute to distinguish place of articulation.

There have been many normalization procedures invented in the history of vowel formant-frequency normalization (see Flynn, 2012: Chapter 6 for a detailed presentation of each). Adank et al. (2004) examined the performance of 11 normalization procedures and found that the Lobanov, Nearey, and Gerstman methods performed strongest (Lobanov, 1971; Nearey, 1978; Gerstman, 1968). After comparing the three sources of variation (vowel, region, and gender) by multivariate analysis, Lobanov-normalization was found to be the best procedure.

Because of this finding, the present study will employ Lobanov-normalization as its normalization method against which to compare the non-normalized burst attributes. Admittedly, it is not known whether the choice of Lobanov-normalization based on the results for *vowels* can justify the use of this normalization method on *consonants*, but given that none of the previous research on plosives appears to have examined normalization, some sort of starting point has to be followed.

The formula for Lobanov-normalization is as follows (Lobanov, 1971: 606):

$$\text{Lobanov} = \frac{x - \mu}{\sigma}$$

The speaker's mean value for the attribute (μ) is subtracted from the value of a given token (*x*), and this is divided by the speaker's standard deviation (σ) for that attribute.

This study also will present the results of a variant type of normalization in which there is no division by the standard deviation. We refer to this normalization as 'Norm':

$$\text{Norm} = x - \mu$$

Where *x* again represents an individual token in the dataset and μ represents the mean value for that attribute in the speech of an individual speaker. The reason for including Norm is that comparing its performance with Lobanov quantifies whether and to what extent the division by the standard deviation (σ) improves the classification accuracy above and beyond that which was yielded by the subtraction of the mean (μ).

In addition to using normalization on burst attributes, the present study also uses normalization on formant attributes as has already been widely practiced in phonetic science. Section 5.4.2 employs Norm above and compares normalization by individual speaker to: (1) no normalization; and (2) normalization by speaker sex. The theoretical motivation for these comparisons is outlined in 5.4.2.

In conclusion, the present study appears to be the first largescale study to examine the effect of speaker normalization on the performance of burst-based attributes. Consequently it is difficult to have a theoretical expectation of whether and to what extent normalization of burst attributes will improve classification accuracy. The motivation for examining normalization is part of the overarching aim of the present study stated in the Overview, namely to examine certain aspects of plosives' place of articulation that appear not to have been investigated by most previous studies.

2.4 Conclusions

These are the key findings of this chapter:

- $F2_{locus}$ is expected to yield higher classification accuracy than $F2_{onset}$ on /b d g/, a hypothesis that will be tested in Chapter 5 (Aim 1 of this thesis).
- $F2_{onset}$, $F2_{mid}$ (and $F3_{onset}$) can classify the place of articulation of /b d g/ well above chance but well below 100%. The highest accuracy was achieved by Sussman et al. (1991) – 77% on highly-controlled data – but most studies have found accuracies of 65-70% (Hasegawa-Johnson, 1996: 25).
- Regarding the burst, many acoustic attributes have been proposed for extracting place of articulation information in the history of phonetics, and there is considerable uncertainty over what burst attributes perform best. Our review of perceptual studies indicates that the burst peak usually contains the most important information for place of articulation.
- This proliferation of burst-based attributes in the history of phonetics has proceeded without comparing the performance of such attributes with a feature set (DCT coefficients) that are theoretically expected to capture more of the variance in the burst envelope than the traditional-phonetic attributes, and which have been widely used in ASR for decades. Thus Aim 4 of this thesis is to compare the performance of the two attribute types on the burst (Chapter 7).
- Another respect in which phonetic and ASR approaches have tended to differ is in their choice of spectral representation. Thus Aim 3 of this thesis is to investigate whether the

choice between a Hz-dB, Bark-phon, and Bark-sone spectrum affects the classification accuracy of burst attributes.

- The acoustics of bilabial bursts have proven difficult to characterize precisely. They appear to be relatively broadband and yet Liberman et al. (1952) succeeded at yielding a bilabial percept using narrowband burst spectra. Results from the Three-Dimensional Deep Search (Li et al., 2010) suggest that the burst is less important for perceiving bilabial place than for the other two places of articulation.
- Forward masking leads us to expect $F2_{onset}$ to be attenuated by the burst in /b d g/. The more intense the burst, the larger this attenuation of $F2_{onset}$ would be. Nevertheless, given that the present study is acoustic rather than perceptual in nature, all potential plosive place of articulation features, including $F2_{onset}$, will be utilized (Chapters 5, 6, and 7).
- Normalization by individual speaker has long been practiced in phonetics for vowel formants. However, it is unclear whether normalization of burst attributes would improve the classification accuracy of such attributes. Aim 2 of this thesis is to test this (Chapter 6).

Chapter 3: Pilot Study

This chapter presents the results of the pilot study. This study was undertaken prior to the main study and explores the performance of various acoustic attributes on the release burst and formant transitions so that any of the attributes that perform particularly poorly at distinguishing place of articulation can be excluded prior to the main study. This pre-trial was necessary in that several formant attributes were original to the present study and needed to be tested on a small scale first to decide whether to bring them forward to the main study.

This chapter begins with the methodology, detailing the participants, recording, material, segmentation, annotation, and data extraction (Section 3.1). Section 3.2 introduces what each of the burst attributes do and presents their classification accuracies as individual attributes. In 3.3 the same is done for the formant-based attributes. In 3.4 the performance of these attributes when combined with each other is examined. Section 3.5 discusses the findings of the pilot study and its limitations. 3.6 concludes.

3.1 Methodology

3.1.1 Participants

There were two participants, one male, one female. Both were from the south of England and both were in their twenties at the time of recording (2015). Due to the small scale of the pilot, the aim was to have relatively controlled data and so speakers were recruited with an accent as similar to each other as was feasible. There were nevertheless some differences in their speech: the female speaker spoke RP whereas the male speaker spoke a more noticeably south-eastern variety, what might be termed south-eastern 'near-RP' in Wells's (1982) taxonomy. For example, his FLEECE vowel is [ɪi] whereas the female speaker's is [iː] or [i̟i]; his NEAR is usually [ɪː] whereas hers is usually [ɪə]. All the other vowels, however, show little difference between the speakers.

3.1.2 Recording

Prior to recording, ethical approval was sought and a full application was not deemed necessary due to the non-confidential nature of the participants' task. Nevertheless a consent form with information about the task was provided to participants to sign beforehand. The speakers were compensated financially for their participation.

The participants were recorded in an anechoic booth using a Roland Edirol R-44 recorder and a Roland CS-50 microphone. The sampling frequency was 44.1 kHz with 16-bit

quantization, recorded in mono channel. The microphone was approximately 10 inches from the speakers' mouths at an angle of appoximately 45˚. The sentences were presented in randomized order one by one on a computer screen using Microsoft PowerPoint.

3.1.3 Material

The material consisted of /hVCV/ sequences inserted into the carrier phrase *They blur* ____. The /h/ before the VCV and the word *blur* were chosen to keep F2 coarticulation from outside the VCV to a minimum (the /ɜː/ of *blur* has a relatively average, neutral F2 value, and /h/ assumes the F2 value of following vowel rather than having an $F2_{onset}$ frequency of its own).

Due to the small-scale nature of the pilot, the aim when designing the material was to have relatively highly controlled data that nevertheless covered a comprehensive range of F2 variation in vowels: in the literature review it was noted that velars in particular have been found to vary greatly in their burst and formant acoustics depending on the backness of the following vowel. Thus three front vowels were chosen (/iː ɪə ɛː/), two back vowels (/ɑː ɔː/), and one central vowel (/ɜː/). Prior to reading the sentences, the speakers listened to a pre-recorded demonstration of the sentences being read by the author in which the nucleus was placed on the final syllable.

The following table shows the entire set of contexts for /b/ (there were analogous words for /d g/):

ˈhiː ˈbiː *hee bee*	ˈhiː ˈbɪə *hee beer*	ˈhiː ˈbɛː *hee bair*	ˈhiː ˈbɜː *hee bur*	ˈhiː ˈbɑː *hee bar*	ˈhiː ˈbɔː *hee baw*
ˈhɪə ˈbiː *heer bee*	ˈhɪə ˈbɪə *heer beer*	ˈhɪə ˈbɛː *heer bair*	ˈhɪə ˈbɜː *heer bur*	ˈhɪə ˈbɑː *heer bar*	ˈhɪə ˈbɔː *heer baw*
ˈhɛː ˈbiː *hair bee*	ˈhɛː ˈbɪə *hair beer*	ˈhɛː ˈbɛː *hair bair*	ˈhɛː ˈbɜː *hair bur*	ˈhɛː ˈbɑː *hair bar*	ˈhɛː ˈbɔː *hair baw*
ˈhɜː ˈbiː *hur bee*	ˈhɜː ˈbɪə *hur beer*	ˈhɜː ˈbɛː *hur bair*	ˈhɜː ˈbɜː *hur bur*	ˈhɜː ˈbɑː *hur bar*	ˈhɜː ˈbɔː *hur baw*
ˈhɑː ˈbiː *har bee*	ˈhɑː ˈbɪə *har beer*	ˈhɑː ˈbɛː *har bair*	ˈhɑː ˈbɜː *har bur*	ˈhɑː ˈbɑː *har bar*	ˈhɑː ˈbɔː *har baw*
ˈhɔː ˈbiː haw bee	ˈhɔː ˈbɪə *haw beer*	ˈhɔː ˈbɛː *haw bair*	ˈhɔː ˈbɜː *haw bur*	ˈhɔː ˈbɑː *haw bar*	ˈhɔː ˈbɔː *haw baw*

Table 3.1: The full set of VCV contexts for /b/. Analogous sequences were produced for /d/ and /g/.

Each VCV sequence was presented three times. Thus 2 speakers × 36 vowel contexts × 3 consonants × 3 repetitions = 648 tokens. Four of these tokens were omitted from the dataset due to the speaker hesitating whilst uttering them. The presentation of the items was randomized.

Some of the above words (e.g. *bee*, *beer*) correspond to real words in English whereas others (e.g. *bair*, *baw*) do not. Nevertheless even the real words are rendered semantically nonce

by virtue of being uttered in the carrier. No attempt was made to control for the real-versus-nonce difference beyond this; the aim was to have wide coverage of F2 variation and so some use of real words was unavoidable while simultaneously maintaining the phonotactic consistency exemplified in Table 3.1 above.

3.1.4 Segmentation

Segmentation was done using the Praat software package (Version 5.4.03, 18 December 2014; Boersma and Weenink, 2014). Although semi-automatic segmentation was initially trialled, the procedure ended up being effectively manual: all boundaries were decided by the researcher. Because the syllables consisted of nonce words, it was decided not to run a more automated forced-alignment segmentation. Several annotation tiers were added to mark the location of subphonemic entities of interest (see Figure 3.1):

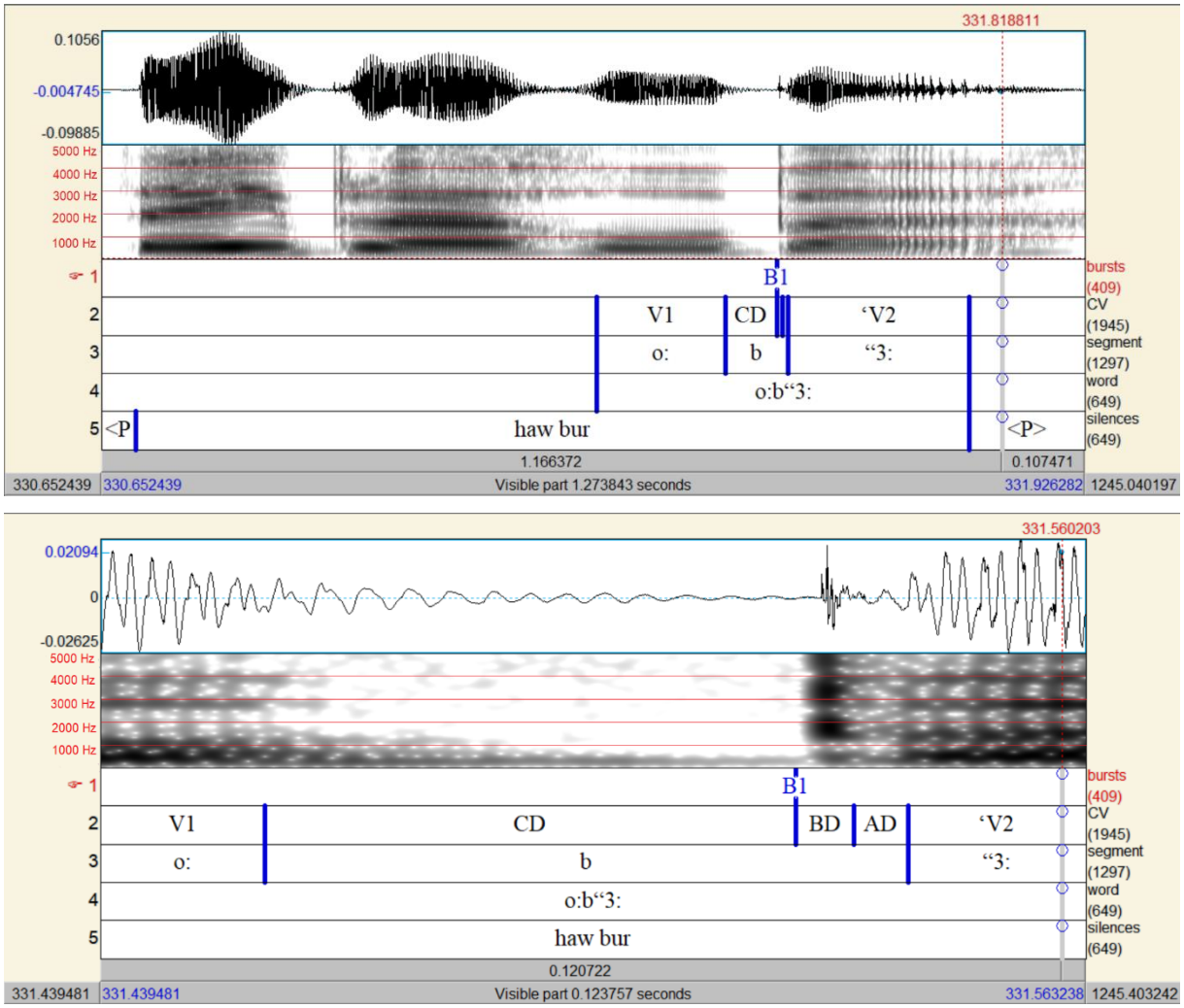


Figure 3.1: Screenshots showing the five annotation tiers used in the present part of the study.
The example is *They blur haw bur* uttered by f01. The top screenshot shows the entire sentence: the 'silences' tier contains an orthographic transcription of the two words in which the plosive is situated (viz. 'haw bur'); the tier also records the duration of the entire sentence (this information was not utilized). The bottom screenshot zooms in on the plosive to show the boundaries of the closure ('CD'), burst ('BD'), and aspiration ('AD').

These subphonemic entities were the plosive's closure (labelled 'CD'), the burst ('BD'), and the aspiration ('AD'). In addition, the boundaries of the surrounding segments were marked. Given that all sequences were VCV, this was simply the two vowels (named 'V1' and 'V2', respectively).

The 'word' tier consists of a phonetic annotation of the same utterance with a lefthand boundary corresponding to the beginning of the vowel preceding the plosive and a righthand boundary corresponding to the end of the vowel following the plosive. The 'segment' tier corresponds to the same information as the word tier except that the boundaries for the individual segments are added. The 'CV' tier's primary function is to mark the beginning and end of the three subphonemic events in the plosive, namely the closure ('CD'), burst ('BD'), and aspiration ('AD'). The 'bursts' tier marks the beginning of each transient in the burst.

Whilst annotating, the dynamic range on the spectrogram was set to 50 dB and the spectrogram displayed frequencies from 0 to 5,000 Hz. The tier 'bursts' marked the onset of the burst using a point tier. It was also used to mark the number of release transients observed in the entire burst ('B1' for the first transient, 'B2' for the second, etc.). This tier was not used in the main study (Chapter 4) as the information on the number of transients was judged not to give important information on place of articulation.[4]

The boundary for the end of the [h] (= the beginning of the vowel V1) was put at the point where the vowel's F1 appeared to reach approximately a similar amplitude on the spectrogram as it has during most of the following vowel. In some cases, such as in the above screenshot, this meant that at the point at which the beginning of the boundary was placed, the F2 had not yet reached its full amplitude, i.e. there appears to have been traces of breathy voice from the [h] remaining in this formant. However, when the annotation was played aloud, no trace of the [h] could be heard.

Regarding the segment tier, it can be seen that it is a straightforward rendering of the three phonemes in the VCV sequence. The beginning of the plosive was taken as the point at which there is a sudden decrease in the amplitude of F3:

[4] Although there appeared to be some tendency for pre-front-vowel /g/-bursts in particular to have more than one transient, there were also occasional instances of /b/ having more than one transient (in the speech of f01). The segmentation of individual transients was thus not carried forward to the main study as it appeared to provide unimportant information on place of articulation (i.e. all three places of articulation usually had release bursts that consisted of only one transient).

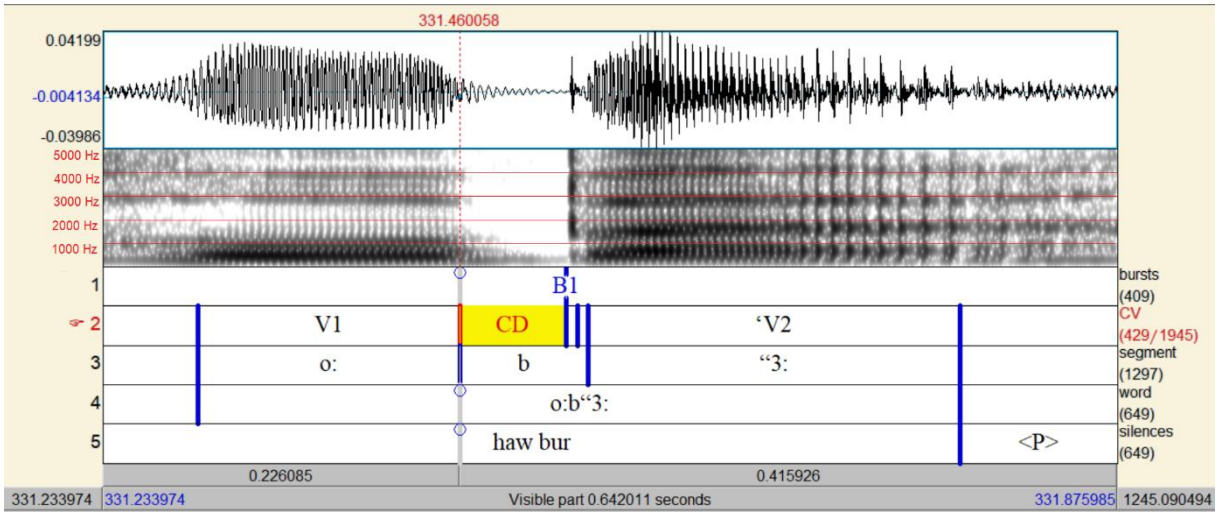


Figure 3.2: Screenshot showing the boundary for the beginning of the plosive (the red dotted line at 331.46 s). Note its coincidence with the decrease of the third formant's amplitude. (The " and ' symbols indicate where the nuclear stress is.)

F3 was chosen because whenever there is a discrepancy in the time at which the formants' amplitudes decrease, it is F3 that decreases first. Thus if F2 or F1 had been chosen as the criterion for segmentation, there would have been cases where the $F3_{offset}$ of V_1 was low in amplitude, leading to potential inaccuracies in the measured frequency of this formant that would have had to be corrected.

Figure 3.2 illustrates a further challenge: deciding the end of the utterance-final vowel. This proved difficult in the speech of the female speaker, who tended to sigh at the end of each sentence. The policy adopted was to put the boundary where the creaky voicing ended. In Figure 3.2 this is straightforward to see because the creaky voice at the end of the vowel produces long glottal cycles that disappear once the voicing has ceased. The male speaker had little or no tendency to sigh at the end of utterances, which made segmenting his utterance-final vowels relatively straightforward.

In the lower screenshot of Figure 3.1, the burst can be seen in both the spectrogram and waveform. As with the other boundaries, the spectrogram was used as the criterion for deciding where the burst began. With hindsight this may not have been the wisest decision as the spectrogram is the derivative of the waveform, not the other way around. For all the other boundaries we have discussed up until now, the discrepancy is inconsequential: the precise duration of the surrounding vowels, for example, was not important in the study. The burst, however, is such an ephemeral acoustic event that it is best to segment it with precision. Thus the segmentation criteria for the burst in the main study are based on the waveform.

The lower screenshot of Figure 3.1 also shows how the end of the burst was segmented. As can be seen, this was put at the point where there is a fairly sudden decrease in the amplitude

of the burst at a wide range of frequencies (for alveolars this decrease in amplitude is located in the frequencies above ca. 3,500 Hz while for velars it is located near the F2 if the following vowel back or near F3 if the following vowel is front).

It can be seen in the lower figure of Figure 3.1 that the boundary marking the commencement of voice onset time was placed at the point where the first periodic black striation appears in the spectrogram. In some cases, there was evidence of acoustic energy in the low-frequency region slightly before this, known as an edge vibration, but this was ignored as the edge vibration is regarded as unimportant (Lisker and Abramson, 1964: 417).

In the main study (see Section 4.3 for further details), the spectrogram rather than the waveform was followed in such matters due to its greater temporal precision.

3.1.5 Annotation

X-SAMPA was used for the phonemic annotation (Wells, 1995) as it can be typed more rapidly than copying and pasting phonetic symbols or using an on-screen keyboard. There were only six vowels and three consonants in the pilot, as indicated in the following table:

IPA symbol	X-Sampa
iː	i:
ɪə	I@
ɛː	E:
ɜː	3:
ɑː	A:
oː	O:
ə	@
ˈ	“
b	b
d	d
g	g

Table 3.2: The annotation used in the pilot study.

3.1.6 Data Extraction

A Praat script was written to extract the data automatically. F1, F2, and F3 were extracted at 10-ms intervals for the final 90 ms of the vowel preceding the plosive (yielding 30 datapoints, 10 for each formant) and for the first 170 ms of the vowel following the plosive (yielding 54 datapoints, 18 for each formant). This meant that the vowel measurements were made with reference to absolute vowel durations rather than relative durations (e.g. percentage of total vowel duration). In this chapter, reference is made to ‘$F2_{mid}$’. In reality, this $F2_{mid}$ is not from the literal midpoint of the vowel but rather from the vowel steady-state. The vowel steady-state

was extracted from 170 ms into the vowel following the plosive and 90 ms before the end of the vowel preceding the plosive. To verify that this part of the vowel was indeed steady, the F2 transitions for the context with the longest formant transitions (namely [do]) were examined and it was found that the formant was steady by 110 ms at the latest, thus meaning that 170 ms was verifiably part of the vowel steady state. The same procedure was followed for the vowel preceding the plosive (e.g. [od]) and it was again verified that 90 ms before the closure was always situated in the vowel steady state. For the main study, it was decided that the use of absolute vowel durations rather than relative durations was unnecessarily out of sync with previous research and relative vowel durations were adopted instead.

For the formant extraction, Praat's formant tracker was used, with the following settings: LPC using the Burg method, 5-ms step, and a 25-ms Gaussian window. For the female speaker, the 'maximum formant' parameter was set to 5,500 Hz, and for the male speaker it was set to 4,500 Hz (the standard setting of 5,000 Hz was tried but 4,500 Hz was noted in a few cases to track the visible formants better). Subsequent to their extraction, formant tracking errors were corrected by examining spectrographic images of each syllable that were outputted by the script that extracted the data.

A DFT spectrum was centred at the middle of the interval marked burst ('BD'). The duration of the window corresponded to the duration of the burst.[5] This criterion was dropped in the main study since it results in the frequency resolution varying in the window from one token to another depending on how long the burst happened to be. The window used was Kaiser2. Prior to spectral analysis the recording was downsampled to 20 kHz and lowpass-filtered at 10 kHz. Pre-emphasis was applied starting from 50 Hz. The initial DFT was smoothed using cepstral smoothing (1,000 Hz).

3.1.7 Statistics

The results will be presented using three statistics. For evaluating the performance of individual attributes, discriminant analysis and Wilks's Lambda will be used, while for evaluating larger groups the statistic of choice is random forests.

3.1.7.1 Discriminant analysis

This statistic is employed to evaluate the ability of a predictor to distinguish a categorical variable (in the present case, place of articulation). For each token in the dataset, the statistic

[5] Except for the following two attributes: CoG10 and CoG+AD. In CoG10, the window was 10 ms long and its lefthand edge corresponded to the beginning of the burst. In CoG+AD the length of the window length corresponded to the burst duration plus the aspiration duration and was centred in the middle of this selection.

classifies that token using all other tokens in the dataset as the training set. This process is repeated for every token in the dataset. The classification accuracy is the number of tokens classified correctly divided by the total number of tokens. One strength of this statistic is that it maximizes the size of the training set, which is particularly beneficial when dealing with small datasets. It will be used in this thesis for the comparison of individual attributes, and for the combined classification of small groups of attributes ($N \leq 3$). For larger groups of attributes, random forests were chosen due to their sophisticated handling of large numbers of correlated attributes (see Section 3.1.7.3 below for details).

The statistic was run in the SPSS statistical package (IBM, 2015, 2016) using the Analyze > Classify > Discriminant command, with the 'leave-one-out classification' option ticked in the 'Classification' dialogue box. These 'cross-validated' results in the output are the discriminant analysis values reported throughout this thesis. The prior probabilities were set to the default, 'All groups equal'. The 'grouping variable' was place of articulation and the independent variable was the particular attribute being examined. In addition, the box for 'Means' was ticked in the 'Statistics…' dialogue box as it provided descriptive statistics, which were used to gain an initial sense of the typical values of each place of articulation for a given acoustic attribute. All other settings in the Discriminant Analysis dialogue box were set as per the SPSS defaults. Throughout this thesis, the results for the discriminant analyses were run with these settings.

In the main study (Chapters 5-7), the 'grouping variable' was set to phoneme (rather than place of articulation) whenever either the voiced or voiceless series was being classified separately. The phoneme tier had values 0-5 (0 = /b/, 1 = /d/, 2 = /g/; 3 = /p/, 4 = /t/, 5 = /k/); thus to examine the voiced series the 'Define range…' option was given a minimum of 0 and a maximum of 2, whereas to examine the voiceless series the minimum and maximum were set to 3 and 5 respectively. When either prevocalic/non-prevocalic or front-vowel/back-vowel or schwa/non-schwa tokens were classified (as in Chapters 5 and 6), this was done by adding a variable to the 'Selection Variable' option, and specifying the relevant value (e.g. '5' for prevocalic, '7' for non-prevocalic).

3.1.7.2 Wilks's Lambda

The second statistic presented in the results is *Wilks's Lambda*, which is used to determine whether the group means (group = places of articulation) differ on a discriminant function (Cramer and Howitt, 2004: 181). At one extreme a score of 0 indicates that the means of groups differ perfectly, whereas a score of 1 indicates that the means of the groups are the same and so do not differ (ibid.). For our purposes the lower the value, the better the acoustic attribute's

performance, since it indicates that the mean values of the three places of articulation are relatively separate.

As with the discriminant analysis, this statistic was run in SPSS and is part of the default output when running the Analyze > Classify > Discriminant command. It will be used in parts of this thesis as one means of evaluating the performance of individual acoustic attributes.

3.1.7.3 Random Forests

For examining the classification accuracy of larger numbers of attributes in combination, this and later chapters employ random forests (Breiman, 2001). This machine learning algorithm is popular because it does not require much manual tuning of settings (Al-Tamimi, 2017: 15), especially in comparison to certain other approaches such as support vector machines. Furthermore, the classification accuracies yielded by the statistic are often substantially higher than approaches that rely on single trees for making classification decisions (James et al., 2013: 303).

Random forests are termed forests as they combine the output of numerous decision trees to make a consensus decision ('forest') for classifying each token. They are termed random because the trees in the forest are formed by randomly selecting subsets of the data. This randomness ensures that the trees have statistical independence from each other. Each time a subset of the predictors in the data is sampled, a decision tree is formed. This is repeated many times and the classification decision is made based on the category (in the present case = bilabial, alveolar, velar) with the largest number of trees (Liaw and Wiener, 2002).

Within each decision tree, every time a split occurs a choice is made randomly from a set of candidates. The number of candidates considered at each of these splits is by convention equal to the rounded square root of the total number of predictors in the dataset (e.g. a dataset of 17 predictors would involve a choice between four candidate predictors at each split, since the square root of 17 when rounded is 4).

The fact that random forests are designed in such a way as to prevent any predictor from being chosen in each tree might seem odd. After all, if there is a particularly good predictor that happens not to be found in the subset of predictors considered in the construction of a particular tree, this would appear to be a lost opportunity. However, as James et al. (2013: 319-320) point out, this is in fact one of the strengths of random forests. If each tree in a random forest could choose from the same set of predictors, then most of them would use the strongest predictor at the top of the decision tree. That would mean that the trees in the forest would be quite alike. But amalgamating the decisions of many closely correlated trees does not reduce the variance as much as using many uncorrelated trees (p. 320). The practical effect of this decorrelating

found in random forests is that weaker predictors have more of a voice in the decision of the forest, which makes the decision more reliable due to the reduced dependence on the strongest predictor.

Random forests require the dataset to be divided into a test set and training set. The training set ('in-bag set') is typically twice as large as the test set ('out-of-bag set'), i.e. the training set is two thirds of the total dataset, the test set one third.

The random forests were performed using the R programming language (R Core Team, 2018). Different implementations of the random forest approach can be availed of in R. The original implementation (Liaw and Wiener, 2002) was found by Strobl et al. (2007) to yield inflated estimates of the predictors' importance, especially when these predictors are heavily correlated. Because of these issues, the original implementation (viz. the R package `randomForest`) was not used in the present study; instead the function `cforest` in the package `party` was chosen (Hothorn et al., 2006).

Another decision regarding random forests is the number of trees to be used to train the model. The procedure used is outlined in Oshiro et al. (2012). This procedure grows 15 different random forests by setting `ntree` to values from 100 to 1,500 in increments of 100. Each of these forests was checked for their predictive power by using the test set as a discriminant analysis (i.e. `OOB = TRUE`; Al-Tamimi, 2017: 16). After this the area under the curve (AUC) was compared (using the `pROC` package in R; Robin et al., 2011). Then an ROC curve (which illustrates the classification model's performance under the range of conditions) for each was generated and a non-parametric Z test of significance was performed on the ROC curves. When this procedure was run on a few cases it was found that 100 trees were enough to reach the highest predictive accuracy. Hence it was decided to use N = 100 trees in all cases.

An attribute importance comparison will also be presented later in this chapter and at the end of Chapter 6. This provides a sense of which attributes contribute to the classification accuracy the most by quantifying the mean loss in classification accuracy when each attribute is removed from the random forest. This supplements the information on individual-attribute performance given by the discriminant analysis results. The attribute that contributes the most is the one whose absence from the random forest results in the largest decrease in classification accuracy. This was done with an AUC-based estimation of the variable importance, i.e. `party`'s `varimpAUC` function and `conditional = T` (the procedure detailed throughout this section follows closely that of Al-Tamimi, 2017: 16; for full details, see the specification in Strobl et al., 2009). The description provided above applies not just to the random forests used in this chapter but also those found in Chapters 6 and 7.

3.2 Burst Attributes

The aim of this section is to compare the performance of various burst attributes relative to each other. Most of these attributes were chosen from the results of previous studies, especially Stevens et al. (1999) and Suchato (2004). However the current study includes the attribute AllPeakHz, which was not examined in those studies. Also, those studies did not examine the accuracy of certain attributes in isolation, such as HiPeak and MidPeak (termed 'Ahi' and 'Amax23' in those studies). These gaps will be addressed.

3.2.1 Description of Burst Attributes

The attributes are as follows (all amplitudes and energies are in dB using the reference level found in Praat 5.4.03's default intensity settings):

1. CoG

This is the burst's centre of gravity, given by the following formula:

$$CoG = \frac{\sum af}{\sum a}$$

This acoustic attribute consists of multiplying the amplitude of each spectral component (*a*) by its frequency (*f*), summing these products, and dividing this sum by the sum of the said components' amplitudes. The amplitudes may optionally be squared, which yields the intensities and has been used in previous studies of plosive place of articulation such as Suchato (2004: 40) on the grounds that it increases the prominence of the spectrum's peaks, which he believes to be particularly informative for place of articulation. It is also followed in the present study, both in the pilot and main study.

Centre of gravity was measured in the following ways: on the burst (CoG+BD) but also on the burst plus aspiration (CoG+AD) as well as on the first ten milliseconds after the beginning of the burst (CoG10). This is a replication of Suchato's work (as noted in 2.3.1, he found CoG10 to be the strongest of the three).

2. CoG10

The centre of gravity was measured not just from the entire burst, but also for a segment of the burst corresponding to its first 10 ms.

3. CoG+AD

In a third measure of the centre of gravity, the centre of gravity was calculated on the burst plus the following aspiration. The difference between burst CoG+AD, CoG10 and CoG is not expected to be large, especially since the aspiration in the burst is relatively short in /b d g/.

4. SD

This is the burst's standard deviation (roughly, how flat or peaky the spectrum is).

$$SD = \sqrt{\frac{\sum a[(f - CoG)^2]}{\sum a}}$$

5. Skew

This measures the skewness (how lopsided the spectrum is) and is expected to yield similar results to CoG (as noted in 2.3.1.8).

$$Skew = \frac{\sum a[(f - CoG)^3]}{[\sum a].[SD^3]}$$

6. Kurt

This measures the burst's kurtosis, which compares the size of the tails relative to the rest of the spectrum.

$$Kurt = \frac{\sum a[(f - CoG)^4]}{[\sum a].[SD^4]} - 3$$

7. AllPeakHz

This measures the frequency of the highest intensity component of the burst in the region from 750 to 8,000 Hz.

8. HiPeakdB

This measures the amplitude of the spectral component with the greatest amplitude in the burst's high-frequency region (which extends from 3,500 to 8,000 Hz). Suchato (2004) did not present the performance of this attribute on its own (instead he presented its performance as part of HiPeak-F1(dB) and HiPeak-MidMean(dB) only) so this fills that research gap. That is, comparing the performance of this attribute with HiPeak-F1(dB) and HiPeak-MidMean(dB) – attributes 13 and 14 below – reveals to what extent subtracting the F1 amplitude or mean mid-frequency improves the attribute's classification accuracy.

9. HiTotaldB

This measures the total amplitude of the spectral components in the burst's high-frequency region. Again, Suchato (2004) did not present the performance of this attribute as a standalone attribute, instead only presenting it as part of the TiltTotaldB attribute.

10. MidPeakdB

This measures the amplitude of the spectral component with the greatest amplitude in the burst's mid-frequency region (which extends from 1,250 to 3,000 Hz).

11. MidTotaldB

This measures the total amplitude of the spectral components in the burst's mid-frequency region. See comments for attribute 9 above regarding Suchato (2004).

12. MidMeandB

This measures the mean amplitude of the spectral components in the burst's mid-frequency region. See comments for attribute 8 above regarding Suchato (2004).

13. HiPeak-F1(dB):

This subtracts the F1 amplitude at the onset of voicing (see 3.1.4 for definition) from the amplitude of the spectral component with the greatest intensity in the burst's high-frequency region (i.e. Attribute 8 above). This attribute is the same as the attribute that Stevens et al. (1999) and Suchato (2004) termed 'Av-Ahi'. (Its performance will be examined in the main study in greater detail, for which see 6.4.7.)

14. HiPeak-MidMean(dB):

This compares the amplitude of the most intense component in the high-frequency region of the burst (attribute 8) with the mean amplitude of the burst's mid-frequency region (attribute 12) by subtracting the latter from the former. The attribute is the same as Stevens et al.'s (1999) and Suchato's (2004) attribute 'Ahi-A23'.

15. MidPeak-F1(dB):

This compares the amplitude of the highest-amplitude peak in the mid frequency region of the burst (attribute 10) with the F1 amplitude at the following vowel's onset by subtracting the former from the latter. This attribute is the same as Suchato's (2004) attribute 'Av-Amax23'.

16. TiltTotaldB:

This compares the energy in the high-frequency region of the burst (Attribute 9) with the energy in the mid-frequency region (attribute 11), by subtracting the latter from the former. This attribute is the same as Suchato's (2004) attribute 'Ehi-E23'. (Its performance will also be evaluated in the main study, where it will be compared with other methods of measuring spectral tilt, for which see 6.4.5.)

3.2.2 Results

These are the results for the discriminant analysis:

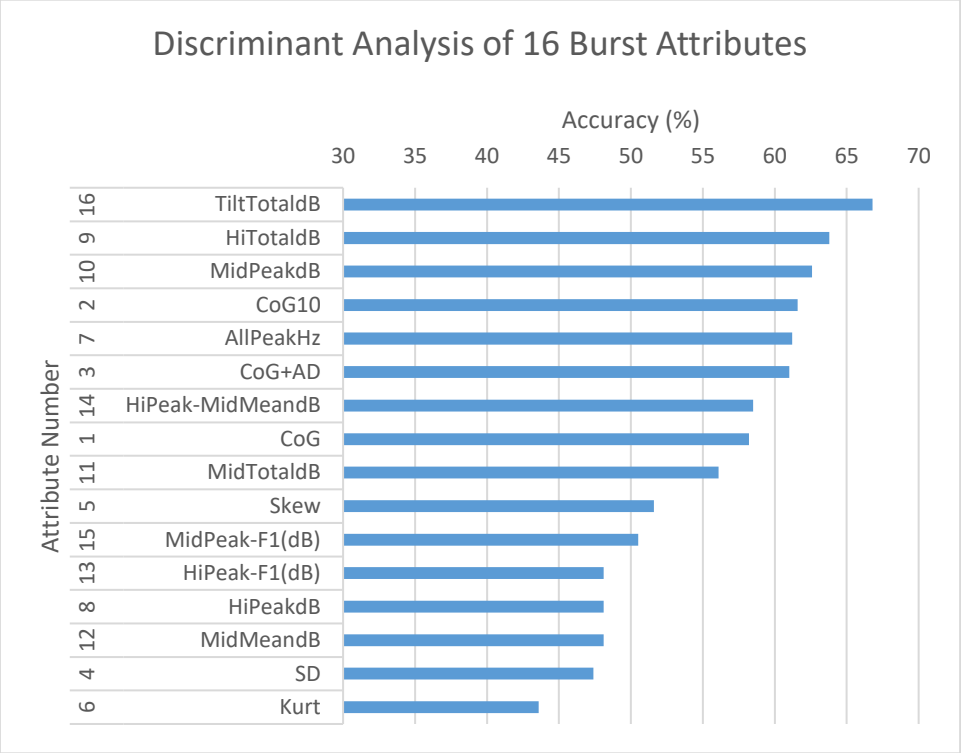


Figure 3.3: Discriminant analysis classification accuracy of the 16 burst attributes at distinguishing the place of articulation of /b d g/.

N = 644.

And here are the results for Wilks's Lambda:

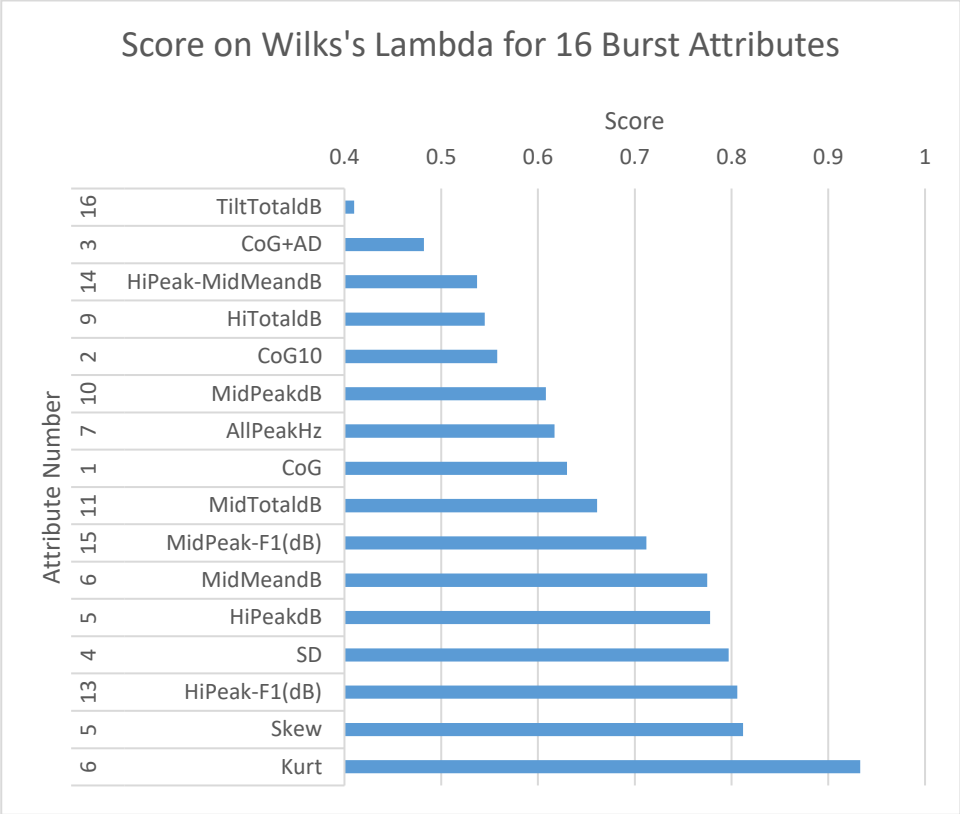


Figure 3.4: Score on Wilks's Lambda for the 16 burst attributes at distinguishing the place of articulation of /b d g/.
N = 644.

As expected, the attributes that have a high score on the discriminant analysis have a low score on Wilks's Lambda. Given that plosive place of articulation consists of a three-way phonemic contrast, an attribute scores worse than chance on the discriminant analysis if its classification accuracy is below 33.3%. All of the above burst attributes surpass this threshold. Nevertheless, there are five attributes that classify below 50%, namely kurtosis, standard deviation, HiPeak-MidMeandB, HiPeakdB, and MidMeandB. On Wilks's Lambda, there are three attributes whose scores are greater than 0.8 (i.e. relatively poor), namely HiPeak-F1(dB), standard deviation, skewness, and kurtosis.

On the discriminant analysis, the strongest attribute is TiltTotaldB with a classification accuracy of 66.8%. Next highest is HiTotaldB (63.8%), followed by MidPeakdB (62.6%), CoG10 (61.6%), and AllPeakHz (61.2%). Suchato (2004) also found TiltTotaldB to be the strongest burst-based attribute under some of the conditions he examined; under other conditions CoG10 was strongest. Suchato did not examine the other spectral moments (namely standard deviation, skewness, and kurtosis) but it is clear that these are among the weakest attributes. It was noted in the literature review that skewness and kurtosis have been argued to

yield little or no extra information beyond what the first two moments provide (Wheeler, 2004: 56).

On Wilks's Lambda, the lowest (= best) scorer is TiltTotaldB (0.41), followed by CoG+AD (0.482) and HiPeak-MidMeandB (0.537).

Overall the results suggest that some kind of measure of spectral tilt is the strongest measure for the release burst. Nevertheless, there were five other attributes whose classification accuracies were within 6 percentage points of this attribute, including centre of gravity and peak, which are quite different attributes both from TiltTotaldB and from each other: centre of gravity is computed using every single spectral component as input, whereas peak picks out the frequency of a single component (the loudest one). Given this diversity of attributes with reasonably strong results, a variety of burst attributes will be brought forward to the main study (see Section 6.4.1 for further details).

3.3 Formant Attributes

In this section the classification accuracy of a variety of formant attributes is compared. We begin by introducing each of the attributes and the rationale for their development.

3.3.1 Description of Formant Attributes

Non-Compound Attributes

1. F2off:

This measures the F2 frequency at the offset of the vowel before the plosive. It is expected to be lowest on average for bilabials, highest for front-vowel velars, with alveolars and back-vowel velars intermediate.

2. F3off:

This measures the F3 frequency at the offset of the vowel before the plosive. This is expected to be higher on average for alveolars than for velars and bilabials.

3. F2on:

The F2 frequency at the onset of the vowel after the plosive (i.e. at the onset of voicing). This is expected to have high values for front-vowel velars, low values for bilabials, with alveolars and back-vowel velars intermediate.

4. F3on:

The F3 frequency at the onset of the vowel after the plosive. Values are expected to be the same as those described for attribute 2 above.

Compound Attributes

5. F3on-F2on:

The F3 frequency at the onset of the vowel after the plosive minus the F2 frequency at the same moment in time. This is expected to be smallest for velars due to 'velar pinch' with larger values for alveolars and bilabials.

6. F3on20-F2on20:

The F3 frequency at 20 ms after the onset of the vowel after the plosive minus the F2 frequency at the same moment in time. As with attribute 5, the highest values are expected for velars with lower values for alveolars and bilabials.

7. F3off-F2off:

The F3 frequency at the offset of the vowel before the plosive minus the F2 frequency at the same moment in time. The expectations for each place of articulation are the same as for (5) and (6).

8. F2off-F2mid:

The F2 frequency at the offset of the vowel before the plosive minus the F2 frequency at the steady part of the same vowel. Based on inspecting Öhman (1966: 160-162), bilabials seem to have more falling transitions (= transitions in which the formant frequency is lower at the part nearest to the plosive than at the vowel midpoint); therefore we expect bilabials to have lower values on average for this attribute than the other places of articulation.

9. F2on-F2mid:

The F2 frequency at the onset of the vowel after the plosive minus the F2 frequency at the steady part of the same vowel. The theoretical expectations for each place of articulation are the same as those for attribute 8.

10. F2off-F2midV2:

The F2 frequency at the onset of the vowel before the plosive minus the F2 frequency of the steady part of the vowel after the plosive. This attribute was originally inspired by a comparison of Öhman's (1966: 160-162) schematized diagrams of [ybo ydo ygo], which show that the difference in F2 frequency between V1end and V2mid is smaller for bilabials than for alveolars. Thus this attribute is expected to have smaller values for bilabials than for alveolars.

11. F2off-F2on:

The F2 frequency at the offset of the vowel before the plosive minus the F2 frequency at the onset of the vowel after the plosive. This attribute arose from the observation in Öhman's (1966) diagrams that the difference in frequency between V1end and V2onset is tends to be less for alveolars than for other places of articulation. This is related to the finding from Delattre et al.

(1955) that the locus theory worked best for alveolars but less well for the other places of articulation, because $F2_{onset}$ in /d/ varies less than in /b g/.

12. F2off-F2off20:

The F2 frequency at the offset of the vowel before the plosive subtracted from the F2 frequency 20 ms earlier. The expected pattern for the three places of articulation is the same as for attribute 8 above.

13. F2on-F2on20:

The F2 frequency at the onset of the vowel after the plosive minus the F2 frequency 20 ms later. The expected pattern for the three places of articulation is the same as for attribute 9 above.

14. F3on-F2on-(F3on20-F2on20):

This subtracts attribute 5 from 6. The attribute is another attempt to capture the velar pinch. The development of this attribute was prompted by the observation of certain cases of /g/ in Öhman's diagrams – such as [øgɑ øgo] – in which F3on-F2on is smaller than F3on20-F2on20 but without the close approximation of F2 and F3 that is considered typical of the velar pinch. Negative values are expected to indicate some degree of velar pinch.

15. F3off-F2off-(F3off20-F2off20):

This is functionally the same attribute as attribute 14 except that rather than being applied to the vowel after the plosive it is applied to the vowel before it. Negative values are again expected to indicate some degree of velar pinch.

16. F2off-F2on-(F2midV1-F2midV2):

This attribute contains some of the same input as attribute 11 and was inspired by the same observation, namely a comparison of Öhman's [ydo ybo]. This comparison suggests that bilabials should have a lower value for the present attribute than alveolars. That is, the frequency difference between V1end and V2start relative to V1mid-V2mid is smaller in [ybo] than in [ydo]. The main question regarding the attribute, as with attribute 11, is whether it can perform well in general rather than just in this particular context of preceding back vowel and following front vowel.

17. F2on-F2mid-(F2off-F2midV1):

This attribute is a compound of attributes 9 and 10. It is a means of comparing the amount of frequency change in the CV transition to the VC transition. Its formulation was prompted by the observation in Öhman's (1966: 160-161) diagrams of sequences such as [oby ody] in which the amount of frequency change in an alveolar transition can be larger than that in a bilabial transition in the same vocalic context.

18. F2off-F2midV1-(F2midV1-F2midV2):

This attribute is a means of comparing the amount of frequency change between the offset of a vowel preceding the plosive and the midpoint of the following vowel, relative to the frequency change between the first and second vowels. The observations that led to the formulation of this attribute are the same as those detailed in attributes 10 and 17 above.

19. F2on-F2mid-(F2midV1-F2midV2):

This attribute is similar to attribute 18 above and was prompted by the same observations.

20. F2off-F2off20-(F2on-F2on20):

This attribute is a means of comparing the amount of frequency change in the VC transition to that in the CV transition. As such it is analogous to attribute 17 above.

21. F2on-(F2mid-F2on), also known as $F2_R$:

This attribute, which will be introduced in detail in Section 5.3, is inspired by the 1950s locus theory presented in the literature review. It is expected to yield relatively low values for bilabials, high values for front velars, with alveolars and back-velars intermediate. Whenever $F2_{onset}$ is *higher* in frequency than $F2_{mid}$, this attribute yields an F2 value that is even higher than $F2_{onset}$. Conversely, whenever $F2_{onset}$ is lower in frequency than $F2_{mid}$, the attribute yields an F2 value that is even lower than $F2_{onset}$. This is how the attribute functions like $F2_{locus}$ (cf. Figure 2.11 on page 24). Note also that the degree to which $F2_{onset}$ is modified by the formula varies depending on how large the difference in frequency is between $F2_{onset}$ and $F2_{mid}$ (again this is in accordance with the locus theory, as can be seen in Figure 2.9).

3.3.2 Results

As with the burst attributes, the performance of each of the formant attributes will be evaluated using discriminant analysis and Wilks's Lambda. Here are the results for the discriminant analysis:

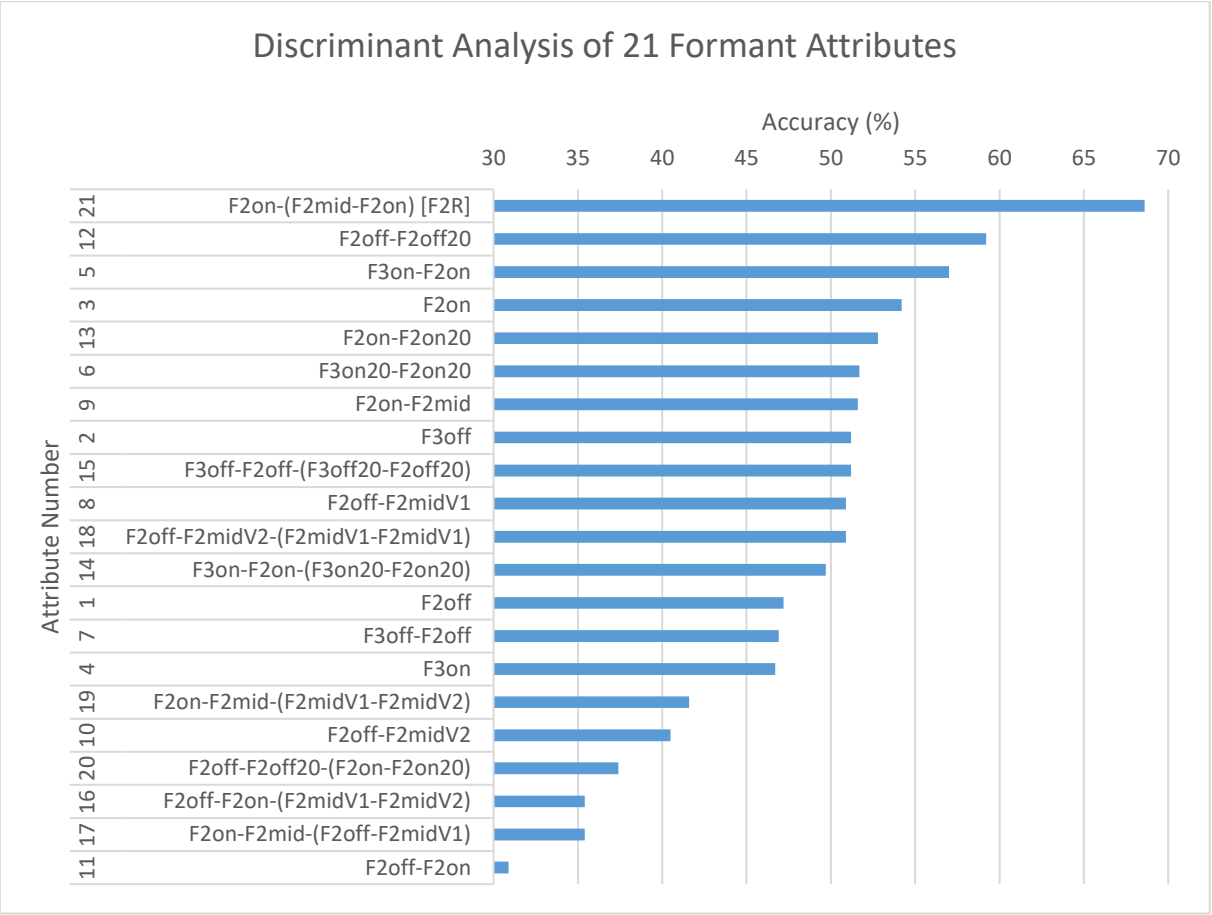


Figure 3.5: Discriminant analysis classification accuracy of the 21 formant attributes at distinguishing the place of articulation of /b d g/.

N = 644.

And here are the results for Wilks's Lambda:

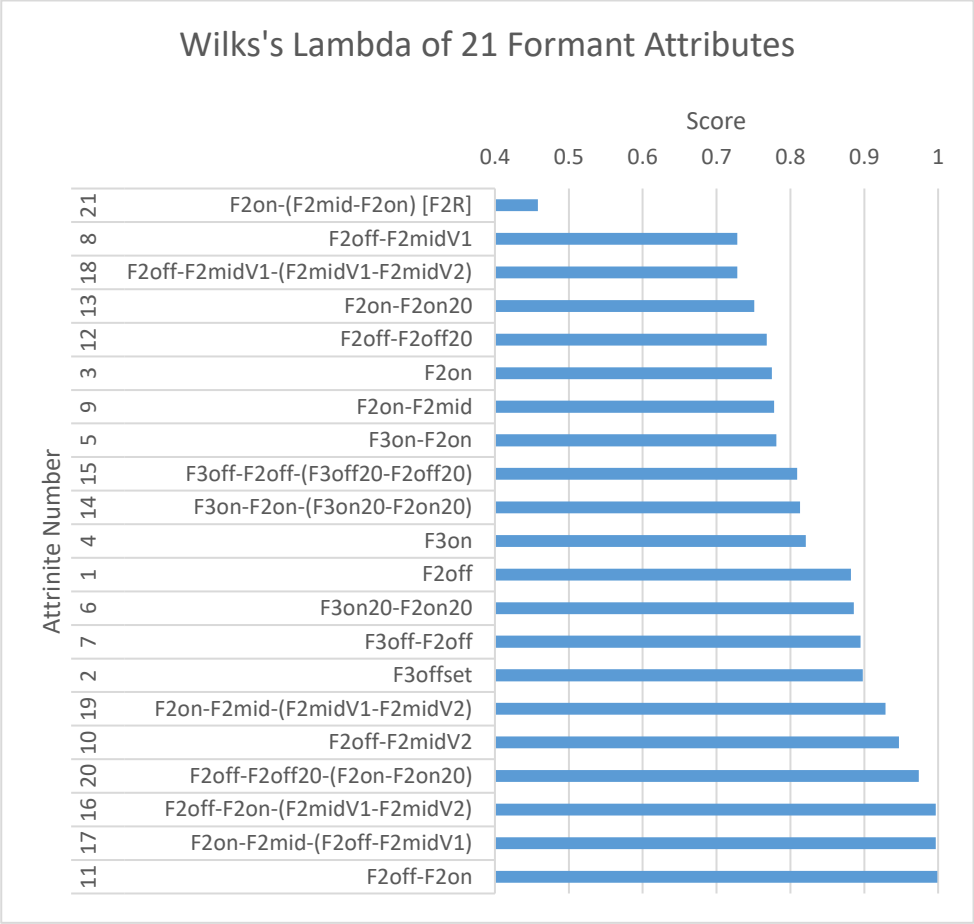


Figure 3.6: Score on Wilks's Lambda for the 21 formant attributes at distinguishing the place of articulation of /b d g/.
N = 644.

As with the burst attributes, the attributes that score highly on the discriminant analysis tend to score lowly on Wilks's Lambda. Given that plosive place of articulation consists of a three-way phonemic contrast, an attribute scores worse than chance on the discriminant analysis if its classification accuracy is below 33.3%. This occurs for one of the attributes, F2off-F2on (attribute 11). However, there are five further attributes that also score poorly, i.e. with a discriminant analysis accuracy below 45% and a Wilks's Lambda above 0.9, namely:

- F2off-F2midV2 [attribute 10]
- F2off-F2on-(F2midV1-F2midV2) [attribute 16]
- F2on-F2mid-(F2off-F2midV1) [attribute 17]
- F2on-F2mid-(F2midV1-F2midV2) [attribute 19]
- F2off-F2off20-(F2on-F2on20) [attribute 20]

Most of these attributes are the most complicated attributes, i.e. the ones that take as input four measurements.

On both the discriminant analysis and Wilks's Lambda there is one attribute that performs far beyond the others, namely F2on-(F2mid-F2son) (to be termed '$F2_R$' for short). Its discriminant analysis classification accuracy is 68.6%, which is 9.4 percentage points higher than the nearest alternatives: F2off-F2off20 (59.2%), F2on-F2on (57.0%), and F2on (54.2%). On Wilks's Lambda, F2on-(F2mid-F2on) scores 0.458, which is again better than the next best attributes by a huge margin (0.27), namely F2off-F2midV1 and F2off-F2midV1-(F2midV1-F2midV2).

Given that the performance of F2on-(F2mid-F2on) greatly surpasses that of all the other formant-based attributes, it will be explored in detail in the main study (Chapter 5), where it will be referred to by the more convenient name of $F2_R$ ('R' being an abbreviation of 'reconstructed', reflecting the hypothetical nature of this F2 frequency).

3.4 Combining Attributes

As described in Section 3.1.7, for large combinations of attributes random forests are employed. The classification accuracy when 16 attributes from the burst are combined is 82.3% with an r^2 of 0.82, indicating that the random forest model fits the data relatively well. Here are the results for the relative contribution of each attribute to the classification accuracy, measured as the mean decrease in classification accuracy when each attribute is removed from the model:

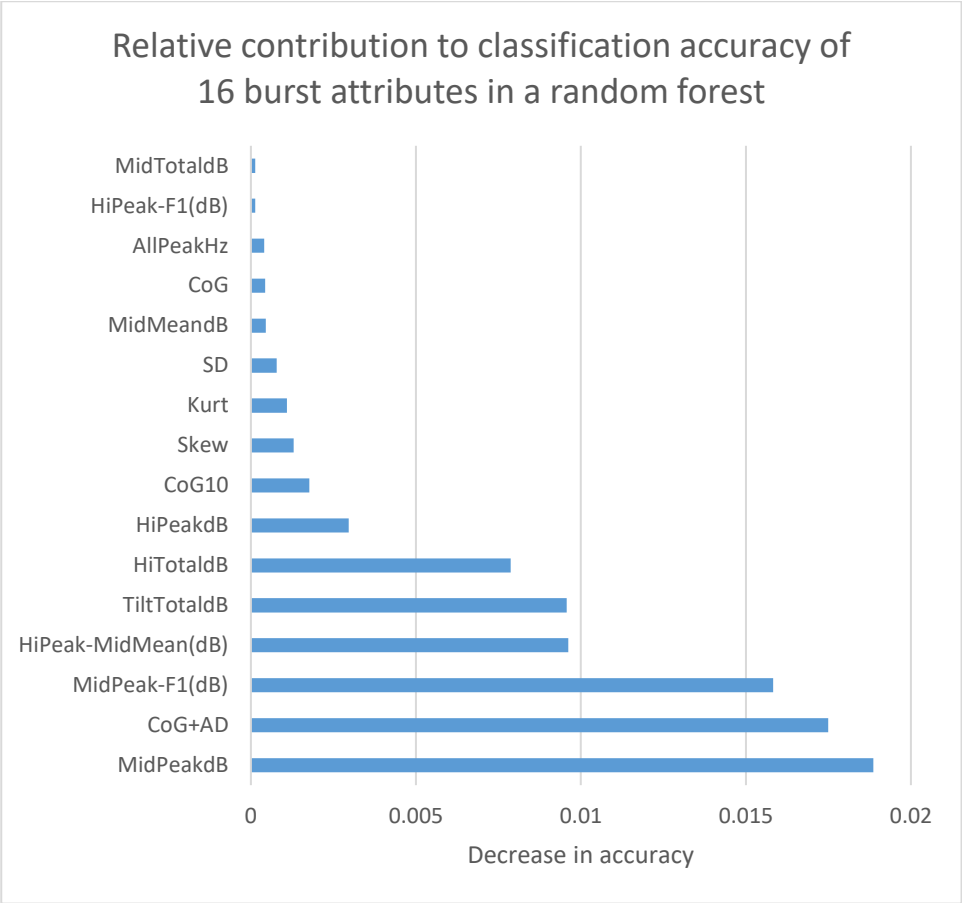


Figure 3.7: Decrease in classification accuracy when each of the 16 burst-based attributes is removed from the random forest.
N = 644.

The results show MidPeak as being the attribute that contributes the most to the classification accuracy. This attribute, recall, measures the amplitude of the peak in the mid-frequency region (between 1,250 and 3,000 Hz). Attributes that contribute almost as much are CoG+AD (which measure the burst and aspiration's centre of gravity) and MidPeak-F1(dB) (which gauges the amplitude of the mid-frequency peak relative to the F1 amplitude at the onset of the following vowel).

The attributes that were strongest on the discriminant analysis are not as strong on the random forest, e.g. HiTotaldB and TiltTotaldB are only in fourth and fifth. AllPeakHz, which was fifth strongest on the discriminant analysis, is now only thirteenth strongest. It is difficult to be sure why there should be such discrepancies between the two statistics but the small size of the present pilot dataset is a possible factor. Another possible factor is the design of random forests: as outlined earlier, it deliberately allows weaker attributes to play a larger role in the decision of the forest by the fact that each decision tree can only choose from the rounded square root of the total number of attributes when picking each attribute to include in the tree.

We now turn to the results from the formant attributes. The combined classification accuracy for the 21 of these was 94.0%. The model's $r^2 = 0.93$, which indicates a close fit to the data. In terms of the relative contribution of each attribute, here are the results:

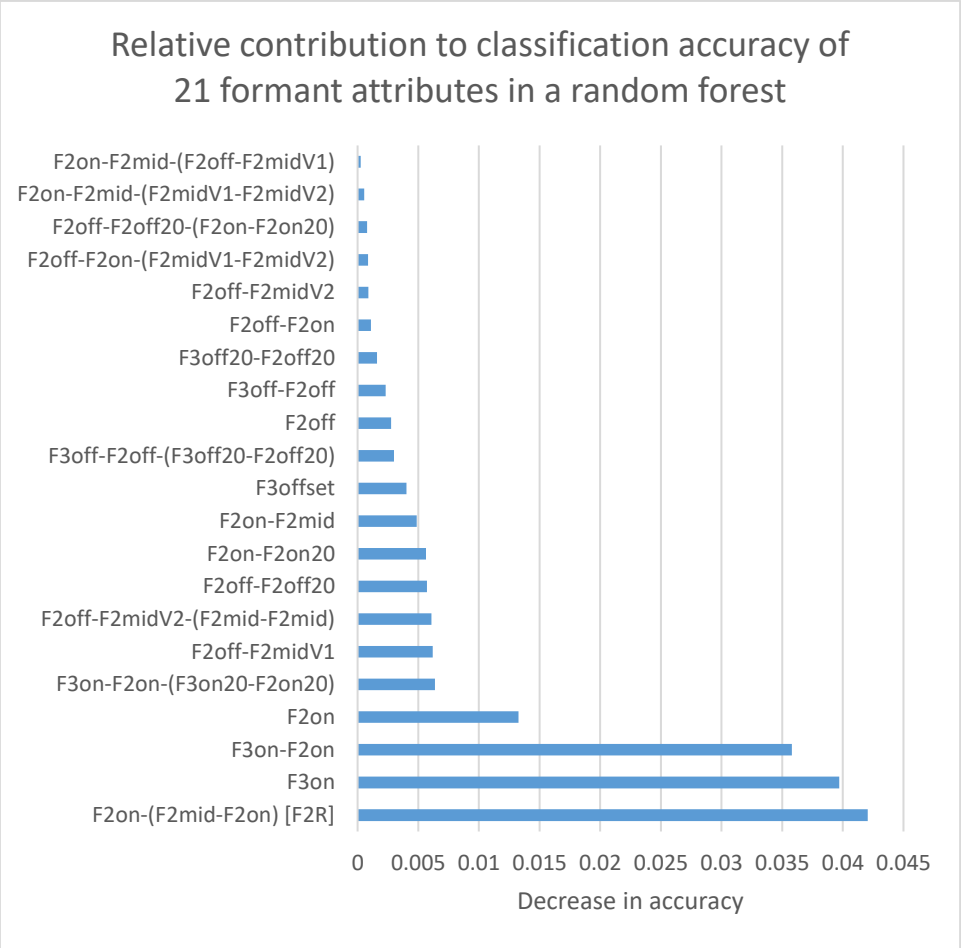


Figure 3.8: Decrease in classification accuracy when each of the 21 formant-based attributes is removed from the random forest.

N = 644.

The strongest four attributes on the discriminant analysis are the same four attributes on the random forest and in the same order. F2on-(F2mid-F2on) (i.e. $F2_R$) is the largest contributor to the classification accuracy, followed closely by F3on. F3on-F2on, which is a measure of velar pinch and which was third in the discriminant analysis, is again in third and contributes almost as much to the classification as the first two attributes. All the remaining 18 attributes contribute little to the classification. The weakest attributes in the discriminant analysis are once again found to be weakest in the random forest.

In the final part of the classification, a random forest was run consisting of all 37 attributes. The classification accuracy yielded was 90.7% with an r^2 of 0.84 between the random forest model and the dataset. When the relative contribution of the attributes was examined, the five attributes that contributed the most to the classification were as follows:

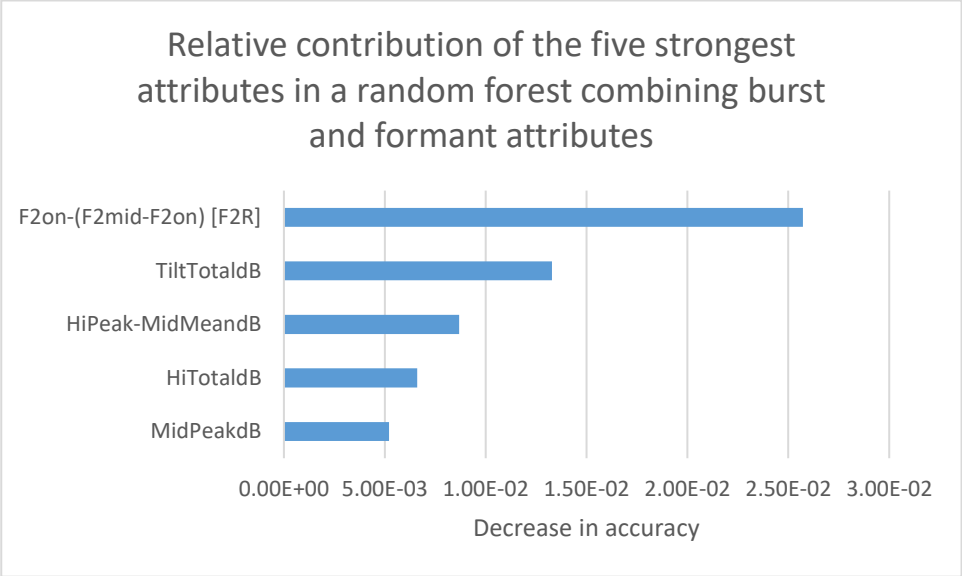


Figure 3.9: Decrease in classification accuracy for the five best-performing attributes in a random forest in which formant and burst attributes were combined.
N = 644.

The results accord broadly with those of the discriminant analysis: F2on-(F2mid-F2on) (i.e. $F2_R$) was the strongest attribute on the discriminant analysis for the formant attributes, and HiTotaldB was the strongest attribute on the discriminant analysis for the burst attributes.

3.5 Discussion

There are a few limitations to the pilot study. The largest one is its statistical power: it consists of the speech of just two speakers. This is addressed by the main study, which consists of the speech of 20 speakers. Another limitation is the artificial nature of the data: nonce VCV sequences. On the one hand these sequences have the advantage of ensuring that the acoustics of plosives adjoining a wide variety of vowel backnesses are tested. On the other hand, because of the semantic emptiness of the words, the speakers tended to read the sequences slowly and carefully, which reduces the amount of coarticulation relative to natural speech. Also, the material involved stressed vowels, so the acoustics of plosives next to schwa was not investigated. In the main study, speakers read semantically meaningful sentences made from actual words of English.

Another limitation of the pilot study was in the choice of formant attributes. In spite of the high classification accuracy when combined in the random forest, quite a large proportion of the formant attributes performed poorly as individual attributes, with six having a discriminant analysis classification accuracy below 45%. The theoretical rationale for many of

these attributes was based on observing in Öhman's (1966) data the formant transitions of a particular place of articulation in a specific vowel context rather than across all contexts. For example it was noted in Section 3.3 that attributes 16 and 17 were developed based on observations in which the plosive was flanked on one side by a front vowel and on the other side by a back vowel. While these attributes may well have classified well in these very specific contexts, ultimately what is useful are attributes that work well in a *variety* of contexts. Of the formant attributes, only F2on-(F2mid-F2on) (i.e. $F2_R$) appears to have done this convincingly, with a discriminant analysis accuracy above two thirds correct.

Another issue with some of the formant attributes (such as attributes 16 and 17) is that they relied on comparing a VC transition to a CV transition. But as we shall see in Section 5.1, the proportion of natural-speech English plosives in which there is both a preceding and following vowel is in the minority (approximately a quarter). Such attributes would not be available whenever one or both surrounding vowels is missing. Given that such cases are in the majority, it raises the question of whether such attributes are in fact worthwhile.

For the burst attributes, one limitation was that the attributes were all derived from a spectrum with linear frequency and logarithmic amplitude. (We shall term this the 'Hz-dB' spectrum.) As was outlined in 2.3.1, phonetics and ASR have tended to use different spectral representations and Aim 3 of the present study is to investigate whether the Hz-dB, Bark-phon, and Bark-sone spectral representations affect the accuracy of burst attributes. For example, it could be the case that the burst peak derived from a Hz-dB spectrum would be at a different frequency from the burst peak that would be found with a more auditorily-oriented spectrum, with knock-on effects for the classification accuracy. Thus Chapter 6 will compare the performance of attributes derived from the Hz-dB spectrum with equivalent attributes derived from the Bark-phon and Bark-sone spectra, with a particular interest in the performance of the AllPeak attribute, given the importance of the peak for plosive perception that was established by Liberman et al. (1952) and by Allen and coworkers (Li et al., 2010; Kapoor, 2010; Xi, 2013).

The present study found that although the formant attributes individually tended to have a lower classification accuracy than the burst-based attributes, when combined they yielded a classification accuracy of 94% as against 82% for the burst-based ones. However, the design of the present dataset favoured the formant attributes, because it consisted exclusively of plosives flanked on both sides by vowels. In the natural data used in the main study, most of the tokens are not of this kind. Thus in natural data there would not be the abundance of formant information for most tokens as there was in the pilot data.

Furthermore, the main study will include the voiceless series /p t k/; in such consonants voicing does not normally begin until several tens of milliseconds after the release of the burst

due to the presence of aspiration. Consequently the formant information is expected to be less informative on place of articulation in this type of plosive.

3.6 Conclusion

In this chapter the ability of a variety of burst and formant attributes to distinguish the place of articulation of /b d g/ has been examined. The results of the discriminant analysis identified F2on-(F2mid-F2on) (i.e. $F2_R$) as the strongest attribute, with a classification that was 9.4 percentage points above the next best formant attribute. The results of the random forest indicated again that this attribute contributed the most to the classification, with F3on contributing the next largest amount.

With regard to the burst attributes, the discriminant analysis did not reveal a clearly superior attribute as was the case with the formant attributes. The strongest attribute was HiTotaldB (found to be the second strongest attribute by Suchato (2004)) but CoG, AllPeakHz, and MidPeakdB were also relatively strong (exceeding 60% accuracy). The spectral moments SD, Skew, and Kurt were relatively weak.

Chapter 4: Methodology

This chapter introduces the main study and summarizes the methods chosen for recording, annotating, extracting, and analysing the data.

4.1 Theoretical Motivation

The overall aim of the main study is to investigate certain aspects of the acoustics of place of articulation in English plosives that have not hitherto been examined by most previous studies. Within this overall aim are contained four specific aims:

1. To test the performance of a technique for collapsing $F2_{onset}$ and $F2_{mid}$ into a single attribute, termed $F2_R$. The development of this technique has been inspired by the observation that $F2_{onset}$ and $F2_{mid}$ tend to be closely correlated (the slope in a regression plot between the two typically being between ca. 0.4 and 0.75). See Section 4.4 for information on the formant information measured, and Chapter 5 for the results.
2. To compare the strength of burst-based attributes at distinguishing place of articulation with and without normalization by individual speaker. This has been done because normalization by individual speaker of formant frequencies has been in widespread use whereas normalization of aperiodic events such as the burst seems to have been less widely studied. See Section 4.5 for burst measurements, 6.3 for the burst attributes, and 6.4.4 and 6.5 for the results of the normalization.
3. To examine the effect of different spectral representations (Hz-dB, Bark-phon, and Bark-sone) on the accuracy of the burst attributes. See 6.4.1 and 6.5 for the results of the comparison.
4. To compare the performance of some traditional burst-based attributes with the first 12 coefficients of the discrete cosine transform (DCT). As we saw in the literature review, phonetic science has a long tradition of developing burst attributes that are tailored to the specific task of extracting place-of-articulation information from the burst, whereas automatic speech recognition (ASR) has long used attributes that function on all kinds of acoustic speech material, and which are theoretically expected to capture more of the variance in the burst than the attributes that have been traditionally used in phonetic science. Hence the DCT features can serve as a benchmark against which to compare the performance of other burst features that have been developed in the history of phonetics. See Section 7.3 for the results.

4.2 Data Collection

4.2.1 Participants

There were 20 speakers in total, 10 of whom were male, 10 female. At the time of recording (April and May 2016) they ranged in age from late teens to late thirties. They came from various locations in Britain: Tyneside (2), Sunderland (1), County Durham (1), Middlesbrough (2), Lancashire (1), South Yorkshire (2), Greater Manchester (2), Merseyside (1), Warwickshire (1), Cambridgeshire (1), Hertfordshire (1), Greater London (1), Hampshire (1), Gloucestershire (2), and Conwy County Borough (1). One of the speakers spent the first 14 years of his life in Florida and shows occasional (southern) US features, e.g. more frequent pre-voicing of /b d g/.

Three of the speakers can be regarded as RP speakers. The rest spoke regional accents to varying degrees. If speakers of only one particular accent had been chosen, then there would be uncertainty as to whether the results would be generalizable across plosives from different accents. The aim was to have a dataset that was representative of the real-life task of recognizing place of articulation from a wide variety of accents.

4.2.2 Recording

Each participant was recorded in an anechoic booth at Newcastle University. Care was taken to have the microphone positioned at approximately the same location and angle vis-à-vis the lips for each speaker, i.e. at a distance of approximately six inches and an angle of approximately 10°. Given that the speakers varied considerably in height, this distance between the microphone and the speaker's lips was held constant across speakers by placing the microphone on a stack of books: the taller the speaker, the larger the number of books placed in the stack.

The speakers were instructed to remain reasonably still whilst reading in order to keep the acoustic conditions as consistent as possible within and between subjects. They were asked to read the sentences at a normal rate, i.e. not too slowly and not too quickly. Nevertheless most speakers read the first few sentences more slowly than subsequent ones, which is to be expected when faced with a new task. Also, a few of the speakers read noticeably more quickly or more slowly than most of the other speakers. Thus in terms of tempo the recorded data are somewhat variable.

Sentences were presented one at a time on a monitor in front of the speaker. The speakers were given a short break after reading half the sentences and another short one before describing the images.

The recorder used was a Roland Edirol R-44 4-channel Portable Recorder, linked to a Roland Edirol CS-50 microphone. The microphone was set to "lo cut" rather than "flat" to reduce the acoustic influence of very low frequencies. For the extent of directional response,

the setting “focus” was used rather than “flat”, which is appropriate for when there is only one speaker using the microphone since it focuses the microphone’s response to sound sources that are close to itself. The sampling frequency was 44.1 kHz and the quantization was 16 bit, with a mono channel. Recordings were saved as .wav files.

The above procedures were intended to minimize those variations in acoustics between speakers that are not related to place of articulation. In reality there were still differences between the speakers, the most important of which is that some speakers spoke louder than others. The significance of this is that it makes amplitude-based acoustic attributes more difficult to compare across speakers than would be the case if the speakers had spoken at the same loudness as each other. Sections 6.4.4-6.4.7 investigate a variety of methods for dealing with between-speaker acoustic variability in burst attributes, while Section 5.4.2 does the same for formant attributes.

4.2.3 Material

The material consisted of two parts: a set of 84 sentences to be read, and a set of five images to be described (see Appendix 1). The subject matter of the sentences was various everyday topics; nevertheless, the aim was to include as many plosive consonants as was reasonable, and so the sentences contain an average of ca. 14 plosives each.

The participants were asked to describe a series of images to obtain a sample of spontaneous speech. Spontaneous speech is thought to be more rapid and to contain more elisions and other connected-speech processes than read speech. However, the image description task did not manage to elicit this level of spontaneous speech; instead it tended to contain pauses and to be at times slow-paced, due to the speakers looking at the images and figuring out what to say about them. Thus the spontaneous speech material tended to be less fluent and less rapid than the read speech, the reverse of what was intended. A spontaneous conversation would have probably elicited more of the desired colloquial style. Thus the image description material was not analysed as it was not felt to be sufficiently rapid and colloquial to be more acoustically challenging than the read speech.

4.3 Annotation and Transcription

From each speaker a total of 30 sentences was annotated. These were annotated 10 at a time, e.g. the first 10 sentences from one speaker would be annotated and then the next 10 would be annotated from another speaker. This process was repeated three times from the 20 speakers, yielding a total of 600 sentences. Manual annotation was chosen rather than automatic or semi-automatic annotation, for three reasons:

(1) It allows for greater precision in the segmentation of the burst, which seemed warranted since it was decided that the burst would be segmented into its transient and frication to allow for the possibility of separate acoustic analysis (see 4.3.5 for details).

(2) Previous studies (e.g. Blumstein and Stevens, 1979; Suchato, 2004; Modarresi et al. 2005) have attempted to measure F2 in the burst. However, in the present study inspection of spectrograms indicated that burst tokens are in fact highly variable in terms of how present F2 is. Thus it was decided to use the occasion of segmentation to rate (on a five-point scale) exactly how present F2 and F3 were on the spectrogram in each burst token (see 4.3.5).

(3) Manual annotation can be a valuable source of hypotheses about how reliable particular cues are likely to be. In addition to the case of burst formants mentioned in (2) above, manual annotation can also be used to make notes on anomalous cases, i.e. plosives with atypical acoustic patterns. Anomalous cases are worth noting precisely because of their challenge for existing theories of place of articulation. Such cases can form the basis for future improvements in methodology and understanding of the complexities involved in the acoustics of place of articulation (see 8.6.2 and 9.2 for examples).

Subsequent to annotation all tiers were checked for errors. Further annotation errors were detected and corrected using the search tool in the Excel file of the extracted data.

Whilst annotating the data, the spectrogram's display was pre-emphasized (which is the default in Praat) and the dynamic range was set to 50 dB. It was occasionally difficult to see whether a bilabial plosive contained a transient, and in such cases the range was temporarily raised to 70 dB to see if it was in fact present. The length of the selection in the screen (i.e. the zoom) was held fairly constant. A variety of zooms were used as necessary; increased zoom was usually only necessary for identifying with precision the beginning of the plosive's transient in the waveform.

As for the transcription, this consisted of a broad and narrow transcription, as described below. Detailed information about the transcription policy adopted in the present study can be found in Appendix 4.

4.3.1 Exclusion of Tokens

As mentioned above there was an average of approximately 14 plosives per sentence. Given that 600 sentences were annotated, this would lead to ca. 8,400 plosives in the study. However, the rules of connected speech mean that plosives can be elided. Approximately two plosives

per sentence were of this type, which yielded 11.91 plosives per sentence (or 7,147 tokens in total).

Out of these 7,147 tokens, various allophones were not brought forward for further examination:

(1) The phone [ʔ], which is a relatively frequent allophone of /t/ (N = 488) and occasionally of /k/ (N = 60) and /p/ (N = 20), was not examined as its acoustic cues are sufficiently different from [p t k b d g] to warrant a separate study (e.g. analysis of glottal pulse shape and F0 would be necessary).

(2) The post-glottalized phones [p͡ʔ t͡ʔ k͡ʔ] occur as allophones of intervocalic non-foot-initial /p t k/ in the speech of four speakers from the North East (viz. f09, m06, m07, m09). They were omitted because their acoustics are sufficiently different from [p t k] that their inclusion could distort the results. Spectrographic inspection of such tokens suggested that the release burst tends to be greatly attenuated in amplitude and to be shorter than in non-glottalized tokens (in other words, the burst appeared to be more click-like). They were too few in number (N = 43) to allow for a separate statistically reliable acoustic analysis.

(3) The tap [ɾ] occurs occasionally in the data as an allophone of /t d/. Given that the tap generally lacks both a transient and frication, its place-of-articulation acoustics were deemed insufficiently similar to [p t k b d g] to warrant inclusion. Taps were relatively infrequent (N = 105).

(4) Laterally released [t^{l} d^{l}] occurred surprisingly rarely (N = 6 and N = 30 respectively). Spectrographic inspection suggested that their acoustics were sufficiently different from [t d] to warrant their exclusion: they appeared to lack the prominent high-frequency energy that typifies orally released [t d] and to instead have lower-intensity, shorter bursts with energy spread more diffusely.

(5) Nasally released [t^{n} d^{n}] (N = 1 and N = 50 respectively). Spectrographic inspection suggested that the release bursts of such tokens (when they were visible on the spectrogram at all) tend to be very low in amplitude and diffuse, i.e. sufficiently different from orally-released [t d] to justify their exclusion. There were also 16 tokens of [b^{m}], which in error were not excluded from the analysis. The inclusion of these tokens, though undesirable, is not expected to affect the results appreciably: only 11 of the 16 contained release bursts. These bursts are likely to have been low in amplitude and diffuse, which is similar to the pattern of orally-released [b] (as shown Section 6.1). The sole instance of [t^{n}] in the dataset should also have been excluded but was not, in error.

(6) Retroflex [ɖ] (N = 8). These occurred before /r/ across a word boundary, e.g. *burgled recently*, *good reason*. Retroflex was deemed to be sufficiently different a place of articulation

from alveolar [d] to warrant exclusion from the analysis. There were no instances of [ɾ]. Note that syllable-initial realizations of /t d/ before /r/ (e.g. *dry*, try) were not included in the present study since they are realized as postalveolar affricates rather than as alveolar plosives.

(7) Approximant realizations of /b/ (N = 23) and /g/ (N = 11). The position of $F2_{onset}$ tended to be difficult to ascertain with precision due to the lack of an abrupt change in amplitude ('edge') and they lacked release bursts.

(8) Miscellaneous cases that were difficult to analyse (N = 19). Many of these cases involved a sequence of segments that was uttered so rapidly it was unclear where to put the segmentation boundaries.

Fricative realizations of [p t k d] (N = 4, 29, 20, and 18 respectively) were included in the analysis. This was justified on the basis that the fricative realizations have the same place of articulation as their plosive equivalents and, consequently, would be expected to have spectra that are sufficiently similar to the frication found in fricated plosives to warrant inclusion, e.g. fricative /t/ shows the high-frequency peak found in plosive /t/. Thus the friction of the fricatives was labelled '2' as per the frication in the burst. However, given that the present study is not concerned with place of articulation per se but rather place of articulation in *plosives*, this decision to keep fricatives is open to question and was arguably mistaken (see 8.6.1 for further discussion).

Also included were ejective realizations [p' t' k'] (N = 6, 20, and 12 respectively) as the overall spectral shape of these plosives were observed to be broadly similar to those of [p t k] given their shared place of articulation. However, this decision is open to debate as some instances in which ejectives appeared to have prominent F2 in the burst were noted. This is unlikely to have affected the classification much given how rare ejectives were (well under 1% of the total).

After all of these omissions and inclusions, the total number of tokens in the dataset was 6,284 of which 5,471 contained a release burst. Note that in the following chapters, the /p t k/ in /sp st sk/ clusters have been included with /b d g/ due to their voice onset times being essentially the same as /b d g/.

The data were annotated manually using a Praat TextGrid, which contained the following five tiers, illustrated in Figure 4.1:

- attribute
- allophone
- phoneme
- word

- comment

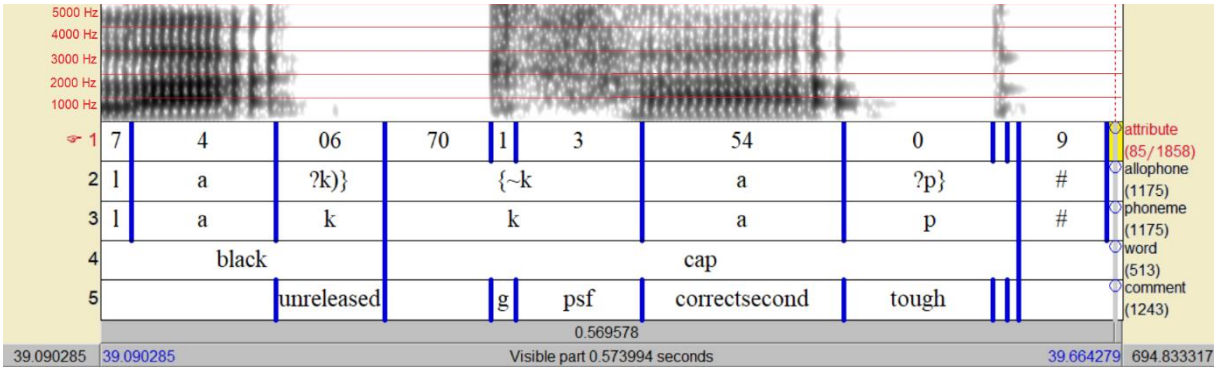


Figure 4.1: Screenshot illustrating the five annotation tiers used in the present study.
This example also illustrates the coding of clusters of plosives – note the labels for the /k/ of *black* ('06') and the /k/ of *cap* ('70') in the attribute tier. Note also the coding of the /a/ of *cap*: given that this vowel constitutes both the following context of the /k/ and the preceding context of the /p/, it is given both a '5' and a '4'.

4.3.2 Attribute Tier

The attribute tier can be thought of as a skeleton of the acoustic features, i.e. its function was to indicate to the Praat script the locations from which to extract the data. It consists of the boundaries of the various phonemic and subphonemic units of interest. Each of these units was labelled a number from 0 to 9.

Number	Meaning
0	the closure of the plosive
1	the transient of the plosive (if any)
2	the frication of the plosive (if any)
3	the aspiration of the plosive (if any)
4	a preceding vowel (if any)
5	a following vowel (if any)
6	a preceding consonant (if any)
7	a following consonant (if any)
8	a preceding pause (if any)
9	a following pause (if any)

Table 4.1: Key of digits used on the attribute tier.

Note that '0' is the only unit that is guaranteed to appear for every plosive.[6] Note also that it was possible for an interval to contain more than one number; for instance if two plosives occur one after the other (e.g. the /kt/ in *doctor, correcting*), then the first one is labelled '06' and the second one is labelled '70' (Figure 4.1 above contains an example of this). The reason for these

[6] For fricative realizations the '0' was added (despite there not being a closure) to preclude the possibility of the Praat script producing an error. This '0' interval was very short (ca. 5 ms).

doubled numberings is that the first plosive is not just a data point in its own right but also constitutes the preceding context of the second plosive, hence the '6'. Likewise the second plosive is not just a token in its own right but is also the following context of the preceding plosive, hence the '7'. Another source of doubled numbering occurs when a vowel is both preceded and followed by a plosive (e.g. the /a/ in *cap* in Figure 4.1, which is preceded by /k/ and followed by /t/) – such vowels are labelled '54' – or when a consonant is abutted on both sides by a plosive (e.g. the /s/ in the phrase *cute speckled*) – such consonants are labelled '76'.

4.3.3 Allophone, Phoneme, and Word Tiers

The allophone tier consists of a narrow phonetic transcription of the plosive and the immediately preceding and following segments. The transcription includes information on devoicing, preceding or following word boundaries, centralization of vowels, and whether the plosive belongs to a stressed syllable (see Appendix 2 for further details). The phoneme tier consists of the same information as the allophone tier except that the three segments were transcribed phonemically (and the information about stress is not included). The word tier transcribes in orthographic form the word to which each annotated segment belongs.

4.3.4 Comment Tier

The comment tier was used for a wide variety of functions:

1. To indicate whether a burst was acoustically typical for its place of articulation;
2. To indicate that a plosive should be omitted, i.e. not taken forward for further analysis; as enumerated in 4.3.1 above, most of these cases involved taps or glottal stops;
3. To indicate that a plosive is unreleased (i.e. does not have a burst; N = 474);
4. To indicate cases in which the formant frequencies given by the tracker were errant;
5. To indicate that a plosive's burst was judged not to be audible. This involved listening to the syllable in which the plosive was found with and without the transient to determine if there was an audible difference. Most cases in which the plosive's transient was judged to be inaudible were bilabial.
6. Various miscellaneous properties of the plosive, e.g. whether it was heavily voiced.

Point 1 requires further elaboration. The aim of the typicality judgement was to find examples of release bursts whose acoustic properties did not match typical spectra. This follows the philosophy that it is atypical cases that most advance understanding of a phenomenon since their lack of correspondence with theory drives the improvement of existing theories (see 8.6.1 and 9.2 for discussion). An alveolar burst was regarded as acoustically typical if it had a

preponderance of energy in the high frequencies (= above ca. 3,500 Hz before front or central vowels, above ca. 3,000 Hz before back vowels) relative to other frequencies. In most cases inspection of the spectrogram on its own was sufficient to establish this, though in unclear cases a Bark-phon spectral slice was examined to determine for definite whether the high frequencies were stronger than the other frequencies.

A velar consonant was regarded as acoustically typical if it had a clear peak of energy in or near F2 (when followed by a back or central vowel or /w r/) or F3 (when followed by a front vowel or /j/). In most cases this was true; there were, however, occasional instances of velars in which there was also a peak in the high-frequency region (typically between 3,500 and 5,500 Hz) that was almost as prominent or more prominent than the peak of energy in the F2 or F3 region. Such cases of two-peaked velar bursts were labelled ‘bimodal’.

A bilabial burst was defined as acoustically typical if it was (1) audible, and (2) appeared to have a relatively flat spectrum.

For further details on the comment tier, see Appendix 3.

4.3.5 Segmentation of the Burst

As mentioned above, the present study sought to segment the burst into its transient and frication. It appears that previous studies have not attempted to do so. Therefore, it is necessary to exemplify the segmentation.

The onset of the transient was placed at the beginning of the first spike in the waveform after the quasi-silence of the closure, following Ladefoged's (2003: 141) and Foulkes et al.'s (2011: 59) segmentation. The offset of the transient (if not followed by frication) was defined as the point at which an abrupt decrease in the amplitude of the transient was apparent. If followed by frication, the end of the transient was defined as the point in which the transient no longer contained energy at a broad range of frequencies (see Figure 4.3 for three examples). This placement of the boundary follows Fant’s (1973: 111) definition of the transient noted in Section 2.1.3 above.

In general, /p b/ tend not to contain frication (Fant, 1973: 113). In the present dataset, of those /p b/ tokens that contain a burst, just 12% of the /p/ tokens and 6% of the /b/ tokens were segmented as containing frication, as can be seen in the following table:

Phoneme	**Total**	**Containing a burst**	*Of which contained a transient*	*Of which contained frication*
/p/	785	665 (84.7%)	637 (95.8%)	77 (11.6%)
/t/	1,310	1,177 (89.8%)	982 (83.43%)	1,103 (93.7%)
/k/	1,128	985 (87.3%)	940 (95.4%)	395 (40.1%)
/b/	1,020	823 (80.7%)	793 (96.4%)	45 (5.5%)
/d/	1,425	1,252 (87.9%)	1,110 (88.7%)	740 (59.1%)
/g/	616	565 (91.7%)	548 (97.0%)	100 (17.7%)
TOTAL	6,284	5,471 (87.0%)	5,015 (91.7%)	2460 (45.0%)

Table 4.2 : Proportion of each plosive phoneme containing a burst, transient, and frication.

There are two main patterns in the data. Firstly, alveolars are the place of articulation with the highest proportion of bursts that contain frication, followed by velars, with bilabials least likely to be fricated. Secondly, there is a trend for the voiceless series to be more likely to contain frication than the voiced series. This is unsurprising given that the impedance at the glottis results in the volume velocity being greater in voiceless stops (Stevens, 1998), and as noted in 2.1.3 greater volume velocity – holding constriction cross-sectional area constant – is one of the aerodynamic conditions that increases the likelihood and/or duration of turbulence.

The rationale behind segmenting the burst into transient and frication will be discussed in greater detail in Section 8.6.2. For now, the primary aim is to describe how the difference between the transient and frication was determined in the spectrogram.

In many cases it was relatively easy to decide that only the transient was present. Here is an example of such a case:

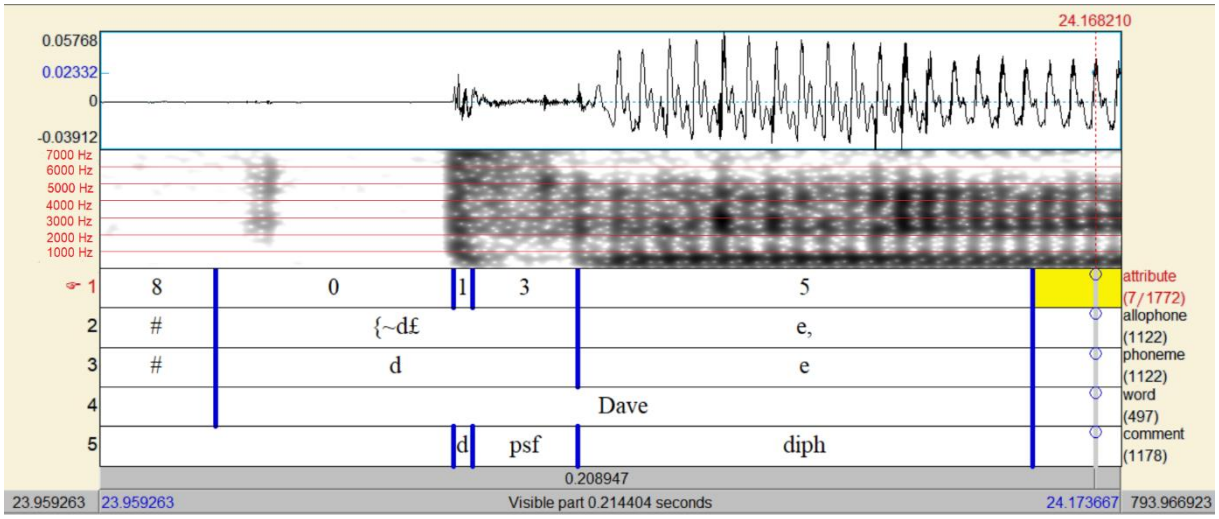


Figure 4.2: Examples of a burst in which it is relatively obvious that only the transient was present. From f03's utterance of the /d/ in *Dave*.

The burst was segmented as containing only a transient as the energy at all frequencies appears to last for approximately the same amount of time. In the following examples this is not the case:

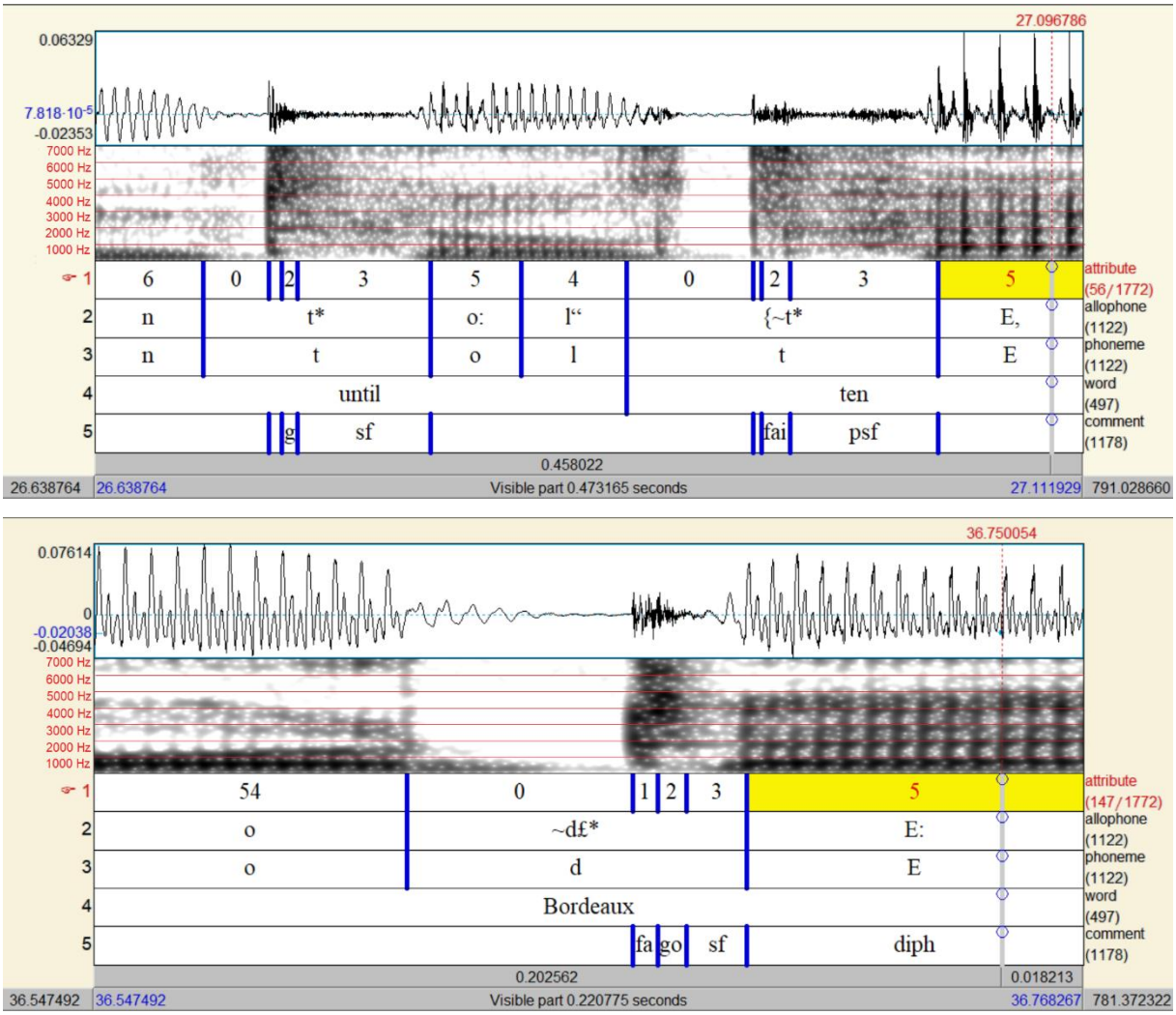


Figure 4.3: Three examples of release bursts that were judged to contain both a transient and frication. This can be seen by noting the blue boundaries on the attribute tier containing the transient (‘1’) and frication (‘2’). The frication can be seen to contain less energy in the F1 region. From f03’s utterance of the /t/ in *until*, the /t/ in *ten*, and the /d/ in *Bordeaux*.

Following Fant’s (1973: 111) definition of the transient as being that part of the burst that contains energy at a wide range of frequencies, the boundary between the end of the transient and the beginning of the frication was placed at the point in which the energy begins to be concentrated at a narrower range of frequencies (which for alveolar place is the high-frequency region above 3,500 Hz), as can be seen in each of the three examples illustrated in Figure 4.3 above (note in particular the lack of energy in the F1 region for the frication relative to the transient).

From the above examples it might appear that the burst really can be segmented into transient and frication straightforwardly. However, difficult cases tended to crop up quite often in practice, such as in the following example:

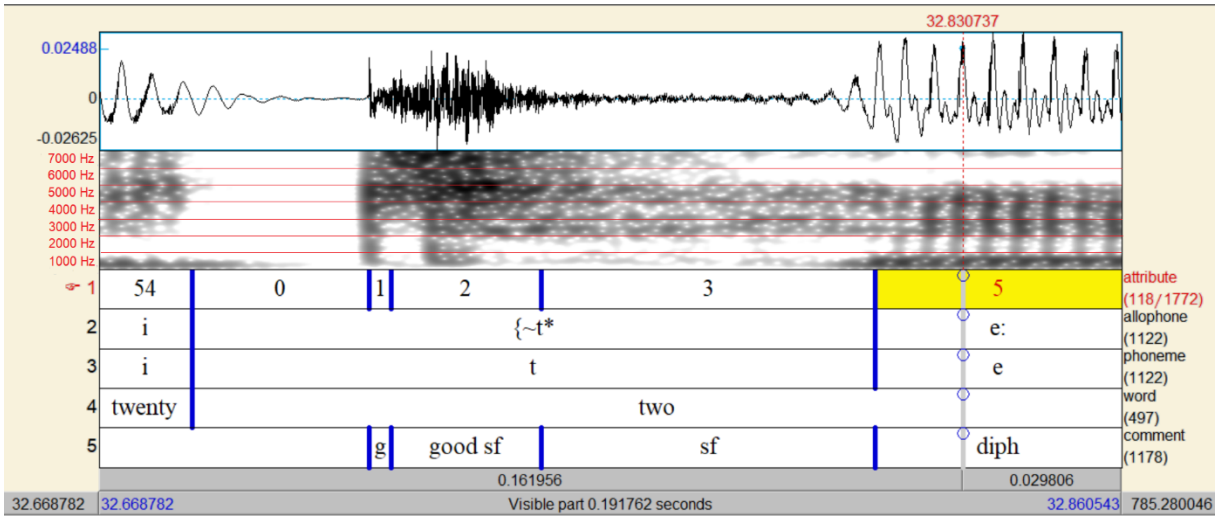


Figure 4.4: Example of inconsistent annotation of the transient and frication.
From f03's utterance of the /t/ in *two*.

In the above example it was decided to place the boundary between the end of the transient and the beginning of the frication at the point where the broad spread of energy disappeared from the spectrum. However, this broad spread of energy can be seen to reappear approximately halfway into the 'frication', which means that this part of the burst meets the definition of the transient, not the frication. But following this definition in the present case would have required segmenting the burst with no less than four labels:

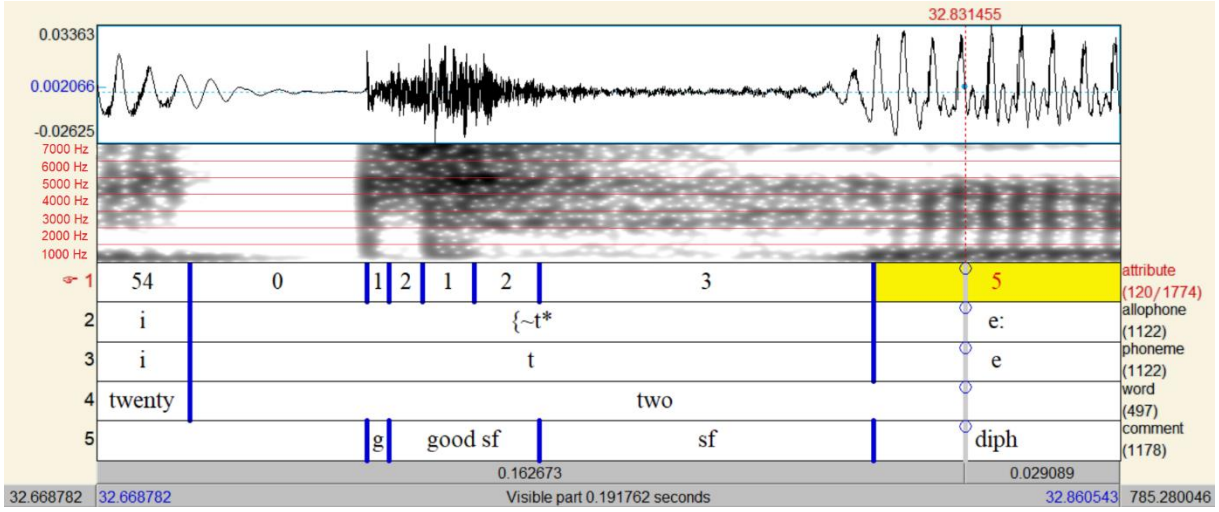


Figure 4.5: Example of a hypothetical annotation of the transient and frication.
Same example as in Figure 4.4.

In all cases, the policy followed was to have only one label of a particular kind in each token (for example if '1' was used it could only be used once, and likewise with '2') to aid the

automatic extraction (i.e. if there were bursts in which '1' or '2' were used more than once, then there would no longer be a one-to-one correspondence between these labels and the number of tokens in the dataset, which would make data analysis considerably more difficult).

In the results for the burst (presented in Chapters 6 and 7), a selection criterion was implemented using an Excel formula of the following form: the burst is represented by the frication whenever only the frication was present (457 cases or 8.4% of the total), and by the transient whenever only the transient or the transient + frication were present (5,015 cases or 91.6% of cases). Thus in the great majority of cases the window that represents the burst information was centred at the onset of the burst. In this respect the issue of choosing between the transient and the frication turned out to be relatively unimportant.

It is already apparent from the example presented above that the segmentation of the burst into the transient and frication was not always straightforward. There is also a more important broader question of whether segmenting the burst into the transient and frication is necessary and/or desirable. In Section 8.6.2 the reasoning behind this segmentation policy will be presented in detail along with a retrospective critique of the reasoning.

4.3.6 Further Segmentation Criteria

Whenever a vowel was a diphthong, only the part of the diphthong closest to the plosive was segmented. In all cases the boundary was put at 50% into the diphthong (the beginning of the diphthong is defined as the beginning of voicing in the case of a diphthong following a voiceless segment and the beginning of an abrupt change in the spectrum in the case of a diphthong following a voiced segment such as a nasasl or lateral).

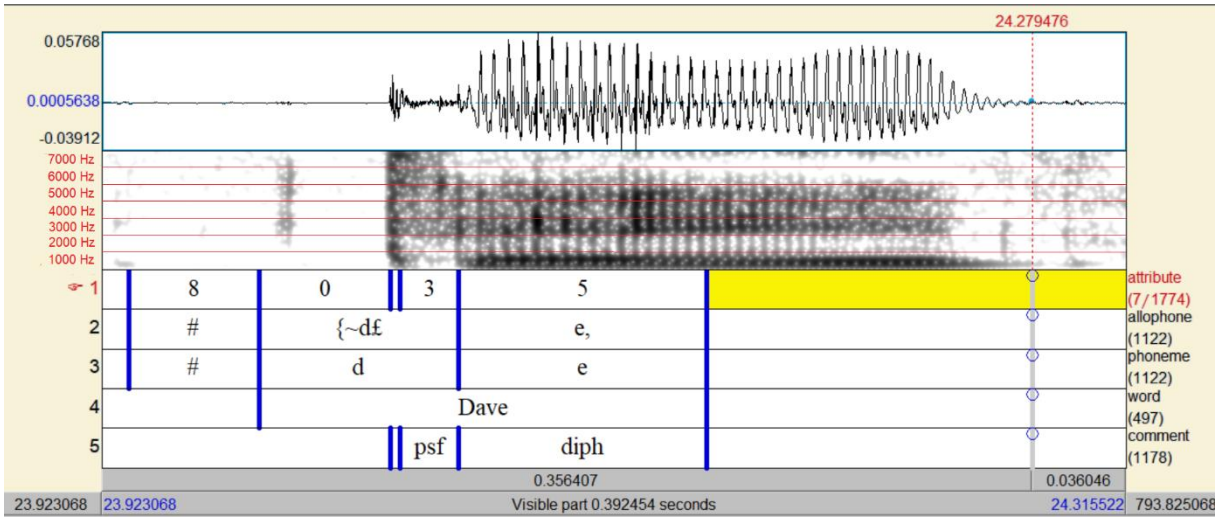


Figure 4.6: Annotation of the diphthong in *Dave*.

Only the part of the diphthong that is closest to the plosive is annotated. Given that the boundary is put at 50% into the vowel, then the diphthong's 'midpoint' is situated at 25% into the overall vowel. Specific cases of monophthongs and diphthongs were investigated to check whether this had an effect on the vowel formant measurements but it did not: the vowel steady-state appeared to be represented by each measurement location.

The above example also illustrates another segmentation policy: the closure duration ('0' in the top tier) of all utterance-initial plosives was arbitrarily set at 50 ms. As for the onset (and offset) of voicing, this was placed at the point where F1 increased notably in amplitude, but excluding edge vibrations (i.e. the same criteria outlined in 3.1.4 for the pilot study). The above example is straightforward as it does not contain an edge vibration. This segmentation criterion corresponds to the point marked 'C' in Foulkes et al.'s (2011: 63) Figure 6.5.

4.4 Formant Measurements

If the plosive was preceded or followed by a pause (labelled '8' and '9' respectively), then no formant information was extracted from such intervals. If, however, the plosive was preceded or followed by a vowel or consonant, then the first three formants were extracted from the midpoint of the segment. The three formants were also extracted from the edge of the segment that adjoined the plosive, i.e. a preceding vowel or consonant ('4' or '6') had its formants extracted at offset, a following vowel or consonant ('5' or '7') at onset. The first three formants were also measured at (1) the onset of the transient; (2) the midpoint of the frication; (3) the onset of the aspiration. The amplitudes of the formants were also extracted at these three time points. However, the formant information extracted from these points was found to be highly error-prone: there is typically no F1 during the burst and aspiration and so the formant tracker would frequently track F2 as though it were F1, and F3 as F2. Thus the formant frequencies extracted by Praat from the burst and aspiration were not reliable enough to be used.

In terms of formant settings, the default Burg formant-tracking algorithm was used and 'Number of formants' was set to 5.0 for all speakers. For the female speakers, 'Maximum formant' was set to 5,500 Hz, while for most of the male speakers it was set to 5,000 Hz. However, for m01 and m04 it was set to 5,250 Hz and for m08 it was set to 4,500 Hz, as use of 5,000 Hz on these speakers appeared in some instances to show reduced accuracy at estimating the formant frequencies. For all speakers the formant frequencies were obtained from a 25-ms Kaiser2 window with a 5-ms timestep and interpolation. The frequencies were converted to Bark in Excel using the same formula found in Praat, namely 7*LN(Hz/650+(SQRT(1+(Hz/650)^2))).

One might wonder whether it is possible to extract formants from the burst or aspiration and – if it is possible – whether a clear, unambiguous formant pattern appears in the burst consistently. At the initial stage of data annotation, the working hypothesis followed was that it is not possible to consistently extract formants from the burst and aspiration. However whilst annotating, examples were noted in which at least one of the formants appeared clearly in the burst. Of particular interest was a fricated /t/ burst from m08 in which the early part of the burst contained a visible F2 but the later part did not. The high-frequency energy in the two parts of the burst looked to be approximately the same in amplitude, which meant that whatever difference in timbre was heard between the two parts of the burst was likely due to the presence of F2 in one part but not the other. When the two parts of the burst were played separately a difference in timbre between the two parts was indeed apparent: the part of the burst with the strong F2 sounded lower in timbre than the part not containing the formant. This indicated that the presence versus absence of the F2 was substantial enough to be noticeable psychoacoustically.

In light of this case it was decided that an examination of the data was warranted, an examination in which the presence of F2 and F3 in the burst would be inspected spectrographically and rated on a five-point scale, with a maximum score indicating that the formant was prominently present and a minimum score indicating that the formant was entirely absent (with intermediate scores indicating varying degrees of ambiguity; see Appendix 3).

The overall conclusion from this examination of the data is that the presence of F2 and F3 in the burst and aspiration is capricious. The burst, recall, may consist of either a transient and/or frication. Let us begin with the results for the transient:

Presence of formant in transient	**F2**	**F3**
no	2131	2737
somewhat	602	694
probable	474	466
definite	543	419
definite and prominent	1035	395
not noted	327	401
total	*5112*	*5112*

Table 4.3: Visibility of F2 and F3 in the release burst's transient.

Cases defined as 'no' were cases where the formant was judged not to be present; 'somewhat' means that the formant appeared to be faintly and/or ambiguously visible; 'probable' means that the formant was probably present but not as unambiguously as in 'definite'. 'Definite and prominent' means that the formant was definitely present *and* appeared to be high in amplitude. 'Not noted' means that the annotator omitted to inspect the spectrogram for the F2 or F3 information. N = 5,112.

The above table shows that of the 4,785 cases in which the visibility of F2 was noted in the burst's transient, in only 1,578 cases (33.0%) was F2 judged to be definitely present. If we liberalize the definition to include cases where F2 was 'probably' present, then the figure rises to 2,052, which is still less than half (42.9%) of the total. Of the remainder 2,131 (44.5%) were judged definitely not to contain F2, with 602 (12.6%) 'somewhat' having F2 (i.e. the F2 was either faintly visible or ambiguously present).

The results for F3 are similar: of the 4,711 cases in which F3 was viewed in the transient, in just 814 cases did it appear to be definitely present (17.2%). If cases where it was 'probably' present are added, the figure rises to 1,280 (27.2%).

The results for the burst's frication:

Presence of formant in frication	**F2**	**F3**
no	509	660
somewhat	505	559
probable	572	474
definite	646	588
definite and prominent	277	168
not noted	56	116
total	*2565*	*2565*

Table 4.4: Visibility of F2 and F3 in the release burst's frication.

Cases defined as 'no' were cases where the formant was judged not to be present; 'somewhat' means that the formant appeared to be faintly and/or ambiguously visible; 'probable' means that the formant was probably present but not as unambiguously as in 'definite'. 'Definite and prominent' means that the formant was definitely present *and* appeared to be high in amplitude. 'Not noted' means that the annotator omitted to inspect the spectrogram for the F2 or F3 information. N = 2,565.

The results for the frication are similar to those for the transient: 36.8% of cases involve F2 definitely being present as against 33.0% in the transient. If 'probable' cases are added to this the figure rises to 59.6%, which is notably higher than the equivalent figure for the transient, (42.9%). This leaves over 40% of cases in which the frication did not appear to contain F2. It should be borne in mind that the number of bursts that contained frication, 2,565, is approximately half the total number of bursts (5,471). Thus even the 59.6% is less than 30% of the total number of bursts. The percentage of cases definitely containing F3 is somewhat lower than the percentage containing F2, which is the same trend as was found in the transient.

Finally, the results for the aspiration:

Presence of formants in aspiration	**F2**	**F3**
no	738	1031
somewhat	595	583
probable	932	662
definite	643	588
definite and prominent	158	64
not noted	450	675
total	*3966*	*3966*

Table 4.5: Visibility of F2 and F3 in the aspiration.

Cases defined as 'no' were cases where the formant was judged not to be present; 'somewhat' means that the formant appeared to be faintly and/or ambiguously visible; 'probable' means that the formant was probably present but not as unambiguously as in 'definite'. 'Definite and prominent' means that the formant was definitely present *and* appeared to be high in amplitude. 'Not noted' means that the annotator omitted to inspect the spectrogram for the F2 or F3 information. N = 3,966.

As with the transient and frication the cases where F2 and F3 were definitely present are in the minority.

In addition to the patchy presence of F2 and F3, there were also cases where the most prominent peak of energy in the F2 and/or F3 regions did not appear to be a genuine formant. That is, a formant-like blob would appear in the F2 or F3 region of the burst but not appear to be contiguous with the F2 in the aspiration or at voicing onset. In Chapter 2 it was noted that release bursts can contain subglottal formants (Blumstein and Stevens, 1979; Fant et al., 1972) whose frequency in alveolars can be ca. 1,200 Hz, i.e. within the F2 region but not the ca. 1,800 Hz value typical of a true alveolar $F2_{onset}$. In the present data set, for F2 there were 81 cases noted of apparently misleading $F2_{onset}$ values in the transient, 2 in the frication, and 0 in the aspiration. For F3 there were 31 cases in the transient, 0 in the frication, and 0 in the aspiration.

Modarresi et al. (2004) recommended measuring the aspiration in the burst for obtaining better results for classifying bilabial place. However, Suchato (2004) found that $F2_{onset}$ in the burst was less successful at distinguishing place than $F2_{onset}$ measured at the onset of the following vowel. This was in spite of the fact that the formant measurements in his study were manually checked, i.e. the lesser success of $F2_{onset}$ in the burst cannot be dismissed as merely being an automatic formant-tracking error. Blumstein and Stevens (1979) found that only a third of /d/ bursts and a sixth of /t/ bursts contained F2. Nossair and Zahorian write (1991: 2981) "For many tokens, particularly labial and alveolar stops, the formants are simply not well

defined in the burst and aspiration segments." The results of the present examination are in accord with these authors.

With all of these considerations in mind it was decided not to use the automatic formant measurements from the burst, frication, and aspiration as acoustic attributes in the present study.

4.5 Burst Measurements

4.5.1 Choice and Positioning of Window

As outlined earlier, the plosive's release burst was segmented into two parts, the transient and the frication (see Section 2.1.3 for an acoustic definition of each and illustration from a spectrogram). For both the transient and the frication, measurements were taken using a window centred at two points: the onset of the transient and the midpoint of the frication. The onset of the transient was chosen to minimize the likelihood of the window including voicing from the following segment. The midpoint of the frication was chosen as this is generally the part of the frication that been examined by the largest proportion of fricative studies (Koenig et al., 2013), and given that a plosive's frication is acoustically equivalent to a brief fricative (Fant, 1973), this is the analogous position to place the window in a plosive. It is possible that placing the window the middle of the frication rather than at the beginning may in a few cases have slightly increased the likelihood of the window containing some unwanted information from the following vowel onset. This is only possible in the case of /b d g/ rather than /p t k/, since the aspiration in the latter series tends to be relatively long (the mean aspiration duration of /p t k/ in the present dataset is 22.4 ms as against 6.2 ms for /b d g/).

The window used on the transient and frication was Praat's 'Kaiser1'. The Kaiser window, like a Gaussian window, is a relatively bell-shaped window with a tall middle and gently tapering sides that are designed to prevent the generation of the frequency splatter that square-shaped windowing of the signal is liable to create. The '1' in 'Kaiser1' refers to the fact that its sigma value is set to $\sigma = 1$. This σ refers to the steepness of the window's slopes; $\sigma = 1$ means that the slopes are relatively shallow. For the Kaiser series of windows, Praat allows the option of having a window slope (σ) anywhere between 1 and 5. These various window slopes were tested extensively before the data extraction phase and it was noted that they gave similar spectra except that the steeper-sloped windows (i.e. Kaiser2 to Kaiser5) tended to give slightly lower-amplitude spectra, which in some cases appeared to make the burst peak slightly less prominent. For this reason, Kaiser1 was chosen.

In terms of the window's length, this was set to 25.6 ms, which was implemented using a Fast Fourier Transform (FFT). The choice of this length is linked to the fact that the signal was downsampled to 20,000 Hz. Downsampling to 20,000 Hz means that the number of

samples in the 25.6-ms window is 1,024, which is a multiple of 2 (viz. 2^{10}). The duration of windows used in automatic speech recognition typically ranges between 20 and 30 ms (Hermansky, 1990: 1739; Huckvale, 2013: 209) and Blumstein and Stevens (1979) also studied plosive bursts using a 25.6-ms window, so in that sense the present duration is relatively conventional. Potential limitations of this window length will, however, be noted in Section 8.6.

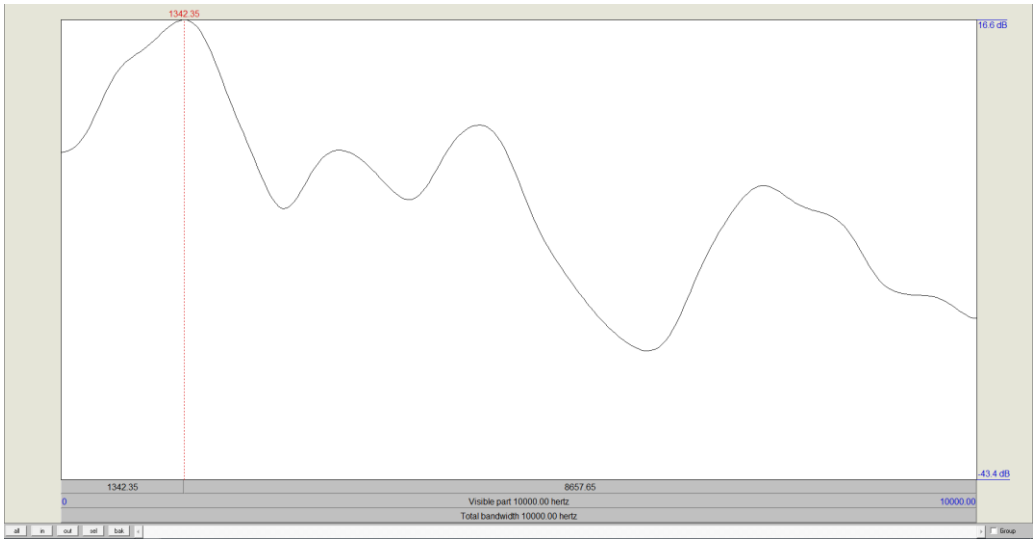
Cepstral smoothing was set to 1,000 Hz. Pre-emphasis was not used on the burst spectrum, which is in accord with certain previous studies of obstruent place of articulation such as Suchato (2004) for plosives and Koenig et al. (2013) for fricatives. In contrast, Blumstein and Stevens (1979) used 'high-frequency pre-emphasis' on their spectra. However their definition of what constitutes 'high-frequency' is not specified, so it is not possible to replicate their application of pre-emphasis. In any event, Blumstein and Stevens used spectra with a sampling rate of 10,000 Hz, which means that the maximum frequency in their spectra was 5,000 Hz. In contrast, in the present study's spectra frequencies were present up to 10,000 Hz, which means that applying pre-emphasis to the spectrum would boost frequencies above 5,000 Hz relative to those below 5,000 Hz (since pre-emphasis, by definition, boosts the spectrum by 6 dB per octave for all frequencies). There is, however, psychoacoustic evidence (presented in 4.5.2) that frequencies above 5,000 Hz are perceived as *less* loud than those between ca. 2,000 to 5,000 Hz, and thus carrying pre-emphasis beyond 5,000 Hz is arguably inappropriate. Furthermore, comparison of attributes with and without pre-emphasis (set from 50 Hz as per the default in Praat) revealed that the performance of the CoGdB attribute at distinguishing place of articulation was markedly poorer with pre-emphasis on: its discriminant analysis classification accuracy worsened from 62% without pre-emphasis to 55% with pre-emphasis. Given that Suchato (2004) found CoG to be the strongest of his attributes, this suggested that the pre-emphasis was adversely altering the burst spectrum. Thus the Hz-dB results in following chapters are for non-pre-emphasized spectra.

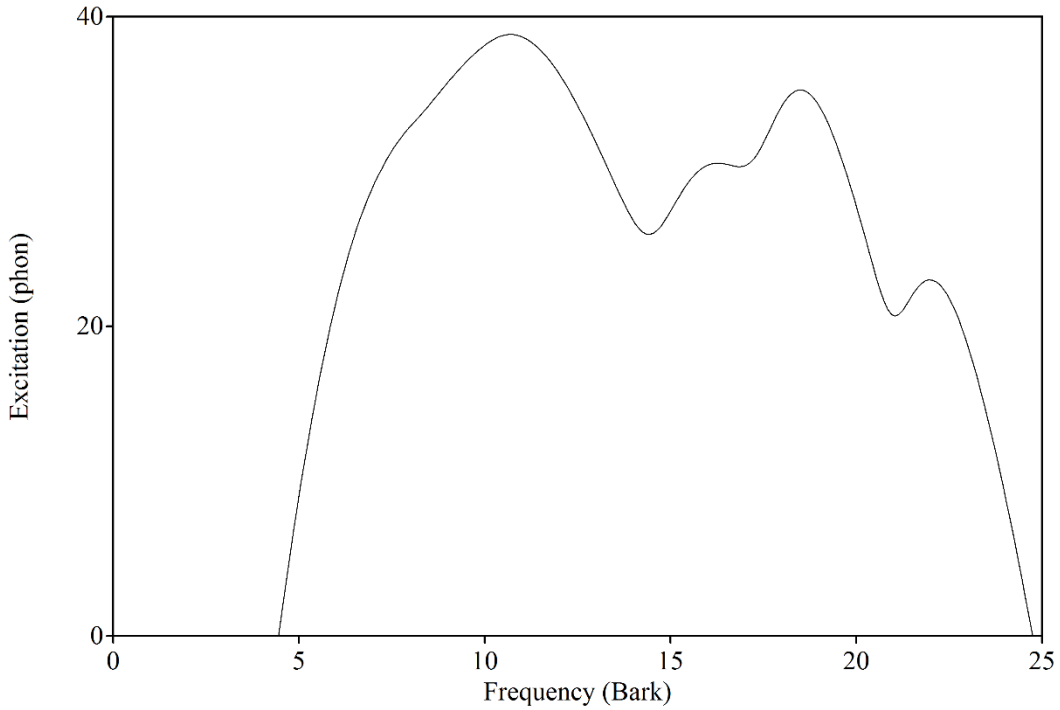
4.5.2 Hz-dB, Bark-Phon, and Bark-Sone Spectra

Aim 3 of the present study is to compare the performance of acoustic attributes derived from three kinds of spectral representation: the widely used Hz-dB spectrum and the psychoacoustically-oriented Bark-phon and Bark-sone spectra. The goal is to see whether these spectral representations affect the classification accuracy of burst-based attributes. For this reason the classification accuracy of equivalent acoustic attributes from the Hz-dB, Bark-phon, and Bark-sone spectra will be compared with each other (Sections 6.4 and 6.5).

In Section 2.3.1.9 we summarized the warping of the frequency and amplitude axes found in perceptual linear prediction (PLP; Hermansky, 1990), which is designed to mimic to an approximate degree the warping of the frequency and amplitude axes that has been documented in human perception by cochlear and psychoacoustic studies (e.g. Zwicker, 1961; Greenwood, 1961; Stevens, 1957). The Praat software programme contains a somewhat similar kind of spectral representation termed 'Excitation', which can be obtained from a regular 'Spectrum slice' object by selecting the slice in the Praat Objects menu and clicking 'Analyse > To Excitation…'. The resulting 'Excitation slice' can then be plotted in the Praat Picture window by selecting the Excitation slice, clicking 'Draw' in the Praat Objects menu, and specifying the desired settings. In this thesis the resulting image will be termed a 'Bark-phon' spectrum, and an example of one can be seen in Figure 4.8 below. In this section this spectrum and the other two spectra are presented in greater detail.

To give the reader a provisional sense of how these spectra look different, Figure 4.8 shows the Hz-dB, Bark-phon, and Bark-sone spectra of a /k/ burst:


1342.35
16.6 dB
-43.4 dB
1342.35
8657.65
0
Visible part 10000.00 hertz
10000.00
Total bandwidth 10000.00 hertz


40
20
0
Excitation (phon)
0
5
10
15
20
25
Frequency (Bark)

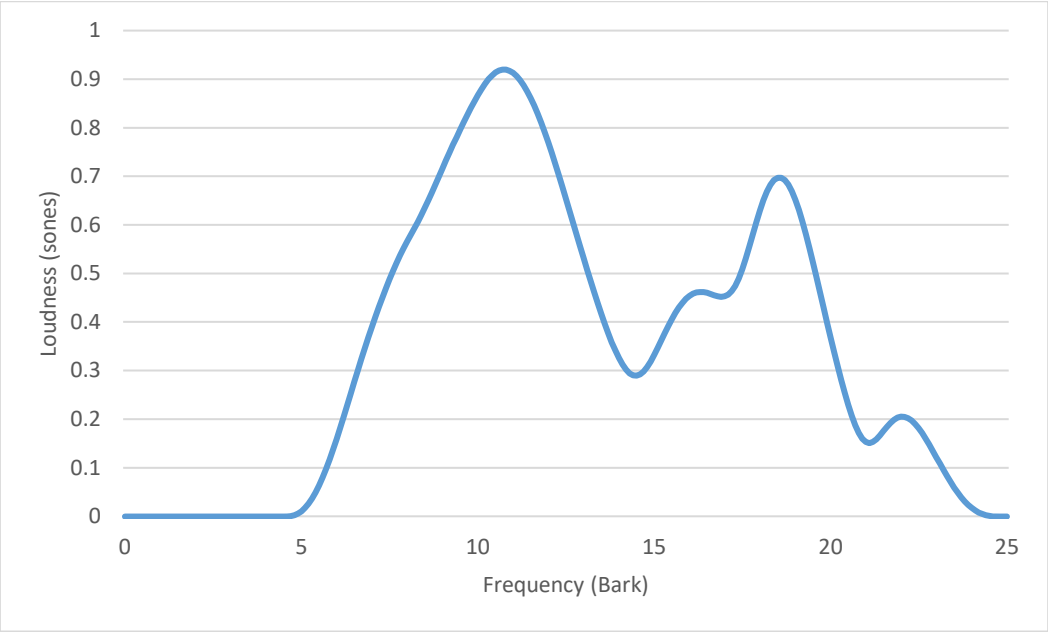


Figure 4.7: Hertz-decibel, Bark-phon, and Bark-sone spectra for a /k/ burst as produced by f01 in the word *called*.

One measure of the ear's ability to discriminate frequency is the just noticeable difference (JND), also known as the difference limen (DL) (Moore, 2012: 204). This refers to the minimum change in frequency needed for the listener to notice a difference, averaged over a number of subjects (who are typically young adults with no known hearing impairment). It has been found that the higher the frequency of the stimulus (in Hertz), the larger the JND. This means that the relationship between frequency on a Hertz scale and frequency on a psychoacoustic scale is not linear.

Over the decades there have been various attempts to make a psychoacoustically-derived frequency scale that would reflect this non-linear frequency selectivity of the human ear. The earliest attempt is the mel frequency scale (Stevens et al., 1937), which is still widely used in automatic speech recognition in the guise of mel-frequency cepstral coefficients (MFCCs) (Davis and Mermelstein, 1980). The experimental methodology on which the mel scale is based has been questioned (e.g. Moore, 2012: 215) and it has been superseded by other attempts to derive a perceptual frequency scale, such as the Bark scale (Zwicker, 1961) and the equivalent rectangular bandwidth (ERB) scale (Moore and Glasberg, 1983b; Glasberg and Moore (1990)). Of these two frequency scales ERB is derived using a more sophisticated experimental technique known as the notched-noise method (Patterson, 1976). However, the ERB and Bark frequency scales are approximately the same shape above 1,000 Hz, i.e. both are close to logarithmic. It is only below these frequencies that the scales differ to an appreciable degree, the Bark scale being less logarithmic than the ERB scale at these frequencies.

Nevertheless the Bark scale is a reasonably close approximation to the psychoacoustic spacing of frequency over most of the audible spectrum.

One might wonder what the purpose is of going to the trouble of using a non-linear frequency scale. Using a statistic known as mutual information to quantify the information dependence between two adjacent frequency regions, it has been found that the higher the frequency, the higher the degree of mutual information between adjacent frequencies, and this degree of increase of mutual information is roughly equivalent to the warping of frequency by psychoacoustically-oriented frequency scales (Kingsbury, 2009: 10:29-11:09; Yang et al., 1999). So when using a Bark or ERB frequency scale, the spectrum is closer to being approximately equally information-dense at all frequencies than it would be with a Hertz scale; on this latter scale, the information density of the spectrum is such that low-frequency bins tend to carry more information than high-frequency bins.

That, then, is the frequency dimension. The other dimension of a spectral slice is magnitude, also referred to as amplitude or intensity (Moore, 2012: 6; 10). It has been found that the magnitude of a sound as measured using *physical* instrumentation such as a microphone can differ in various ways from the *psychological* percept of loudness. The first important respect in which loudness and intensity differ arises from the fact that as sound passes through the outer and middle ear, the intensity of some frequencies is boosted more than others (Moore, 2012: 44; 57-61). Or to put it differently: the ear's responsiveness to a spectral component varies depending on the component's frequency. For example, imagine listening to a 1,000 Hz pure tone that is just about audible. Would one be able to hear a 100-Hz tone if it were of the same intensity? No: to be just audible, the 100-Hz tone would need to be approximately 25 dB greater in level than the just-audible 1,000 Hz tone. In contrast a 3,000 Hz pure tone would be still be audible even if it were 10 dB less intense than the just-audible 1,000 Hz tone.

This variation in loudness at different frequencies has been examined by psychoacousticians since the early twentieth century and has led to the development of a representation known as equal-loudness contours. These contours represent the relative loudness of pure tones coming from a front direction in a free field (Moore, 2012: 135) and hold for stimuli longer than 500 ms (Fastl and Zwicker, 2007: 203). The contours have been tweaked over the years as more data and better modelling have been implemented. The latest edition of the contours is shown here:

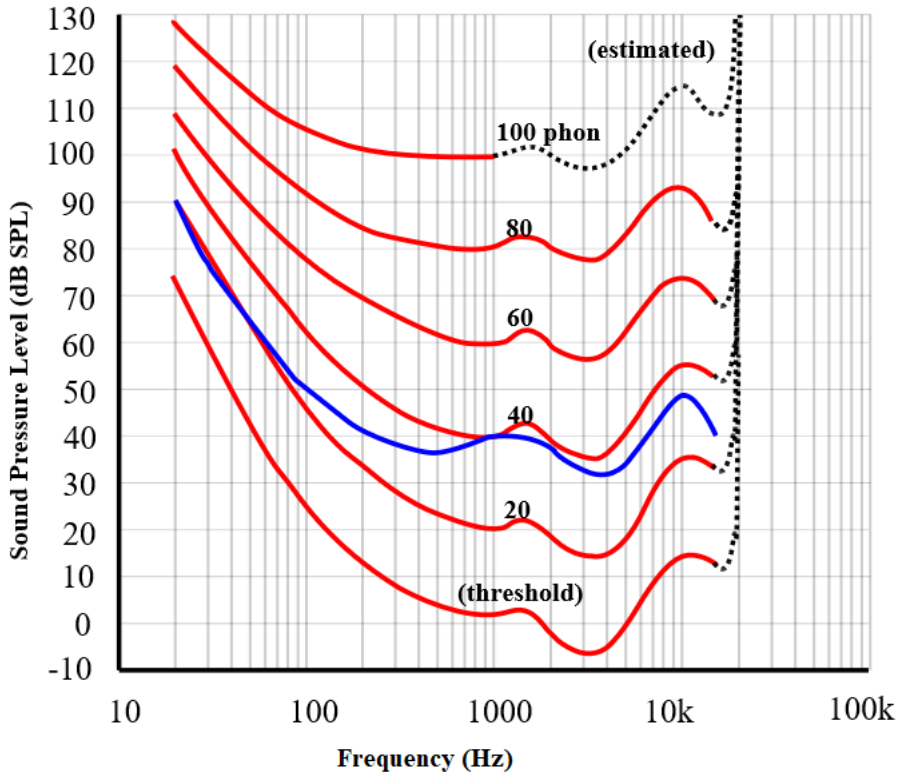


Figure 4.8: The latest edition of the equal-loudness contours (ISO 226, 2003).
The red curves show the loudness in phons, while the horizontal grey lines show the intensity of the input sinusoid. For example, a 200-Hz sinusoid with an intensity of 70 dB SPL corresponds to a loudness of 60 phons. (The blue line shows the shape of an obsolete edition of the contours.)

The point of reference around which the contours are based is a 1,000-Hz tone whose sound pressure level is 40 dB SPL: such a sound is defined as having a loudness of 40 phons. All other sound intensities in the chart are expressed relative to this benchmark. The experimental procedure is that a subject listens to a pure tone of known frequency and intensity and has to adjust the loudness of a 1,000-Hz tone until it sounds equally loud as the test tone (Moore, 2012: 134). The sound level of this modified 1,000-Hz tone is logged and the procedure is repeated for a wide variety of tones of different frequencies and sound levels. The results for a sample of listeners are averaged together and smoothed to form equal-loudness contours. The red curves in Figure 4.8 show the loudness in phons while the horizontal grey lines show the intensity of the input sinusoid. For example, a 200-Hz sinusoid with an intensity of 70 dB SPL corresponds to a loudness of 60 phons, which can be seen in Figure 4.9 by using the vertical and horizontal grey lines to locate the point representing a 200-Hz sinusoid at 70 dB SPL and then seeing which of the red contours intersects this point.

The Praat manual does not state what edition of the equal-loudness contours it uses. However, it appears from Boersma (1998: 106) to be that of Fletcher and Munson (1933). The Praat equal-loudness contours are roughly similar in shape to the contours shown in Figure 4.8 above. The shape of the contours in Praat can be illustrated by generating a click train, then placing a 25.6-ms window on the centre of one of the pulses in the train and plotting the resulting spectrum. A click is the ideal sound to use because in a Hz-dB spectral slice it has a perfectly flat spectrum. Thus the shape that the click assumes in the Bark-phon spectral slice indicates how Praat's equal-loudness contours warp the spectrum.

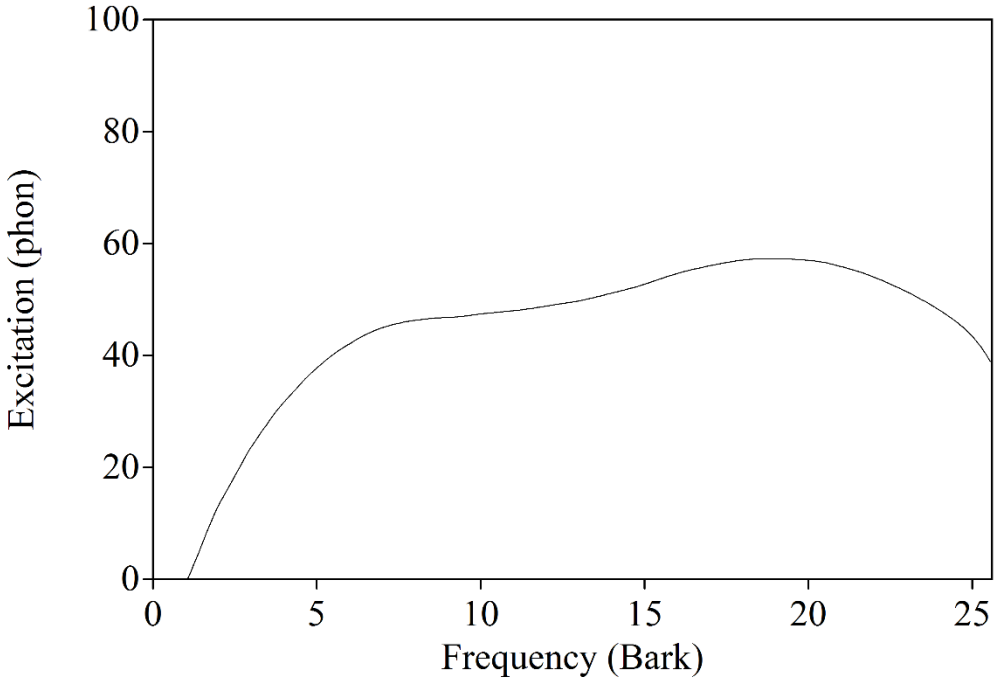


Figure 4.9: Bark-phon spectral slice (termed 'Excitation pattern' in Praat) for a single broadband pulse of a click train, windowed with a 25.6 Kaiser1 window.

It is evident that the shape of the equal-loudness contours used by Praat follows a broadly similar shape to the ISO 2003 version. For example, the low-frequency components are reduced in amplitude relative to the mid-frequency and high-frequency components. There are, however, some differences in detail: Praat's contours boost sound the most at around 18-19 Bark (or ca. 4,200-4,800 Hz), whereas the 2003 contours boost the most around 3,000 to 3,500 Hz. Another difference is that the 2003 contours have a slight kink in them between 1,000 and 2,000 Hz that is not found in the Praat contours.

Thus far our discussion of the phon and decibel scales has focused on their differences. What they have in common is that they both compress the magnitude axis logarithmically. That is, each increase of 10 dB corresponds to a 10-fold increase in the pre-log-compressed intensity.

After the logarithmic compression a change from 40 to 50 dB (or 40 to 50 phons) amounts to a 25% increase in value, and a change from, say, 50 to 60 dB corresponds to a 20% increase.

Do human listeners perceive, say, a 60 dB sinusoid as being 20% louder than a 50 dB sinusoid? No. It turns out that – at least for moderately intense sounds, i.e. those above 40 dB but below 80 dB SPL (Moore, 2014: 6) – listeners perceive each 10-dB (or 10-phon) increase as approximately corresponding to a doubling of loudness (Moore, 2012: 137; Fastl and Zwicker, 2007: 206; see also Lyon, 2017: 80-81, Stevens (1957), and Stevens (1972)). The sone scale captures this, e.g. an increase in loudness from 40 to 50 phons amounts to an increase from 1 to 2 sones.

Let us now discuss how to convert from the phon to the sone scale. The formula used in the present study is as follows. The formula for values greater than 40 phons is the following (Fastl and Zwicker, 2007: 207; Sengpiel, undated):

(1) $\text{sone} = 2^{(phon - 40) / 10}$

For values below 40 phon, the formula is as follows:

(2) $\text{sone} = phon^{2.86} - 0.0005$

In Praat the formula in (1) is applied to all phon values whether they be above or below 40 phons (Boersma and Weenink, 2014). The fact that Praat uses the same phon-to-sone frequencies at all sound levels leads to discrepancies below 40 phons between the sone values it produces and the sone values produced by the standard formula, as shown in the following table:

phons	**sones (standard formula)**	**sones (Praat formula)**
40	1	1
30	0.44	0.5
20	0.13	0.25
10	0.02	0.13
5	0.003	0.06

Table 4.6: Discrepancies between the sone values yielded by the Praat formula and the official formula.

The standard formula compresses the values at low sound levels much more than the Praat formula. This means that using the Praat version of the phon-to-sone formula would have weakened the way that the sone scale compresses low amplitudes together. Lyon (2017: 80-81)

criticizes the decibel (and hence phon) scale for making low amplitudes too prominent and larger amplitudes not prominent enough. This can be seen in the following comparison:

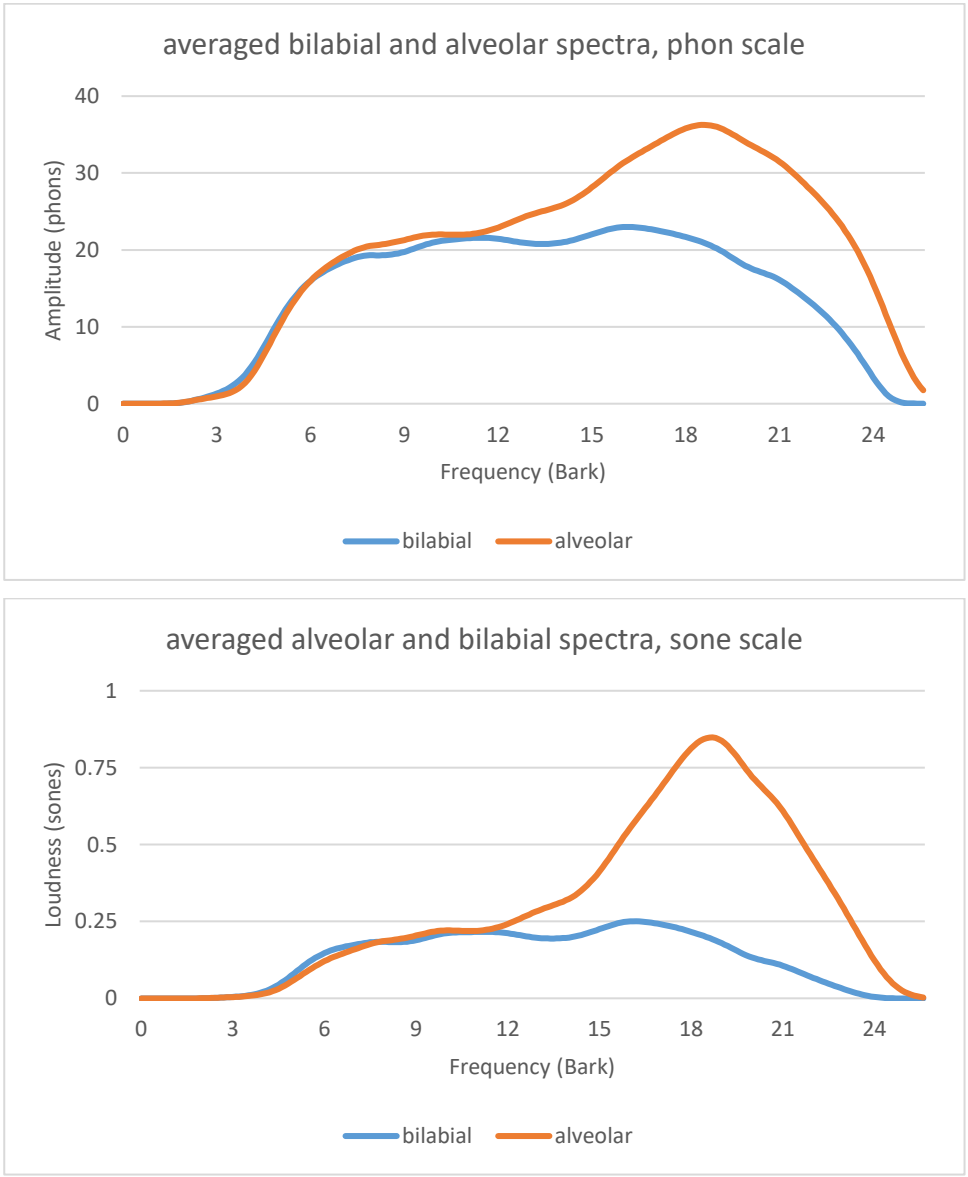


Figure 4.10: Comparison of the mean bilabial and alveolar burst spectra on the phon and sone scales. The sone scale makes the alveolar's high-frequency peak more prominent, and makes the bilabial's entire spectral envelope appear smaller, at least relative the alveolar one. Bilabial N = 1,490; alveolar N = 2,429.

The sone scale enhances the difference between the mean alveolar and bilabial spectra, both by shrinking the low-amplitude body of the bilabial envelope and by enhancing the high-frequency peak of the alveolar spectrum. This difference between the bilabial and the alveolar spectra would not have been as large if the Praat version of the sone-to-phon formula illustrated in Table 4.6 above had been utilized.

It was therefore decided to convert the sound level from phons to sones in Excel after the data had been extracted from Praat using the formulae in (1) and (2) rather than letting Praat

do the conversion with its own formula so that the standard phon-to-sone formula could be used.

The default setting in Praat (followed in the present study) is for the Bark-phon spectrum to have a resolution of 0.1 Bark, i.e. a frequency component every 0.1 Bark, yielding 256 datapoints in total. Three frequency ranges examined in the present study are 'All' from 6.9 to 22.4 Bark (750 to 8,000 Hz), 'High' from 16.7 to 22.4 Bark (approximately 3,500 to 8,000 Hz), and 'Mid' from 9.9 from 15.6 Bark (ca. 1,250 to 3,000 Hz). Note that the High and Mid are the same size as they both contain 58 spectral components.

There are two kinds of amplitude-based attributes that will be derived from each of these regions (on both the Bark-phon and Bark-sone representations): the total loudness and the peak loudness. To get the total loudness, we sum up the loudness (in sones) of the 58 components and divide this sum by 0.1 (since there is a component every 0.1 Bark). The resulting value is often called the sound's 'loudness density' in the hearing literature (e.g. Moore 2012: 140), i.e. it is the loudness per critical band (i.e. per Bark unit). To obtain the peak loudness is more straightforward and can be implemented in Excel using the MAX function followed by the frequency range in parentheses. In terms of frequency-domain attributes, these involve picking the frequency of the peak amplitude in a specified frequency region. (These were implemented with formulas containing Excel's MATCH, INDEX, and MAX functions.)

It is important to bear in mind that although the Bark-phon and Bark-sone representations used in the present study attempt to emulate certain aspects of psychoacoustics, they should not be regarded as a faithful simulation of the auditory periphery. More faithful simulations include Lyon (2017), who provides a comprehensive computational implementation of the auditory periphery, and Moore (2014), who summarizes the history of loudness models developed by the Cambridge auditory research group. A discussion of all the non-linearities (e.g. automatic gain control, adaptation) in the mapping between physical intensity and the psychological percept of loudness is beyond the remit of the present thesis.

Chapter 5: Formant Frequencies

In this chapter we examine the extent to which plosive place of articulation can be distinguished by formant frequencies. The focus will primarily be on the voiced series /b d g/ since the voiceless plosives /p t k/ contain aspiration, which cuts off the voiced formants for several tens of milliseconds after the release of the plosive. Because of this we do not expect the formant frequencies to play much of a role in distinguishing place of articulation for these consonants.

The greater part of this chapter will be concerned with $F2_R$, which is an approach to formant transitions that is reminiscent of the $F2_{locus}$ concept that was reviewed in Section 2.2.1. Recall that $F2_{locus}$ is a reconstructed abstract frequency that occurs ca. 50 ms prior to the observed onset of a formant transition. It was hypothesized that $F2_{locus}$ should be able to classify place of articulation with greater accuracy than $F2_{onset}$. This is among one of several aspects of $F2_R$ that will be tested in the present chapter.

Before turning to $F2_R$ proper, in Section 5.1 mixed-effects modelling will be employed to probe V_1CV_2 sequences. This is done in light of Öhman's (1966) finding that V_1 influences the frequency of $F2_{onset}V_2$ considerably. The results of this modelling will be used to decide whether the influence of V_1 on $F2_{onset}$ is as large as what Öhman found and, consequently, whether it is important to incorporate V_1 in the $F2_R$ formula.

Following this, the rationale for investigating $F2_{locus}$ will be presented (Section 5.2), and in 5.3 I develop the formula for exploring $F2_{locus}$, termed $F2_R$. Most of the rest of the chapter (5.4) examines the performance of $F2_R$ under a variety of conditions; in each case the performance of $F2_R$ at distinguishing place of articulation is compared to the established practice of using the two attributes $F2_{onset}$ + $F2_{mid}$. This allows us to see whether the collapsing of $F2_{onset}$ and $F2_{mid}$ into a single attribute is viable.

In 5.4.1 the basic picture of $F2_R$ for both voiced and voiceless stops is established, and analogous attributes for the F1 and F3 transitions are also examined, termed $F1_R$ and $F3_R$. In 5.4.2 the extent to which formant normalization helps to improve the classification accuracy is investigated and in 5.4.3 the same question is posed regarding splitting the classification by vowel backness. In 5.4.4 $F2_R$ is examined in greater depth by comparing the mean $F2_R$ frequencies of front vowels and back vowels using the theoretical underpinning of the 1950s locus theory as a guide. In 5.4.5 the classification accuracy of place of articulation using information from VC transitions is compared to the accuracy obtained from the CV transitions. In 5.4.6 the performance will be examined of the formant distances 'F2 – F1' and 'F3 – F2' at distinguishing place of articulation and the $F2_R$ approach will be applied to them to determine whether their classification accuracy is stronger than using single formants as attributes. In 5.4.7

we scrutinize the F2 properties of schwa and compare it to non-schwa vowels to see if schwa has a different $F2_R$ pattern to other vowels.

Later on in the chapter (5.4.8) a more advanced type of $F2_R$ is formulated in which the distance (in milliseconds) between $F2_{onset}$ and $F2_{mid}$ is taken into account, and test whether its performance constitutes an improvement over the simpler kind of $F2_R$ that does not incorporate time.

At the end of the chapter the role of the arbitrary constant *c* in the $F2_R$ formula is discussed, including whether it is really needed (5.4.9). In the final section (5.4.10) V_1CV_2 tokens are examined to quantify the improvement in classification accuracy caused by having information from both the V_1 and V_2 contexts relative to V_1 or V_2 on their own.

5.1 Modelling VCV Sequences

5.1.1 Background

In the literature review the work of Öhman (1966) was presented, who analysed the formant transitions associated with Swedish /b d g/ in intervocalic context. This study has been highly influential in phonetic science: as of August 2018 it has been cited almost 1,300 times. Nevertheless several features of the study were noted: (1) it consisted of the speech of just a single speaker; (2) that speaker was the author; (3) the material consisted of nonce V_1CV_2 sequences; (4) the speaker said the syllables in a monotone with equal stress on both syllables; (5) the data consisted of three repetitions of 5×5 V_1-V_2 contexts for three different consonants, i.e. the sample size was N = 225.

In contrast, the data for intervocalic /b d g/ in the present dataset are as follows: (1) the speech of 20 speakers, none of whom are the author; (2) the material comprises real words in semantically meaningful sequences; (3) the speakers were not required to pronounce the material in a particular manner; (4) the data were repeated just once by each speaker and N = 758, over three times the size of Öhman's dataset. It is intended that these differences will yield data that are more representative of real-life speech than Öhman's study.

Furthermore, the present study employs linear mixed-effects modelling, which was not in use at the time of Öhman's study. This statistical model was chosen because it deals efficiently with the dependencies that exist when repeated measurements are taken from the same speaker and/or the same word (more elegantly than traditional methods such as by-items and by-subjects averaging; Winter 2013: 5).

As I noted in the literature review, Öhman's (1966) study is significant in that it was the first study of plosives' formant transitions in V_1CV_2 sequences rather than the CV sequences that had been examined before. Lindblom and Sussman (2012: 4) note that the effect of the

study was to end the debate over an invariant locus frequency. This was because Öhman's diagrams showed substantial influence of the F2 frequency in the preceding vowel (henceforth '$F2_{mid}V_1$') on the F2 frequency at the onset of the following vowel (henceforth '$F2_{onset}V_2$'). For example, here are the diagrams for /obo/ and /ybo/:

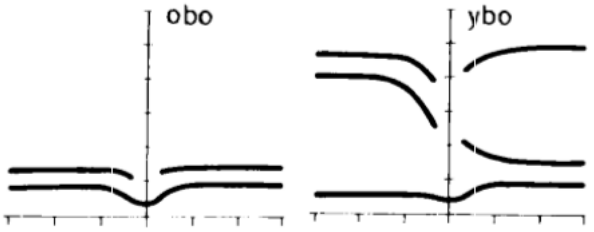


Figure 5.1: Öhman's (1966: 160) stylized spectrograms for /obo/ and /ybo/.
Each image is the result of averaging together the hand-traced formants on spectrograms for three repetitions.

In Figure 5.1 above $F2_{onset}V_2$ is approximately 650 Hz in /obo/ but 1,050 Hz in /ybo/. $F2_{mid}V_1$ is approximately 700 Hz in /obo/ and 1,900 Hz in /ybo/. Thus there is a 1,200-Hz difference in frequency between the two vowel midpoints corresponding to a 400-Hz change in the frequency of $F2_{onset}V_2$. That is, the change in frequency to $F2_{onset}V_2$ caused by the preceding vowel is approximately 0.33.

The present study employs the Bark scale and on this scale the ratio between the above four values is 0.43 rather than 0.33, which suggests slightly more acoustic coarticulatory influence. But in any event the point is that whichever frequency scale is adopted, the coarticulatory influence of V_1 on $F2_{onset}V_2$ in Öhman's data is large.

Here are Öhman's diagrams for /odo ydo ogo ygo/:

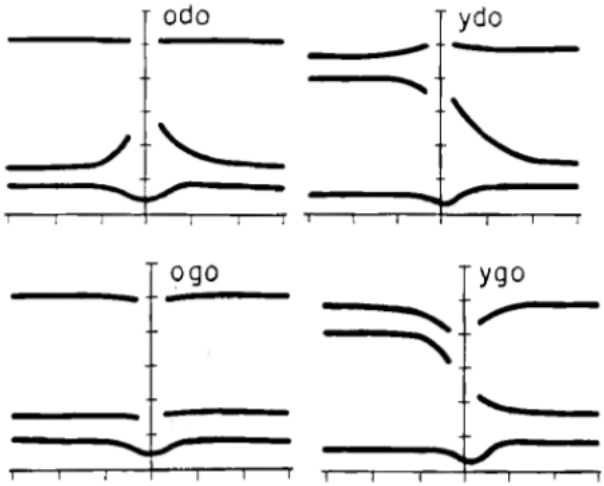


Figure 5.2: Öhman's (1966: 161-162) diagrams for /odo ydo ogo ygo/.

The difference in $F2_{onset}$ between /ydo/ and /odo/ is 400 Hz. When this is divided by the difference in frequency between /y/ and /o/, the resulting ratio is 0.33. This result is the same as those for /obo ybo/ above. However, when the Bark scale is used the figure is 0.28 (as against

0.43 for /b/ above). Regardless of which frequency scale is used, the point is that the influence of $F2_{mid}V1$ on $F2_{onset}V2$ is again substantial. The equivalent ratios for /g/ are 0.21 (using Hertz) or 0.26 (using Bark).

This cursory exploration of Öhman's data does not do his findings justice. Nevertheless it illustrates the general point: the acoustic influence of $F2_{mid}V_1$ on the frequency of $F2_{onset}V_2$ in his data is substanial. This suggests that a model that utilizes place of articulation and $F2_{mid}V_2$ in modelling $F2_{onset}V_2$ would be less accurate than one that also incorporated $F2_{mid}V_1$.

However, as noted at the beginning of this section, there are several important differences between Öhman's data and the present study's data such that it is something of an open question whether the acoustic influence of $F2_{mid}V_1$ on $F2_{onset}V_2$ really is as substantial in naturally-occurring English words read by 20 speakers as Öhman found it to be in nonce Swedish words uttered with equal stress in a monotone by himself.

This question is relevant to deciding the design of $F2_R$. Specifically it is important to determining exactly how large on average the acoustic influence of V_1 is on the onset of V_2 and, consequently, whether it is essential to incorporate this effect into the $F2_R$ formula.

5.1.2 Results

As mentioned above the present section employs linear mixed-effects modelling, using the function `lmer` in the `lme4` package (Bates et al., 2015) in the R statistical program (R Core Team, 2018). The aim is to quantify how much the slope between $F2_{onset}V_2$ and $F2_{mid}V_2$ changes once $F2_{mid}V_1$ is added to the model. Thus we begin with a model where $F2_{mid}V_2$ is the sole fixed effect and add $F2_{mid}V_1$ as a second fixed effect. In all the models to be presented, $F2_{onset}V_2$ is the dependent variable. The random effects will also be the same in all models, namely speaker (N = 20) and word (N = 140). Random slopes were used for $F2_{mid}V_1$ in the two random effects (when random slopes for both $F2_{mid}V_1$ and $F2_{mid}V_2$ were used, the model failed to converge). `REML = TRUE` for those individual models but for the chi-squared comparison of models, `REML = FALSE`, following the guidelines of Winter (2013: 12-13).

We begin with the results for a model in which the sole fixed effect is $F2_{mid}V_2$. In this model the estimated slope associated this effect and $F2_{onset}V_2$ is 0.68 (± 0.02 standard error). That is, for every 1 Bark that $F2_{mid}V_2$ is increased, $F2_{onset}V_2$ increases by approximately 0.68 Bark. When the same model is run again except that the second fixed effect of $F2_{mid}V_1$ is added, the estimated slope associated with $F2_{mid}V_2$ becomes 0.65 (± 0.03 standard error). That is, the estimated slope has changed by 0.03 Bark, a surprisingly small change.

The estimated slope for $F2_{mid}V_1$ is 0.15 (± 0.03 standard error), which is over four times smaller than the slope for $F2_{mid}V_2$. That is, for every 1 Bark that $F2_{mid}V_1$ is increased, $F2_{onset}V_2$

increases by 0.15 Bark. Thus $F2_{mid}V_1$ has less than a quarter the effect on $F2_{onset}V_2$ as does $F2_{mid}V_2$. An ANOVA was used to compare this model in which $F2_{mid}V_1$ is included as a second fixed effect with the model in which $F2_{mid}V_2$ was the sole fixed effect. This showed that the effect of $F2_{mid}V_1$ on $F2_{onset}V_2$ is highly statistically significant ($\chi^2(1) = 30.3$, $p < 0.001$). Nevertheless, influence of the preceding vowel on $F2_{onset}$ is less than a quarter of that of the following vowel.

When a model is run in which the sole fixed effect is $F2_{mid}V_1$, the slope increases from 0.15 (± 0.03 standard error) to 0.26 (± 0.04 standard error). The residual error ε associated with each model increases from 0.61 standard deviations to 0.72 standard deviations. This is in contrast to the results for $F2_{mid}V_2$, in which the slope (recall) only increased fractionally with the exclusion of $F2_{mid}V_1$ as a fixed effect, from 0.65 to 0.68. Also, ε only increased minimally from 0.61 to 0.62 standard deviations.

These results, taken together, suggest that the role of $F2_{mid}V_1$ in altering the frequency of $F2_{onset}V_2$ is highly statistically significant but small. The value 0.15 is substantially lower than the 0.26 to 0.43 values that were noted above in the cursory analysis of Öhman's (1966: 160-162) diagrams of /obo ybo odo ydo ogo ygo/.

Nevertheless the results thus far have lumped all three places of articulation into a single model. We now look at the results for each place of articulation. As before the dependent variable is $F2_{onset}V_2$ and the random effects are speaker and word (with random slopes again for $F2_{mid}V_1$).

For /d/ the slope for $F2_{mid}V_2$ when it is the sole fixed effect is 0.44 (± 0.03 standard error). When $F2_{mid}V_1$ is added as a second fixed effect, this changes to 0.43 (± 0.03 standard error). As for $F2_{mid}V_1$, its slope is 0.12 ± 0.05 when it is the sole fixed effect and shrinks somewhat to 0.10 ± 0.05 when $F2_{mid}V_2$ included as a second fixed effect. Thus the results for /d/ are similar to the overall results discussed above in that the slope for $F2_{onset}V_2$ is over four times the size of the slope for $F2_{onset}V_1$. This again indicates a minor role of the preceding vowel in determining the frequency at the onset of the following vowel.

The main difference between the results for /d/ and those for /b d g/ is that $F2_{onset}$ is less correlated with both $F2_{mid}V_2$ and $F2_{mid}V_1$. This is hardly surprising given the recurrent finding in locus-equation research (Sussman et al., 1991; Lindblom and Sussman, 2012) that /d/ has shallower slopes than the other two places of articulation. In any event, perhaps the more important point for the present chapter is that the influence of $F2_{mid}V_1$ on $F2_{onset}V_2$ both in the case of /d/ and /b d g/ is small (namely 0.10 and 0.15 respectively).

An ANOVA was run to compare the mixed-effects model in which $F2_{mid}V_1$ was included as a second fixed effect to the model in which $F2_{mid}V_2$ was the sole fixed effect, and

it was found that the difference between the two models is statistically significant ($\chi^2(1) = 4.4$, $p < 0.05$, N = 352).

We now turn to the results for /b/. When $F2_{mid}V_2$ was the sole fixed effect, the slope was 0.70 ± 0.02. When $F2_{mid}V_1$ was added as a second fixed effect, the slope of $F2_{mid}V_2$ changed only fractionally, to 0.69 ± 0.02. This is exactly parallel to the results for /d/ in which the slope again changed by just 0.01 Bark under the same conditions. As in the case of /d/, the difference between these two models is nevertheless statistically significant ($\chi^2(1) = 10.2$, $p < 0.01$, N = 264).

As for $F2_{mid}V_1$, when this was the sole fixed effect the slope was 0.39 ± 0.06, which shrunk considerably to 0.12 ± 0.03 once $F2_{mid}V_1$ was included. This 0.12 slope is almost as small as the 0.10 slope found for /d/ under the same condition.

The mixed-effects model for /g/ when random slopes were included (as done for /b d/) did not converge. Hence the results for /g/ will be presented with a model in which the slopes are fixed. The estimated slope for $F2_{onset}V_2$ when it was the sole fixed effect in the model was 0.78 ± 0.05, which shrunk to 0.74 ± 0.05 when $F2_{onset}V_1$ was included as a second fixed effect. This is a somewhat larger change than for /b d/ (0.04 as opposed to 0.01) but is nevertheless a modest change. The difference between the two models, as in the case of /b d/, is statistically significant ($\chi^2(1) = 7.7$, $p < 0.01$, N = 142).

Regarding $F2_{mid}V_1$, when this was the sole fixed effect the slope was 0.38 ± 0.08, and once $F2_{onset}$V2 was included as a second fixed effect this shrank to 0.14 ± 0.05. This is extremely similar to the picture for /b d/: when $F2_{mid}V_1$ is included the model changes only marginally but when $F2_{mid}V_2$ is included it changes considerably.

In summary, the results have shown that, although the effect of $F2_{mid}V_1$ on the frequency of $F2_{onset}V_2$ is statistically significant, it is small. It is approximately four to five times smaller than the effect of $F2_{mid}V_2$ on $F2_{onset}V_2$. This finding will influence the implementation of $F2_R$ adopted in this study.

5.1.3 Discussion

One way to think about the result in more concrete terms is that Kewley-Port and Zheng (1999), who used four vowels synthesized from a female speaker and embedded them in a sentence identification task, estimated the just-noticeable difference for formant discrimination in vowels to be 0.28 Bark. Admittedly the present examination is concerned with $F2_{onset}$, not with the steady part of the vowel as Kewley-Port and Zheng were. Presumably the just-noticeable difference would be even larger for $F2_{onset}$ than for the steady part of the vowel, since F2 tends to change more rapidly in the transition than at the steady part of the vowel, which would give

the listener fewer opportunities to estimate its frequency than when the formant is static, since a longer stimulus provides more detection opportunities (Moore, 2012: 65; Viemeister and Wakefield, 1991).

Nevertheless, if we use Kewley-Port and Zheng's 0.28-Bark figure provisionally as a rough guide, the difference in frequency between $F2_{onset}V_2$ and $F2_{mid}V_1$ in the present study's data would have to be at least 2 Bark for /g/ (slope = 0.14), 2.5 Bark for /b/ (slope = 0.12), and 3 Bark for /d/ (slope = 0.10) before the listener would stand a reasonable chance of detecting any effect of V_1 on the frequency of $F2_{onset}V_2$. Kewley-Port and Zheng also note that the smallest vowel contrast in their American English dataset is twice the JND (0.56 Bark). To generate such a change in the frequency of $F2_{onset}V_2$, the difference in frequency between $F2_{onset}V_2$ and $F2_{mid}V_1$ in the present dataset would have to be between ca. 4 and 6 Bark.

How frequently does such a large difference in F2 frequency between V_1 and V_2 occur in practice? Out of the 354 intervocalic /d/ tokens, the mean absolute difference in frequency between $F2_{onset}V_2$ and $F2_{mid}V_1$ is 1.20 Bark. There is only *one* token in which the difference is greater than 6 Bark and a further 34 tokens in which it is greater than 3 Bark. This suggests that only around 10% of the /d/ data contain a frequency difference between $F2_{onset}V_2$ and $F2_{mid}V_1$ that would be large enough to result in a big enough change to $F2_{onset}V_2$ to be detected by a listener, at least taking Kewley-Port and Zheng's (1999) work on vowel-steady-state formant discrimination as a rough guide. Given that $F2_{onset}$ discrimination could well be poorer than discrimination of F2 at the vowel steady state, the percentage of /d/ tokens in which $F2_{onset}$ is perceptibly affected by V_1 could be even less than the back-of-an-envelope 10% figure calculated above.

For /g/ there are just four cases out of 142 in which the difference between $F2_{mid}V_1$ and $F2_{onset}V_2$ is greater than 4 Bark and a further 32 in which it is greater than 2 Bark. So only approximately a quarter of the intervocalic /g/ tokens would involve a difference in frequency between $F2_{onset}V_2$ and $F2_{mid}V_1$ that would be large enough for $F2_{mid}V_1$ to change $F2_{onset}V_2$ sufficiently to meet Kewley-Port and Zheng's (1999) detectable threshold for formant frequency change.

Finally, the estimated slope between $F2_{onset}V_2$ and $F2_{mid}V_1$ in /b/, recall, was 0.12. For this to reach Kewley-Port and Zheng's just-noticeable difference, the difference in frequency between $F2_{onset}V_2$ and $F2_{mid}V_1$ would have to be at least 2.33 Bark. Of the 264 intervocalic /b/ tokens, there are 56 tokens where this is the case, or 21% of the total.

The estimates of the last three paragraphs are coarse and provisional. Nevertheless they suggest that the influence of $F2_{mid}V_1$ on $F2_{onset}V_2$ is not likely to be large, as was also suggested by the relatively shallow slopes found in the mixed-effects models.

Consequently, in the rest of this chapter the modelling of the relationship between $F2_{onset}V_2$ (henceforth '$F2_{onset}$') and $F2_{mid}V_2$ (henceforth '$F2_{mid}$') will be performed without regard to the effect of $F2_{mid}V_1$ on $F2_{onset}V_2$, since on average this effect appears to be modest enough to discount. Furthermore, it should be remembered that the above analysis – involving intervocalic /b d g/ – comprises just 758 of the 6,284 tokens in the dataset. Thus even if the influence of $F2_{mid}V_1$ on $F2_{onset}V_2$ *had* turned out to be considerable, it would have affected just 12% of the overall dataset.

The results for intervocalic /p t k/ have not been presented as there are theoretical reasons to expect the influence of $F2_{mid}V_1$ on $F2_{onset}V_2$ to be even less than those for /b d g/. This is because the distance in time between $F2_{onset}V_2$ and $F2_{mid}V_1$ is even larger in /p t k/ than in /b d g/.

	V1 duration (ms)	*Closure duration (ms)*	*VOT duration (ms)*	*V2 duration (ms)*
/b d g/	59	52	16	84
/p t k/	62	62	53	63

Table 5.1: Mean segmental durations for intervocalic /b d g/ and /p t k/.
/b d g/ N = 758; /p t k/ = 939.

Table 5.1 shows that the mean distance between $F2_{mid}V_1$ and $F2_{onset}V_2$ is 97 ms for /b d g/ but 146 ms for /p t k/. In contrast the distance between $F2_{mid}V_2$ and $F2_{onset}V_2$ is 42 ms and 32 ms respectively. Thus the figures above indicate that the influence of $F2_{mid}V_1$ on $F2_{onset}V_2$ is likely to be even less in /p t k/ than what has been established for /b d g/.

A further difference between /b d g/ and /p t k/ is in the location of $F2_{onset}V_2$ relative to the release of the burst. We can see this in the above table in terms of the voice onset time: the mean VOT for /p t k/ is 53 ms but 16 ms for /b d g/. This is a substantial difference, and as will be shown in 5.4.1 it leads to $F2_{onset}$ having a reduced role for distinguishing place of articulation in /p t k/ than in /b d g/.

To return to the main point: the acoustic influence of V_1 on $F2_{onset}$ appears to be small enough to justify not incorporating it in the $F2_R$ formula to be developed in the present chapter (Section 5.3). That is, $F2_{onset}$ and $F2_{mid}$ alone will be used as the inputs to $F2_R$.

5.2 The Curse of Dimensionality

In Chapter 2 we identified a theoretical reason why $F2_{locus}$ is worth testing, namely that the 1950s locus theory predicts that it will classify place of articulation more accurately than $F2_{onset}$. Another reason is to reduce the number of acoustic attributes. It was seen in Chapter 2 that most

existing studies have used $F2_{onset}$ and $F2_{mid}$ as features for place of articulation (locus equations). It was also noted, however, that the variation in $F2_{onset}$ is heavily dependent on $F2_{mid}$, due to the close correlation between the two (linear-regression plots for a given place of articulation typically having a slope between ca. 0.4 and 0.75). Given that $F2_{onset}$ is correlated with $F2_{mid}$ to such a considerable extent, it suggests that collapsing the two attributes into a single attribute might be viable.

In automatic speech recognition (ASR) there can be even more correlation of formant attributes than there is in locus equations in that it is common to have as input a frame (= a new sample of the speech spectrum) every 10 ms (Huckvale, 2013: 209). Thus if a vowel were 80 ms long there would be four or five frames from $F2_{onset}$ to $F2_{mid}$. The spectral envelope between neighbouring frames is often very similar: they usually contain much of the same information. In statistical terms, the information in such neighbouring frames is not independent of each other. This dependence between neighbouring frames is a major difficulty in much automatic speech recognition: in a large-scale review of the current difficulties in ASR, Morgan et al. (2013: 2) write, "it is likely that explorations of methods for properly representing the conditional dependence between frames and phones (given the state) should have a major effect" in improving the robustness of ASR.

The present study is not concerned with ASR but it *is* concerned with the general question (common to both ASR and phonetic science) of how to represent spectral information (including formant information) in a compact manner. Huckvale (1996: 4) describes the case for dimension minimization as follows:

"The experience of pattern recognition shows that it is very hard to deal with large observation vectors: simply because there are too many possible combinations of fine-grained detail. In high dimensional spaces it is impossible to measure distances because the relative importance of the dimensions cannot be estimated."

This problem is known as the "curse of dimensionality" (Bellman, 1961).

Another challenging feature of formant transitions is that they smear together information about at least two phonemes, the plosive and the following vowel (and to some degree also the preceding vowel, as Öhman (1966) showed). So not only is the information in neighbouring frames along the formant transition not independent of each other, but the formant transitions contain information about more than one phoneme. This is a further reason why it would be helpful to develop a means of separating out as far as possible the information in the transitions associated with the plosive from that associated with the adjoining segments.

To summarize: all else being equal, the fewer the number of dimensions needed to successfully distinguish categories, the better.

5.3 A Formula for Finding $F2_{locus}$

In Section 2.2.1 we derived a key prediction from the 1950s locus theory: $F2_{locus}$ should classify place of articulation more accurately than $F2_{onset}$. Let us begin by re-examining the diagram illustrating the theory:

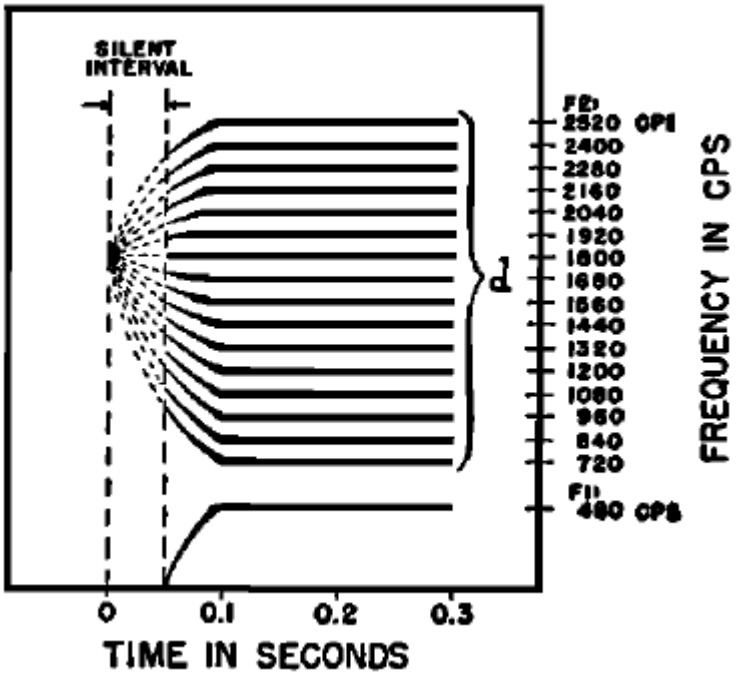


Figure 5.3: Schematic diagram of the locus theory for a /d/ that is paired with a range of vowels that vary in backness.
That is, the vowels vary in their F2 frequency but not their F1 frequency. The F2 transitions for all the vowels begin at the same frequency of 1,800 cps (1,800 Hz), at least if one traces their trajectory to an unobserved point in time approximately 50 ms prior to the vowel onset. This point is known as the F2 locus frequency ($F2_{locus}$). $F2_{locus}$ is believed to be different for the three places of articulation. From Delattre et al. (1955: 771).

In the present study we simplify our task of obtaining $F2_{locus}$ by using just two points along the formant trajectory, $F2_{onset}$ and $F2_{mid}$. Naturally $F2_{locus}$ could also be obtained using the full formant contour but in this study our aim is to find a way of obtaining $F2_{locus}$ knowing only $F2_{onset}$ and $F2_{mid}$. Our task is this: how can we use $F2_{mid}$ to modify $F2_{onset}$ in such a way as to change it into $F2_{locus}$?

To answer this question, let us turn our attention to Figure 5.3 above. What generalization can we make about the sixteen F2 transitions? Answer: the larger the difference in frequency between $F2_{onset}$ and $F2_{mid}$, the larger the difference in frequency between $F2_{onset}$

and $F2_{locus}$. This means that if we want to change $F2_{onset}$ into $F2_{locus}$ using $F2_{mid}$, the degree to which $F2_{onset}$ will have to change depends on the size of the frequency difference between $F2_{onset}$ and $F2_{mid}$. In other words, the first part of our technique is to subtract $F2_{onset}$ from $F2_{mid}$:

(1) $F2_{difference} = F2_{mid} - F2_{onset}$

The second part of the technique is to subtract this $F2_{difference}$ from $F2_{onset}$:

(2) $F2_{reconstructed} = F2_{onset} - F2_{difference}$

The output of (2) extrapolates the F2 transition backwards in time. Remember, however, that because we do not observe $F2_{locus}$, there is nothing to tell us exactly *how* far back in time we should go to obtain $F2_{locus}$. In Figure 5.1 above, the amount of time required (i.e. the part labelled 'silent interval') is posited as 50 ms, but this is a schematic diagram based on artificial stimuli, not an empirical fact. Thus it seems wise to run a variety of $F2_{locus}$ formulae in which the degree to which $F2_{difference}$ modifies $F2_{onset}$ is varied by use of a constant. Let us rewrite (2) as follows:

(3) $F2_{reconstructed} = F2_{onset} - (F2_{difference} \times c)$

Note that we shall refer to the family of acoustic attributes derived from (3) as '$F2_{reconstructed}$', or '$F2_R$' for short. The constant *c* in (3) could be thought of as a time constant, though in practice this notion is complicated by the fact that in the present data set, the distance in time between $F2_{onset}$ and $F2_{mid}$ varies from one token to another (since vowels vary in duration). This means that the amount of time that *c* corresponds to would vary from one token to another. Hence it is probably best to think of *c* as an arbitrary constant. Nevertheless, it *is* possible to make a variant of (3) in which the value of *c* is varied depending on the distance in time between $F2_{onset}$ and $F2_{mid}$, which will be explored in 5.4.8.

For now, then, we take *c* to be an arbitrary constant. A range of variants of (3) is run in which the value of *c* is varied from 0 to 3 in increments of 0.2, which yields 16 results. This range of tested values is sufficiently large to examine the effect of $F2_{difference}$ on $F2_{onset}$ thoroughly. When $c = 0$, of course, it means that our $F2_{locus}$ acoustic attribute is in fact $F2_{onset}$. However, as the value of *c* is gradually increased, the acoustic influence of $F2_{difference}$ on $F2_{onset}$ is increasing. The value of *c* for which the classification accuracy is greatest could be regarded as $F2_{locus}$ (at least from the pragmatic point of view of optimizing the quality of the F2 feature),

but as we shall see in 5.4.4 $F2_R$ in practice it yields frequencies that are different from those schematized in the $F2_{locus}$ concept in Figure 5.3. Therefore it is best to think of $F2_R$ as similar to but different from $F2_{locus}$.

Note that a given value of *c* will be indicated by adding it to the end of the attribute's name. For example, the variant of (3) in which *c* = 1.2 will be referred to as '$F2_R$1.2'.

5.3.1 Tests of Statistical Significance

In this chapter and the next, the performance of many acoustic attributes on discriminant analysis will be compared to each other in terms of their accuracy at distinguishing place of articulation. However, the fact that one attribute classifies more accurately than another does not necessarily mean that the difference in classification accuracy is statistically significant.

To test whether the difference in classification accuracy between two discrimint analysis classifications is statistically significant, the present thesis employs the McNemar test (more precisely, the mid-*p* variant of the test). In a highly cited paper in the machine learning field, Dietterich (1998) investigated five hypothesis-test statistics for comparing the classifications of two classifiers. For cases in which the data are limited and the classifier can only be tested a single time (i.e. only one test set), he recommends the McNemar test. Fagerland et al. (2013a) compare four different variants of the McNemar test and provide a mathematical description of each (pp. 2-3). They recommend the mid-*p* test and provide an ancillary document for how to compute this variant of the McNemar test in different statistical packages such as SPSS and R (Fagerland et al., 2013b).

The statistic takes as input some of the information in a contingency table. The contingency table compares the two classifiers and counts the number of cases that (1) both classifiers classified correctly; (2) classifier A classified correctly but classifer B classified incorrectly; (3) classifier B classified correctly but classifier A classified incorrectly, and (4) both classifiers classified incorrectly. Of these four values, it is (2) and (3) that matter. Under the null hypothesis, these two numbers should be the same (Dietterich, 1998: 1901). The formula is as follows (ibid.):

$$\frac{(|n_{01} - n_{10}| - 1)^2}{n_{01} + n_{10}}.$$

The numerator subtracts (2) from (3), takes its absolute value, subtracts 1, and squares the result. The denominator adds (2) and (3). The resulting value is distributed as χ^2 with 1 degree of freedom. If it is greater than 3.84 the null hypothesis can be rejected (ibid.).

5.4 Results

5.4.1 F_R in CV Formant Transitions

In this section we begin our exploration of the F_R concept on F1 ('$F1_R$'), F2 ('$F2_R$'), and F3 ('$F3_R$'). The phonetic context examined (N = 1,535) is all instances of /b d g/ before a non-schwa vowel. (The schwa data is dealt with in 5.4.7.)

5.4.1.1 $F2_R$

Here are the results for $F2_R$:

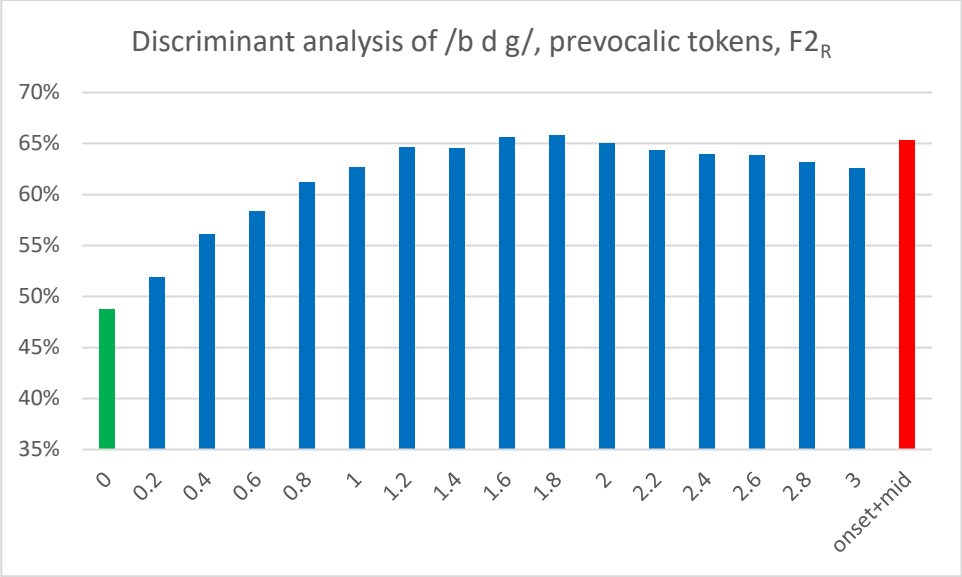


Figure 5.4: Discriminant analysis classification accuracy of $F2_R$ for distinguishing prevocalic tokens of /b d g/. The bar in green (termed '$F2_R0$') indicates the classification accuracy of $F2_{onset}$; the bar in red indicates the classification accuracy when both $F2_{onset}$ and $F2_{mid}$ are used; the bars in blue represent the variants of $F2_R$, namely $0 < c \geq 3$ and increases in increments of 0.2. N = 1,535.

The classification accuracy for all variants of $F2_R$ – which ranges between 51.6% and 65.8% – is better than $F2_{onset}$ (48.8%, i.e. the green bar represented by '0'). Thus the answer to the question in 2.2.1 as to whether $F2_{locus}$ classifies place of articulation better than $F2_{onset}$ is yes. The greatest improvements to the classification accuracy can be seen for those variants of $F2_R$ in which the value of *c* is between 0.2 and 1.2. As the value of *c* is increased beyond 1.2, the improvement in classification accuracy levels off; values of $F2_R$ between 1.2 and 2 perform similarly.

What is more surprising about $F2_R$ than the fact that it classifies place of articulation better than $F2_{onset}$ is the fact that it classifies place of articulation as well as $F2_{onset}$ + $F2_{mid}$. Indeed, $F2_R1.8$ (the strongest performing variant of $F2_R$ in Figure 5.2) classifies 65.8% of the 1,535 /b d g/ tokens correctly, which is marginally higher than the 65.3% accuracy that is

obtained when $F2_{onset}$ and $F2_{mid}$ are used together (= the red bar in the above histogram). This 0.5-percentage-point difference is not statistically significant. Instead, the significance of the result lies in the fact that a single attribute is classifying as well as two heavily correlated attributes, which means that the collapsing of the two attributes into a single attribute seems to be viable.

Nevertheless one might wonder whether the improvements in classification accuracy of $F2_R$ over $F2_{onset}$ are statistically significant. The McNemar test was used for this purpose. In each case the input to the McNemar test was the casewise classification results (i.e. a column of correctly and incorrectly classified tokens) of two adjoining variants of $F2_R$. Thus $F2_{onset}$ was compared with $F2_R0.2$, $F2_R0.2$ was compared with $F2_R0.4$, $F2_R0.4$ was compared with $F2_R0.6$, and so on. This series of tests revealed that the improvement in classification accuracy yielded by $F2_R$ in Figure 5.4 above was statistically significant at the $p < 0.001$ level for all increments of $F2_R$ between 0 and 0.8. The difference in classification accuracy between $F2_R0.8$ and $F2_R1.0$ was also significant, though only at the $p < 0.01$ level. The difference in classification accuracy between $F2_R1.0$ and $F2_R1.2$ was highly significant ($p < 0.001$)

However, all increments of $F2_R$ beyond $F2_R1.2$ are not statistically significant at the $p < 0.01$ level, being either not statistically significant at all or only statistically significant at the $p < 0.05$ level. Given the high number of statistical significance tests performed, it seems wise not to set the threshold for statistical significance at $p < 0.05$ but rather at $p < 0.01$, which will reduce the risk of a Type 1 error. Following this criterion, it can be said that incrementing $F2_R$ beyond $c = 1.2$ does not improve the classification accuracy to a statistically significant degree.

The classification accuracy of $F2_{onset}+F2_{mid}$ (i.e. the red bar in Figure 5.4) was compared with the classification accuracy of $F2_R1.8$ (that is, the variant of $F2_R$ that yielded the highest classification accuracy). This result ($p = 0.61$) is not significant, and likewise when the classification accuracy of $F2_{onset}+F2_{mid}$ was compared with $F2_R1.2$ ($p = 0.59$). In contrast when $F2_{onset}+F2_{mid}$ was compared to $F2_R1.0$ and $F2_R0.8$, the difference in classification accuracy was significant, at the $p < 0.05$ and $p < 0.001$ levels respectively. In conclusion, the classification accuracy of $F2_R$ can be regarded as equivalent to that of $F2_{onset}+F2_{mid}$ provided that $c = 1.2$ or more, with $c = 1.0$ being statistically significantly lower only if one sets the threshold of significance at $p < 0.05$ rather than the $p < 0.01$ used in the present study.

The results discussed above were for /b d g/. Here are the equivalent results for /p t k/:

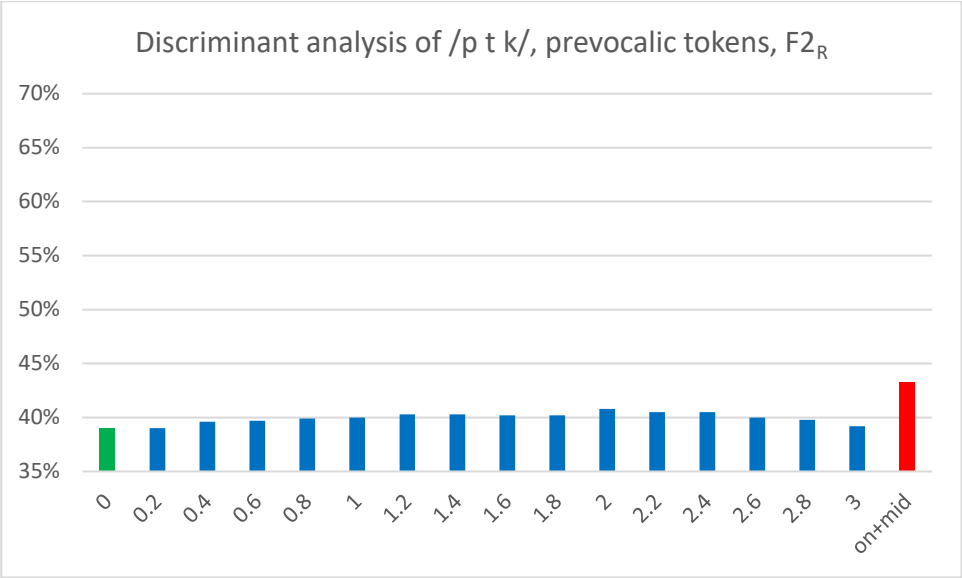


Figure 5.5: Discriminant analysis classification accuracy of $F2_R$ for distinguishing prevocalic tokens of /p t k/. The bar in green indicates the classification accuracy of $F2_{onset}$; the bar in red indicates the classification accuracy when both $F2_{onset}$ and $F2_{mid}$ are used; the bars in blue represent the variants of $F2_R$, namely $0 < c \leq 3$ and increases in increments of 0.2. N = 1,460.

The classification accuracy of $F2_R$ does not amount to much of an improvement over $F2_{onset}$. $F2_{onset}$ + $F2_{mid}$ themselves classify poorly (43.3%), which is 22 percentage points less than was found in Figure 5.4 above for /b d g/. Thus $F2_R$ might be performing poorly here, but the ingredients from which $F2_R$ is made – $F2_{onset}$ and $F2_{mid}$ – are themselves performing poorly.

In 5.1.3 a prediction was laid out for why formant information would classify the place of articulation of voiceless stops poorly, namely the fact that, due to their longer aspiration, $F2_{onset}$ occurs several tens of milliseconds after the release of the burst in /p t k/ (Fant, 1973: 64). The mean aspiration duration for /p t k/ in the present dataset is 22.4 ms (N = 2,830) whereas for /b d g/ it is just 6.2 ms (N = 2,641).

The focus in the rest of this chapter, then, will be on /b d g/.

5.4.1.2 $F3_R$

Figure 5.6 below displays the results for the same context as Figure 5.4 above, but this time for F3:

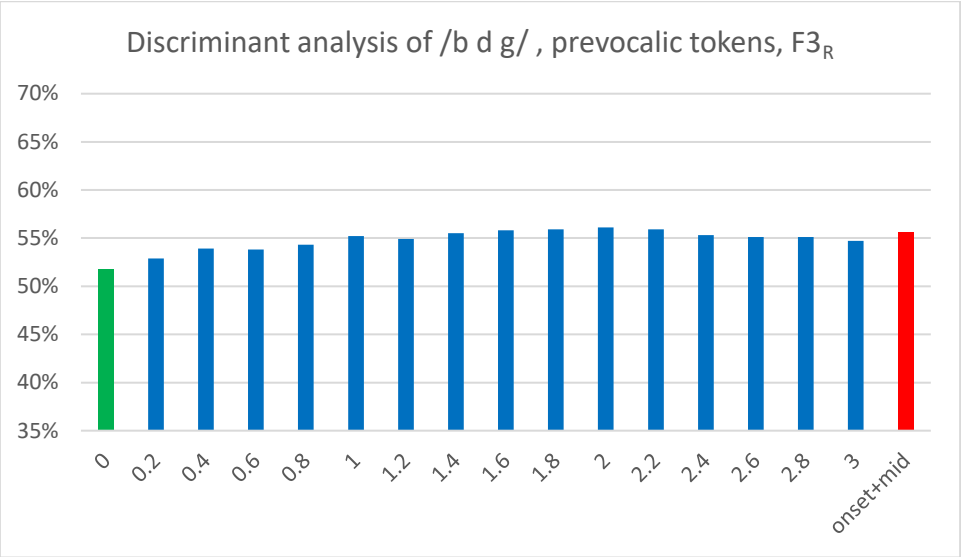


Figure 5.6: Discriminant analysis classification accuracy of $F3_R$ for distinguishing prevocalic tokens of /b d g/. The bar in green indicates the classification accuracy of $F3_{onset}$; the bar in red indicates the classification accuracy when both $F3_{onset}$ and $F3_{mid}$ are used; the bars in blue represent the variants of $F3_R$, namely c = 0.2 to 3 and increases in increments of 0.2. N = 1,535.

All variants of $F3_R$ classify place of articulation more accurately than $F3_{onset}$. The strongest variant ($F3_R2$) classifies 56.1% of the 1,535 tokens correctly, which is marginally higher than combining $F3_{onset}$ and $F3_{mid}$ separately (55.6%). The improvement in classification is smaller for $F3_R$ than $F2_R$, though this does not seem to be due to $F3_R$ per se. Rather, the contribution of $F3_{mid}$ to improving the classification of $F3_{onset}$ is smaller (4% as against 17%) than the improvement yielded when $F2_{mid}$ is combined with $F2_{onset}$. (This can be seen by comparing the green and red bars in Figure 5.6 with their equivalents in Figure 5.4.) Thus the ability of $F3_R$ to classify more accurately than $F3_{onset}$ is dependent on the accuracy of $F3_{mid}$, and analogous comments can be made for $F2_R$. This is unsurprising in that $F2_R$ and $F3_R$ are a modification of $F2_{onset}$ and $F3_{onset}$ based on $F2_{mid}$ and $F3_{mid}$.

In terms of statistical significance, almost none of the neighbouring increments of $F3_R$ in Figure 5.6 above yields a statistically significant change to the classification accuracy than the previous increment. The only minor exceptions are as follows: the difference between $F3_{onset}$ and $F3_R0.2$, $F3_R0.8$ and $F3_R1.0$, and $F3_R2.2$ and $F3_R2.4$ are significant at the $p < 0.05$ level. However, as noted in 5.4.1.1 above, the threshold of statistical significance has been set at $p < 0.01$ due to the large number of significance tests performed.

The classification accuracy of $F3_{onset}$+$F3_{mid}$ was compared with a variety of $F3_R$ variants. It was found that any variant of $F3_R$ greater than or equal to 0.4 did not yield a statistically significantly different classification accuracy from using $F3_{onset}$+$F3_{mid}$. Thus nearly all of the variants of $F3_R$ shown in Figure 5.6 above have a statistically equivalent classification

accuracy to $F3_{onset}$+$F3_{mid}$. The only exceptions are $F3_R0.2$ – whose classification accuracy is statistically significantly lower than that of $F3_{onset}$+$F3_{mid}$ at the $p < 0.05$ level – and $F3_R0$ (i.e. $F3_{onset}$), $p < 0.01$. Nevertheless the overall picture is that incorporating $F3_{mid}$ in the classification (whether in the form of $F3_{onset}$+$F3_{mid}$ or $F3_R$) boosts the classification accuracy to only a minor degree, and in most cases this improvement is not statistically significant. This finding dovetails with Lindblom's (1990) decision not to use $F3_{mid}$ in his multidimensional formant-based classification space for distinguishing /b d g/.

In any event, the most important observation from Figures 5.4 and 5.6 is that $F2_R$ and $F3_R$ classify more accurately than $F2_{onset}$ and $F3_{onset}$, and indeed classify at rates that are comparable those for $F2_{onset}$ + $F3_{onset}$ and $F2_{mid}$ + $F3_{mid}$ combined. Thus the two features can be collapsed into one without compromising classification accuracy.

5.4.1.3 $F1_R$

For the sake of completeness we now present the results when the F_R technique is applied to F1. We noted briefly in Chapter 2 that there is little theoretical reason to expect F1 to contribute much (if anything) to the classification of place of articulation. Nevertheless, let us turn to the results for $F1_R$:

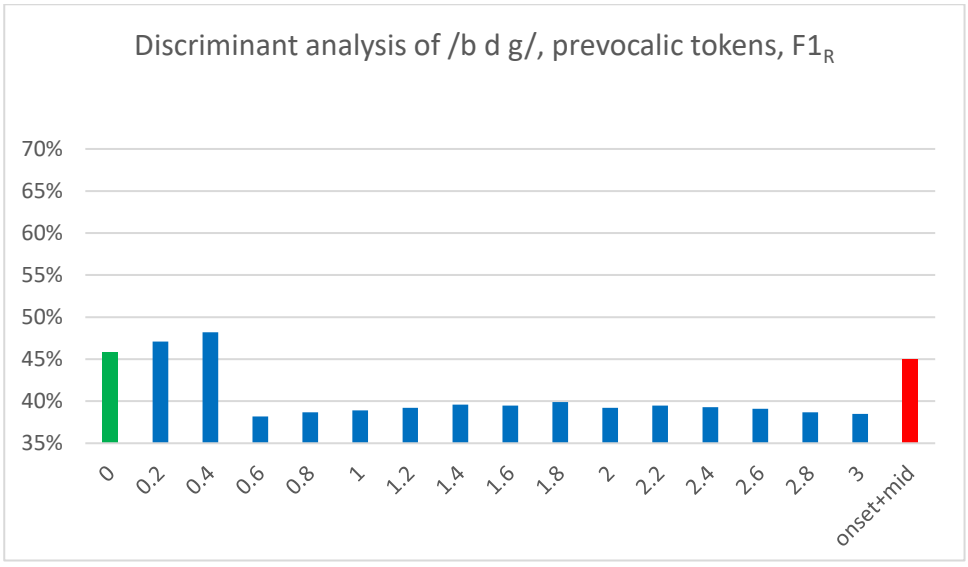


Figure 5.7: Discriminant analysis classification accuracy of $F1_R$ for distinguishing prevocalic tokens of /b d g/. The bar in green indicates the classification accuracy of $F1_{onset}$; the bar in red indicates the classification accuracy when both $F1_{onset}$ and $F1_{mid}$ are used; the bars in blue represent the variants of $F1_R$, namely c = 0.2 to 3 and increases in increments of 0.2. N = 1,535.

$F1_{onset}$ does classify place of articulation better than chance. However, including $F1_{mid}$ in the classification with $F1_{onset}$ weakens the classification, and this is reflected in the classification rates of all but two of the variants of $F1_R$ (namely $F1_R0.2$ and $F1_R0.4$). Thus it seems that,

unsurprisingly, F1 formant transitions contribute little to classifying place of articulation accurately since, even at its strongest, its classification is 8 percentage points lower than F3 and 18 percentage points lower than F2. Consequently the remainder of this chapter will centre on F2 and F3 only.

5.4.2 Formant Frequency Normalization

5.4.2.1 Theoretical Background

In this section, the examination of F2 above is repeated but this time employing techniques of normalization. Vowel normalization is not the primary topic of this chapter, and a full comparison of the many normalization techniques that have been used in phonetics over the decades is beyond the remit of the present work (see Chapter 6 of Flynn, 2012 and Adank et al., 2004 for such investigation). Instead this section will confine itself to two specific aspects of vowel normalization:

(1) How much does it improve the classification accuracy of place of articulation?

(2) Does normalization change the value of *c* for which the classification accuracy of $F2_R$ is highest?

Speakers vary in the length of their vocal tracts. Everything else being equal, the longer the vocal tract the lower the speaker's mean formant frequencies (Flynn, 2012; Turner et al., 2009). Certain articulatory settings are also associated with changes to the formant frequencies, e.g. lowered larynx results in lower mean formant frequencies relative to a neutral laryngeal setting (Laver, 1980: 29, Lindblom and Sundberg, 1971: 1176). One aim of normalization is to reduce or eliminate such individual differences (Adank et al., 2004).

As noted there are very many vowel-formant normalization techniques already in use. Because normalization is not the primary topic of this chapter, this section is confined to examining four methods of normalization, to wit:

(1) $F2_R - \mu F2_{sex}$

(2) $F2_R - \mu F2_{individual}$

(3) $F2_R - \mu F3_{sex}$

(4) $F2_R - \mu F3_{individual}$

The first, $\mu F2_{sex}$, consists of obtaining the mean F2 frequency for each sex (based on all tokens of $F2_{mid}$) and subtracting it from the $F2_R$ value of each token. The second method of normalization, $F2_R - \mu F2_{individual}$, uses the same technique as the first but instead of obtaining the mean F2 value of each *sex*, it uses the mean F2 value of each *individual speaker*. That is,

there are twenty values of $\mu F2_{individual}$ in the present data set (one for each of the twenty speakers) as opposed to just two values for $\mu F2_{sex}$ (one for males, one for females). The third method of normalization, $\mu F3_{sex}$, is equivalent to $\mu F2_{sex}$ except that mean F3 values are used rather than F2 values. The fourth and final method, $\mu F3_{individual}$, is equivalent to $\mu F2_{individual}$ except that the values are from F3 rather than F2.

There are three theoretical motivations behind the examination of these four normalizations. The reason for comparing F3 and F2 is that F3 varies less from one vowel to another than F2: the standard deviation of $F3_{mid}$ in vowels in the present dataset is 0.84 Bark whereas that of $F2_{mid}$ is 1.91 Bark (N = 3,172). Therefore, when a listener is listening to an unfamiliar voice (and lacking visual cues to the person's identity), obtaining an accurate estimate of the speaker's mean formant frequencies could be done more rapidly for F3 than for F2. This suggests that mean F3 would be a better benchmark for normalizing formant-frequency tokens than F2. Nevertheless, the matter is empirical in nature and thus both methods of normalization will be trialled.

As for the comparison of male-female normalization with individual-speaker normalization, the reasoning is as follows. The difference between male and female F0 is so large that a voice can be identified as male or female almost instantly: Baken and Orlikoff (2000: 177) state the F0 range in male speech runs from 85 to 155 Hz, whereas for females the range is from 165 to 255 Hz. In contrast formant frequencies overlap greatly between the genders and the difference in mean values is much smaller (10 to 20% on average; Titze, 1994: 187). This raises the following question: to what extent is it enough to know a speaker's sex to normalize their formant frequencies? Or to put it differently: how much of a gain in classification accuracy is yielded by tailoring the normalization to the individual speaker relative to the more rapidly achieved task of normalizing by the speaker's sex?

The last two paragraphs have explained the reason for comparing F2 and F3 as normalizers as well as the comparison of normalization by speaker sex and individual speaker. However, the choice of subtraction in the normalization formula also requires explanation. Turner et al. (2009: 2) outline the 'fixed-formant-pattern hypothesis': when plotted on a logarithmic frequency scale, the formant pattern of a given vowel for speakers with different vocal tract sizes is expected to be the same, the only difference being in the location of that pattern along the frequency axis. Figure 5.8 illustrates the 'formant pattern' notion:

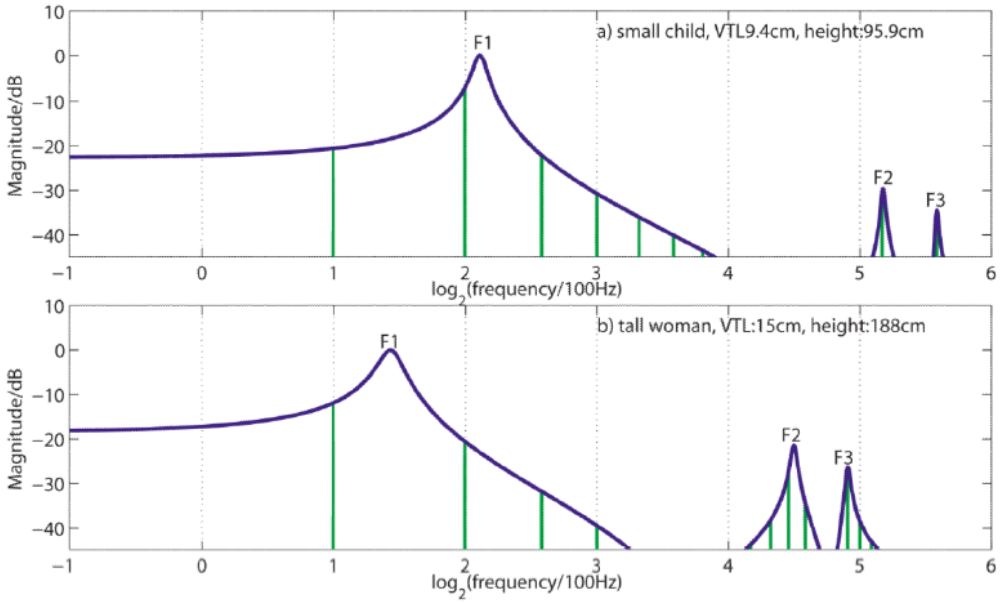


Figure 5.8: Schematic illustration of the fixed-formant-pattern hypothesis (Turner et al., 2009). Magnitude spectra (vertical lines) and spectral envelopes (smooth lines) of two synthetic /i/ vowels like those that might be produced by (a) a small child and (b) a tall woman. Note how all three formants have shifted to the same degree along the log-frequency axis for the tall woman relative to the small child, which is the theoretical assumption underpinning the fixed-formant-pattern hypothesis. Adapted from Turner et al. (2009: 17).

The present study does not employ the log-frequency scale; instead it employs the Bark scale. However, above 1,000 Hz the Bark scale is reasonably close to a log-frequency scale (compare 10^3 Hz = 8.5 Bark; $10^{3.5}$ Hz = 16.0 Bark; 10^4 Hz = 24.0 Bark). Given that F2 is normally greater than 1,000 Hz, we can treat the Bark-transformed F2 and F3 frequencies of the present data as close enough to logarithmic to conform to the fixed-formant-pattern hypothesis.

For the normalization formulas (4)-(7) above, the fixed-formant-pattern hypothesis leads us to expect that when mean F2 or F3 is subtracted from the formant frequencies associated with a given vowel, the difference in frequency should be approximately the same regardless of the speaker.

In any event, the aim is to identify which of the four normalization techniques improves the classification accuracy of $F2_R$ and $F3_R$ the most. A secondary aim is to see whether normalization changes the value of *c* at which the performance of $F2_R$ peaks. Recall that the classification results in the previous section involved classifying together the speech of 20 speakers without any normalization, not even for sex. The results of the present section, then, should be taken as a better indicator of the value of *c* for optimizing the classification of place of articulation.

5.4.2.2 Results on $F2_R$

We begin with the results for $F2_R - \mu F2_{sex}$:

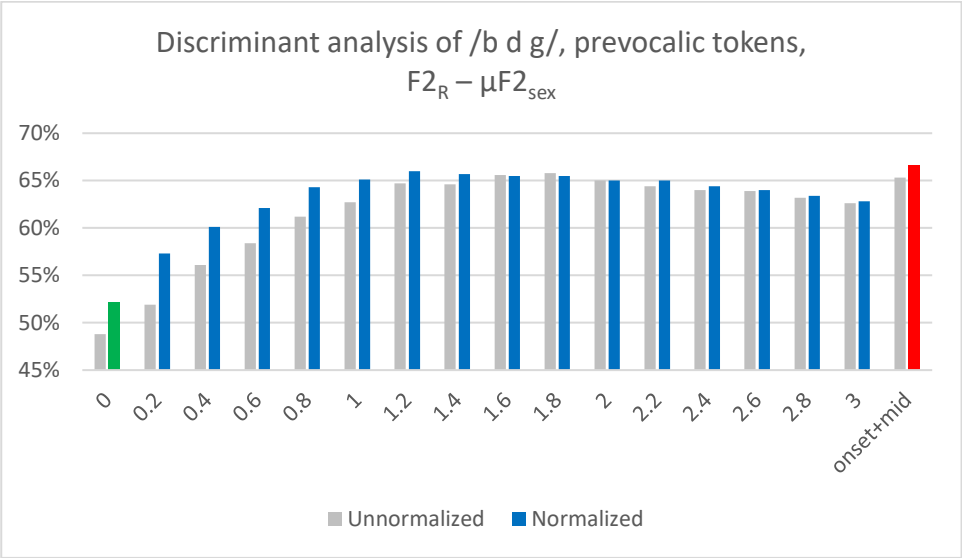


Figure 5.9: Discriminant analysis classification accuracy of normalized $F2_R$ for distinguishing prevocalic tokens of /b d g/.

The normalization consists of subtracting the mean $F2_{mid}$ frequency of the speaker's sex ($\mu F2_{sex}$) from each token (F2). The bar in green indicates the classification accuracy of $F2_{onset}$; the bar in red indicates the classification accuracy when both $F2_{onset}$ and $F2_{mid}$ are used; the bars in blue represent the variants of $F2_R$, namely c = 0.2 to 3 and increases in increments of 0.2. N = 1,535.

This normalization method is an improvement over having no normalization: across the 17 conditions in Figure 5.9 the mean improvement is 1.5 percentage points. However, the peak classification accuracies of the normalized and unnormalized data are very similar: c = 1.2 (66.0%), which is very similar to the peak unnormalized value (c = 1.8, 65.8%).

Furthermore, when each variant of $F2_R - \mu F2_{sex}$ is compared to the equivalent $F2_R$ variant in Figure 5.9 (e.g. $F2_R1.0$ versus $F2_R1.0 - \mu F2_{sex}$), the difference in classification accuracy is in all nearly all cases not statistically significant. The only major exception is $F2_R0.2 - \mu F2_{sex}$ versus $F2_R0.2$ in which the former classifies significally better than the latter at the $p < 0.01$ level. The variants $F2_R0 - \mu F2_{sex}$, $F2_R0.4 - \mu F2_{sex}$, and $F2_R0.6 - \mu F2_{sex}$ also classify significantly better than the equivalent non-normalized variants but the difference is only significant at the $p < 0.05$ level. In conclusion, there is at best weak evidence that subtracting $\mu F2_{sex}$ from F2 yields a meaningful improvement in classification accuracy.

Here are the results for $F2_R - \mu F2_{individual}$:

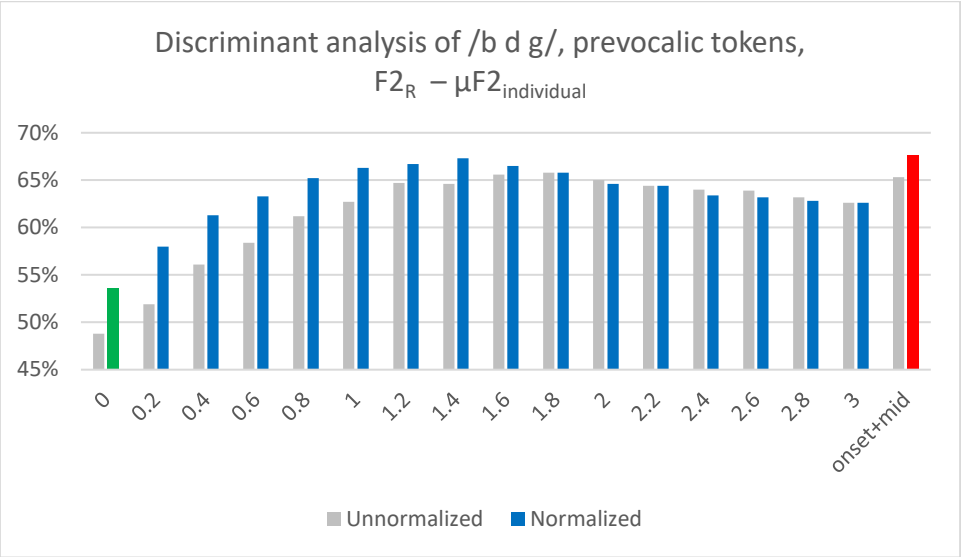


Figure 5.10: The same classification conditions as described in Figure 5.9 except that the mean F2 value used in the normalization has been calculated separately for each speaker.
As opposed to using the same mean for all speakers of a given sex. N = 1,535.

The performance of $\mu F2_{individual}$ is better than that of the unnormalized data, which is unsurprising. The more interesting question is to what extent the attribute is an improvement over $\mu F2_{sex}$. In other words, to what extent does tailoring normalization to an individual speaker improve the classification accuracy over tailoring it to the speaker's sex? The answer (for F2 at least) is: not much. The mean classification accuracy of $\mu F2_{individual}$ over the 17 conditions is 63.7% as against 63.2% for $\mu F2_{sex}$, an improvement of 0.5 percentage points. This is one third the size of the 1.5 percentage-point improvement of $\mu F2_{sex}$ over the unnormalized data. Thus it appears that most of the improvement in classification accuracy comes from normalizing by speaker sex, with diminishing returns for normalizing by individual speaker. Ten times as much normalization (20 normalizations for each speaker as against two normalizations for each sex) adds only an extra third onto the improvement in classification accuracy.

In terms of their peak classification accuracies, the difference between the two normalization techniques is somewhat larger, namely 1.3 percentage points ($\mu F2_{individual}$ peaks at 67.3% when c = 1.4 whereas $\mu F2_{sex}$ peaks at 66.0% when c = 1.2).

A common trend in both normalization techniques is that the peak classification accuracy of $F2_R$ (for c = 1.2 or 1.4) is very similar to using $F2_{onset}$ and $F2_{mid}$ as classifiers: for $\mu F2_{sex}$ the peak classification (66.0%) is 0.6 percentage points less than $F2_{onset} + F2_{mid}$, while in $\mu F2_{individual}$ the peak classification (67.3%) is 0.3 percentage points less than $F2_{onset} + F2_{mid}$. To a first approximation, then, $F2_R$ yields about the same classification accuracy as $F2_{onset} + F2_{mid}$.

However, in terms of statistical significance, the picture with $\mu F2_{individual}$ is similar to that shown above for $\mu F2_{sex}$. When the classification of each variant of F2 – $\mu F2_{individual}$ is

compared to its non-normalized equivalent, the difference in classification between the two equivalents is in most cases not statistically significant. As with $\mu F2_{sex}$, the only exceptions to this generalization are for low values of $F2_R$: the classification accuracies of $F2_R0 - \mu F2_{individual}$ and $F2_R0.2 - \mu F2_{individual}$ are statistically significantly higher than their non-normalized equivalents ($p < 0.001$) and the same is true of $F2_R0.4 - \mu F2_{individual}$ and $F2_R0.6 - \mu F2_{individual}$ ($p < 0.01$). However, for all other values of $F2_R$ the normalization does not yield a statistically significant improvement in classification accuracy. In conclusion, there is at best modest evidence that subtracting $\mu F2_{individual}$ improves the classification accuracy of F2 to a statistically significant degree.

Here are the results for the two methods that use mean F3, beginning with $\mu F3_{sex}$:

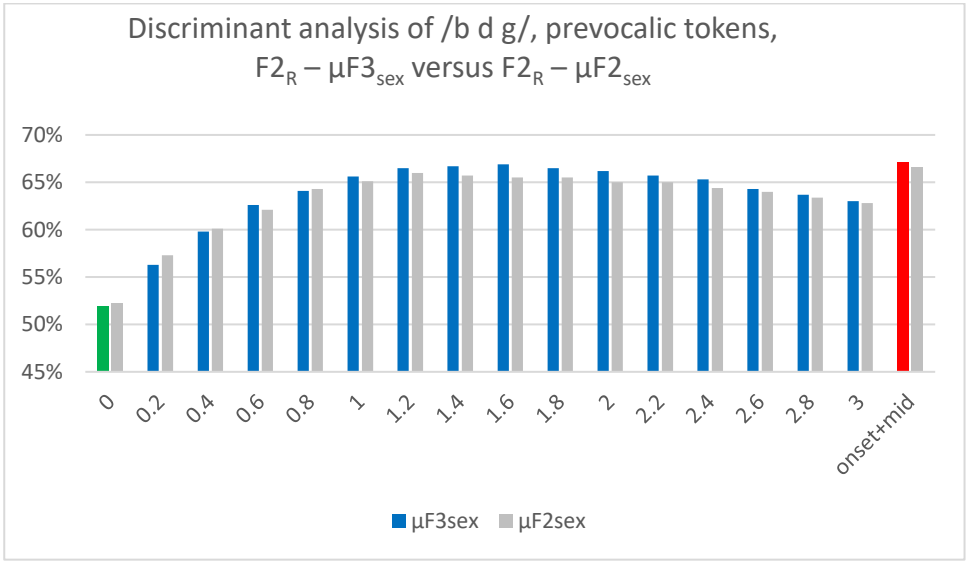


Figure 5.11: The same classification conditions as described in Figure 5.9 except that the mean formant value is F3 rather than F2.
N = 1,535.

It is apparent that the two normalization methods perform very similarly, though with the optimal performance of $\mu F3_{sex}$ being better than that of $\mu F2_{sex}$ (66.9% for $F2_R1.6 - \mu F3_{sex}$ as against 66.0% for $F2_R1.2 - \mu F2_{sex}$). The classification accuracy of $\mu F3_{sex}$ is greater than that of $\mu F2_{sex}$ for 13 out of the 17 conditions shown in Figure 5.11. The average classification of the two is 63.7% and 63.2% respectively.

In terms of statistical significance, the picture with $\mu F3_{sex}$ is broadly similar to the one found for $\mu F2_{sex}$ and $\mu F2_{individual}$ above. That is, it is only for low values of $F2_R$ that there is evidence that subtracting $\mu F3_{sex}$ yields a statistically significant improvement in classification, namely for $F2_R0$ ($p < 0.01$) and $F2_R0.2$, $F2_R0.4$, and $F2_R0.6$ ($p < 0.05$). In sum, there is only slight evidence that subtracting $\mu F3_{sex}$ from F2 improves the classification accuracy in a statistically significant manner.

Here are the results of the final normalization method, $F2_R - \mu F3_{individual}$:

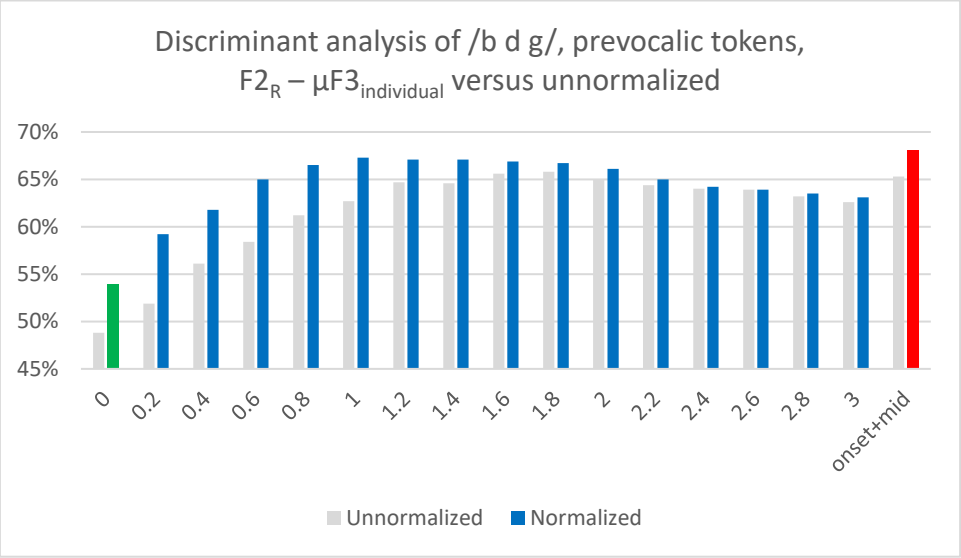


Figure 5.12: Comparison of $F2_R - \mu F3_{individual}$ and the unnormalized data under the same conditions as described in Figure 5.9.

N = 1,535.

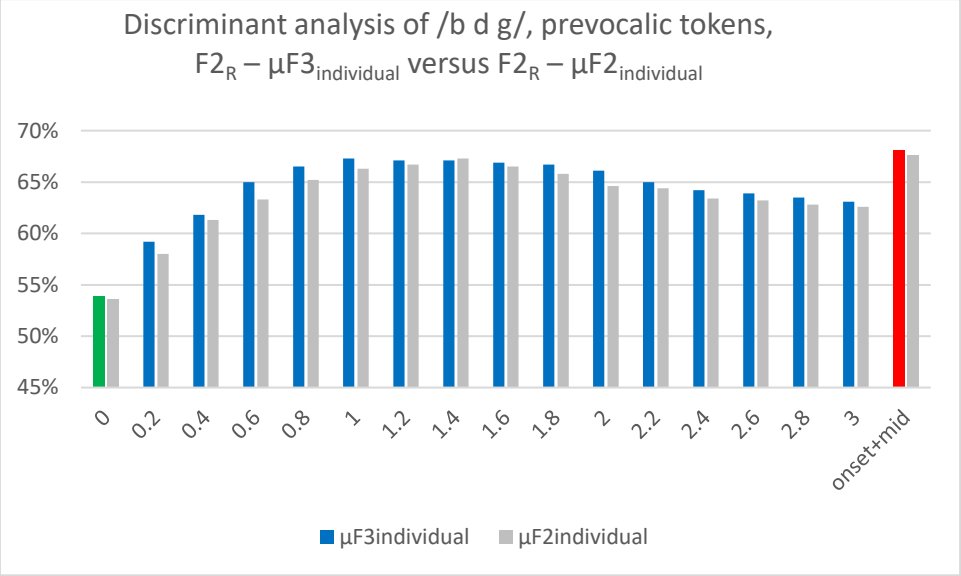


Figure 5.13: Comparison of $F2_R - \mu F3_{individual}$ and $F2_R - \mu F2_{individual}$ under the same conditions as described in Figure 5.9.

N = 1,535.

This normalization method yields the greatest improvement over the unnormalized condition of the four kinds of normalization, with the improvement being particularly evident for values of *c* between 0 and 1 (for these six variants, $F2_R - \mu F3_{individual}$ has an average classification accuracy of 62.3%, which is a 5.8 percentage-point improvement over unnormalized $F2_R$ in the same conditions, as well as being a 2.1 percentage-point improvement over $\mu F3_{sex}$ or $\mu F2_{sex}$, and a 1.0 percentage-point improvement over $\mu F2_{individual}$). Over the entire 17 conditions the

mean classification accuracy of $\mu F3_{individual}$ is 64.4%, which is 0.7 percentage points higher than $\mu F3_{sex}$ and $\mu F2_{individual}$, and 1.2 percentage points higher than $\mu F2_{sex}$.

In terms of statistical significance, the trend for normalization to only yield statistically significant improvements in classification for relatively low values of $F2_R$ is again apparent. However, the statistical significance is higher than for the other three normalizations, being at $p < 0.001$ for $F2_R0$, $F2_R0.2$, $F2_R0.4$, and $F2_R0.6$. For $F2_R0.8$ and $F2_R1.0$ the improvement in classication is again significant ($p < 0.01$). There is thus stronger evidence for $\mu F3_{individual}$ improving classification than for the other three normalizers trialled.

The first overall conclusion is that F3 appears to be a better benchmark for normalization than F2, though only by a slight amount (between 0.5 and 0.7 percentage points). This difference might be related to the fact tha F3 varies less in speech than F2, which means that an accurate estimate of a sex's or individual's mean F3 can be obtained using a smaller sample of tokens.

The second conclusion is that tailoring the formant normalization to the individual speaker does improve the classification accuracy over normalizing speakers of the same sex together, but the improvement is marginal: 0.5 percentage points for μF2 and 0.7 percentage points for μF3. One way of thinking about this result is to remember that normalizing by individual speaker requires 20 normalizations in the present dataset whereas normalizing by sex requires only two. The two individual normalizations (namely $\mu F2_{individual}$ and $\mu F3_{individual}$) improve the classification accuracy by an average of 2.4 percentage points over 34 conditions, while the two sex normalizations ($\mu F2_{sex}$ and $\mu F3_{sex}$) improve it by an average 1.8 percentage points. Thus approximately three quarters of the improvement in classification accuracy is coming from the tailoring to speaker sex, with the remaining quarter coming from the tailoring to the individual speaker. It seems, then, that the bulk of the improvement in classification accuracy caused by formant normalization can be generated by using something as simple as subtracting the mean formant value of the speaker's sex.

The value of *c* for which the performance of $F2_R$ is strongest was 1.2 for $\mu F2_{sex}$, 1.6 for $\mu F3_{sex}$, 1.4 for $\mu F2_{individual}$, and 1.0 for $\mu F3_{individual}$. On the unnormalized data the value was $c = 1.8$, which is marginally higher.

For all four normalization techniques, the peak performance of $F2_R$ comes close to that of $F2_{onset} + F2_{mid}$, as was found in the unnormalized results. For $\mu F3_{individual}$ the discrepancy is 0.8 percentage points, which is somewhat larger than for $\mu F3_{sex}$ (0.2 percentage points), $\mu F2_{individual}$ (0.3 percentage points), and $\mu F2_{sex}$ (0.6 percentage points). Overall the discrepancy in classification accuracy is relatively small, which indicates that $F2_{onset}$ and $F2_{mid}$ can be collapsed into a single attribute with little effect on classification accuracy.

5.4.2.3 Results on $F3_R$

We now normalize $F3_R$. The normalization chosen is $\mu F3_{individual}$ since this has been established as being the strongest normalization method.

Here are the results:

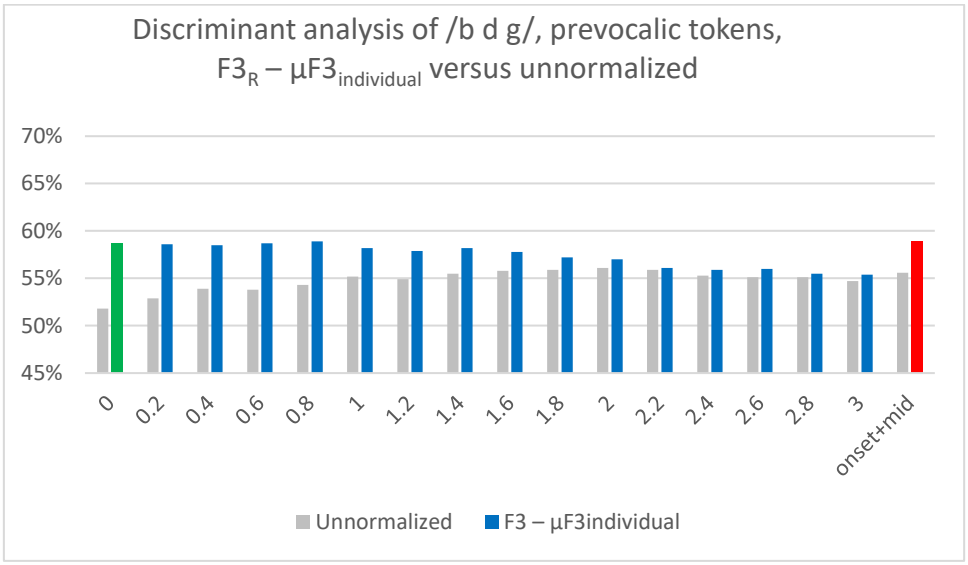


Figure 5.14: Normalization of $F3_R$ using $\mu F3_{individual}$.
The 17 conditions are the same as those described in Figure 5.9. N = 1,535.

After normalization, the classification accuracy of $F3_{onset}$ jumps from 51.8% to 58.7%, an improvement of nearly 7 percentage points. This result is highly statistically significant ($p < 0.001$). When we examine the performance of normalized $F3_R$ over the other 16 conditions, we see that it does not improve in the manner that we saw for $F2_R$: $F2_R$ improved more and more as the value of *c* was incremented from 0 to 1, whereas $F3_R$'s classification accuracy stays approximately the same: the classification accuracy of $F3_{onset}$ and $F3_{mid}$ combined (58.9%) is effectively the same as $F3_{onset}$ on its own (58.7%). This suggests that F3 transitions differ from F2 transitions in that the formant frequency in the middle part of the vowel does not boost the classification accuracy.

This may at first seem surprising but it is consistent with some previous findings. Recall from Chapter 2 that Lindblom (1990, 1996) replotted Öhman's (1966) $F2_{onset}$, $F2_{mid}$, and $F3_{onset}$ values and showed they were sufficient for discriminating /b d g/, i.e. he did not need to utilize $F3_{mid}$. Why might this be? Sussman et al. (1998: 251) plotted $F3_{onset}$ against $F3_{mid}$ for /b d g/:

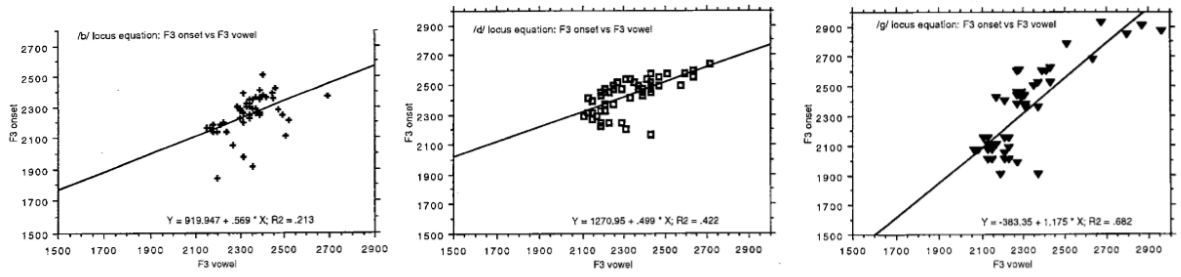


Figure 5.15: Locus equations of $F3_{onset}$ as a function of $F3_{mid}$ for /b d g/, from Sussman et al. (1998: 251). The authors note that there is greater variability (= deviation from the regression line) for F3 locus equations than F2 locus equations.

The authors found that F3 locus equations showed greater deviation from linearity than F2 locus equations, which is apparent from the above diagram in the considerable number of datapoints that lie far away from the regression line, especially for /b/. This is unlike the F2 locus equations in Section 2.2.3, where there is often striking linearity, especially for /b d/.

Stevens (1998: 342) measured the F2 and F3 transitions of [ɑbɑ ɑdɑ ɑgɑ] and notes that the F2 transition moves rapidly relative to the F3 transition, and that the F3 transition is smaller in extent. The results of his acoustic simulation of [bi bə bɑ] (p. 341) show the same slowness of F3 transitions relative to F2 transitions. This slowness of the F3 transition means, of course, that $F3_{onset}$ and $F3_{mid}$ would be expected to be less different from each other than $F2_{onset}$ and $F2_{mid}$, which would mean that $F3_{mid}$ would add less information to $F3_{onset}$ than $F2_{onset}$ does to $F2_{mid}$.

These considerations, taken together with the present study's results, all point to F3 playing a smaller role than F2 for distinguishing the place of articulation of /b d g/. More specifically, $F3_{mid}$ appears to contribute nothing beyond what $F3_{onset}$ contributes, at least when frequencies are normalized (by individual speaker). Correspondingly, $F3_R$ is no better than using $F3_{onset}$ on its own. This is inevitable since the ingredients of $F3_R$ are $F3_{onset}$ and $F3_{mid}$. But given that $F3_{mid}$ turns out to be irrelevant for identifying place of articulation, $F3_R$ in turn is not an improvement over $F3_{onset}$ alone.

Interestingly it was the normalized version of $F3_R$ that revealed this unimportance of $F3_{mid}$ most clearly. The unnormalized $F3_R$, as we see in the grey bars of Figure 5.14, had the illusion of showing a 4% improvement over using $F3_{onset}$ on its own, which disappeared once $F3_R$ was normalized.

5.4.3 Separating Vowels by Backness

The classification will now be split into two classifications: one for front vowels, the other for back vowels, with their respective accuracies added to yield a single figure. Doing so will

answer the question of to what extent this separation of vowel contexts helps the classification accuracy of F2 and F3. The theoretical motivation for doing so is the observation from Figure 2.13 of Section 2.2.1 that the F2 transitions for velars appear to point to different frequencies before front vowels and back vowels (around 2,500 to 3,000 Hz for the former, 1,500 to 2,000 Hz for the latter). Given this lack of a single locus frequency for velars, we expect the separation by vowel backness to aid the classification accuracy, as has been found by previous studies such as Suchato (2004). The main question is by how much.

Back vowels are defined as [ɑ ɔ o u] or any lowered and/or centralized variants thereof, e.g. [ü], [o̟]. Front vowels are any vowels that do not belong to this category, be they central or front. This particular division of vowels into front and back was prompted by the observation (whilst annotating the dataset) that the F2 transitions of velars seem to point to a relatively high frequency whether they occur before front vowels or central vowels (e.g. the /ɜː/ in British English *girlfriend*), whereas before back vowels they usually appear to point to a much lower frequency. Thus it seemed warranted to put central vowels in the same category as front vowels rather than with the back vowels. In contrast, Sussman et al. (1991: 1316) appear to have lumped central vowels with back vowels rather than front vowels. It is not known whether or to what extent this decision over how to class central vowels alters the results.

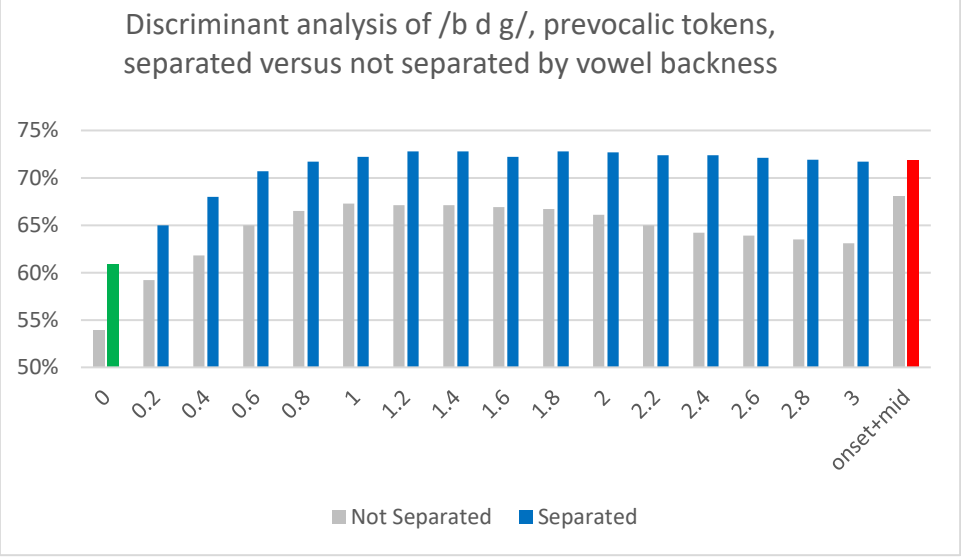


Figure 5.16: Classification accuracy of normalized $F2_R$ before and after separation by vowel backness. The results for front and back were run separately and summed. Both series normalized by $\mu F3_{individual}$. N = 1,535.

The separation of front-vowel and back-vowel contexts improves the classification accuracy considerably: the mean accuracy over the 17 conditions in Figure 5.16 is 64.4% prior to backness separation, 70.2% afterwards. The classification accuracy of the separated $F2_R$ values remains consistently high for a wide range of values of *c*: notice that the blue bars between *c* =

1.2 and $c = 2.8$ are within 1 percentage point of each other. The highest classification accuracy on the unseparated data was for $c = 1$ (67.3%) and for $c = 1.2$ and 1.4 (72.8%) in the data separated by vowel backness. Overall the results indicate that separating vowels by backness leads to a notable improvement in the ability of the F2 transition to distinguish place of articulation. Interestingly, this improvement in classification accuracy (averaging 5.8 percentage points over the 17 conditions examined) is over twice as large as the improvement in classification accuracy yielded by normalization (which ranged from 1.5 percentage points for $\mu F2_{sex}$ to 2.7 percentage points for $\mu F3_{individual}$).

In terms of statistical significance, when the 17 separated-versus-unseparated pairs in Figure 5.16 are compared the difference in classification is in every single case statistically significant at the $p < 0.001$ level. Contrast this with the best of the normalization procedures ($\mu F3_{individual}$), where the difference in classification accuracy between the normalized and non-normalized conditions was statistically significant only for low values of $F2_R$ (1.0 and lower). Thus the evidence that separating vowel contexts by backness improves classification accuracy is stronger than the evidence that normalizing frequencies improves classification accuracy.

The fact that separating by vowel backness leads to this substantial improvement in classification accuracy suggests that $F2_R$ does not eliminate vowel-dependent coarticulation in its entirety. In a way this should not be surprising, since in Section 2.2.1 it was noted that the velar F2 transition before front vowels points to a different frequency than that before back vowels, reflecting the fact that phonetically these are different places of articulation (palatal and velar).

We now examine the results for F3. We have established that when F3 is normalized, $F3_{mid}$ does not add anything to the classification beyond what $F3_{onset}$ contributes, and consequently the classification accuracy of $F3_R$ does not exceed that of $F3_{onset}$. Because of this the results presented henceforth for F3 are those for $F3_{onset}$:

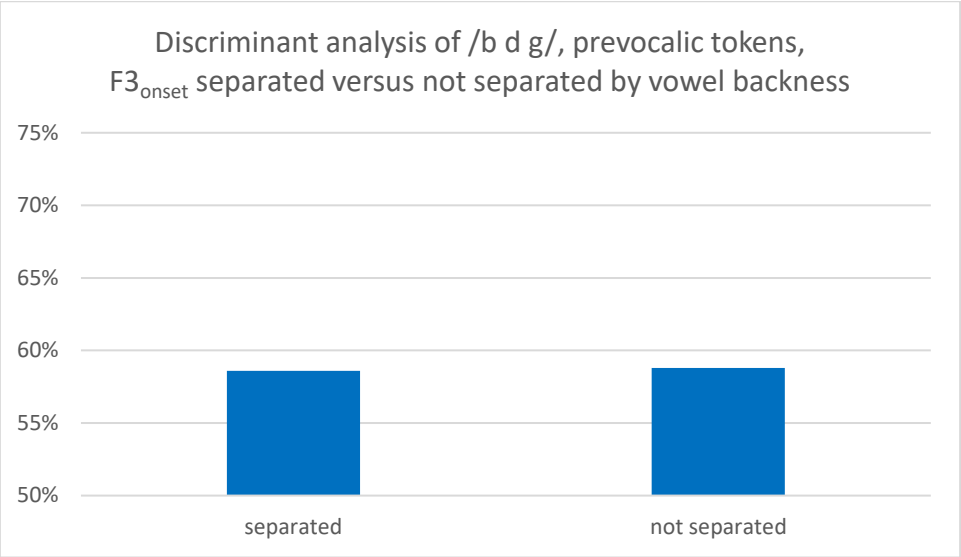


Figure 5.17: The classification accuracy of normalized $F3_{onset}$ on prevocalic /b d g/ when separated by backness versus not separated by backness. Both normalized by $\mu F3_{speaker}$.
N = 1,535.

As we can see, the separation by vowel backness does not yield an improvement in the classification accuracy of $F3_{onset}$; in fact it worsens it marginally (from 58.8% to 58.6%). This result is in marked contrast to the results for F2 above, and suggests that whatever the variation in $F3_{onset}$, it is not governed by vowel backness.

Let us add these $F3_{onset}$ results to the $F2_R$ results to quantify its improvement in classification accuracy:

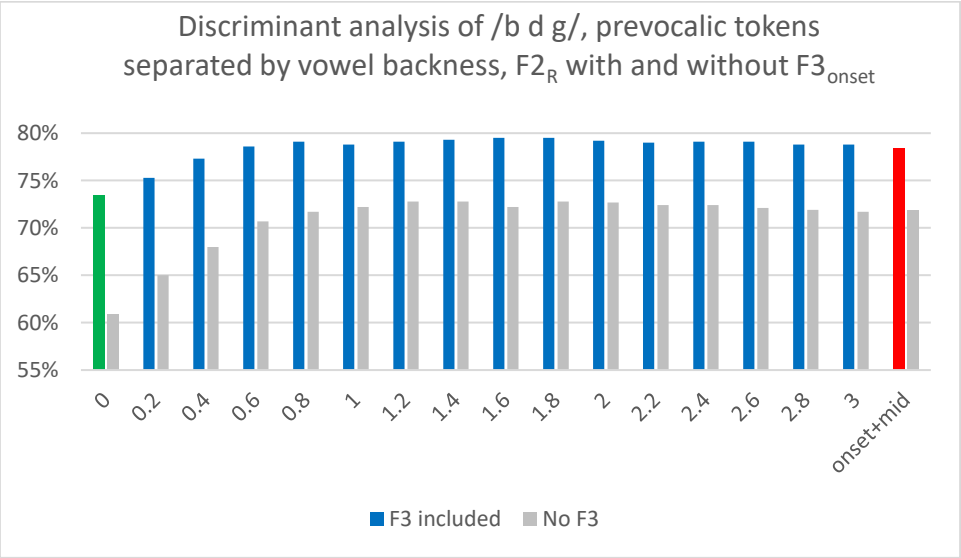


Figure 5.18: Comparison of the classification accuracy of normalized $F2_R$ separated by vowel backness with and without the addition of normalized $F3_{onset}$.
Both series normalized by $\mu F3_{individual}$. N = 1,535.

The improvement yielded by the inclusion of F3 is considerable across all conditions, being at its largest when combined with $F2_R0$ (i.e. $F2_{onset}$) in which the classification improves by 12.5 percentage points, and at its smallest when combined with $F2_{onset}$ and $F2_{mid}$ (6.5 percentage points). The improvement in classification accuracy yielded by the inclusion of $F3_{onset}$ is somewhat larger than the improvement seen above when the classification was separated by vowel backness, which yielded a mean improvement of $F2_R$ over the 17 conditions of 5.8 percentage points, as opposed to the present 7.4 percentage-point improvement.

With regard to statistical significance, the classification accuracy after the inclusion of $F3_{onset} - \mu F3_{individual}$ is significantly different at the $p < 0.001$ level for all 17 pairs shown in Figure 5.18. This is unsurprising given that the inclusion of $F3_R$ has boosted the classification more on average than splitting by vowel backness, which itself showed a $p < 0.001$ significance in all cases.

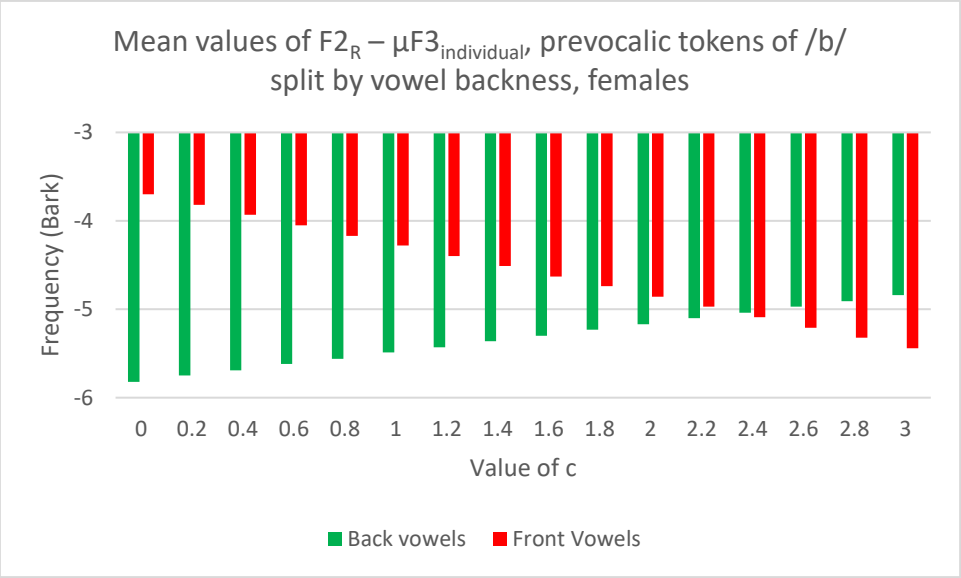
All values of $F2_R + F3_{onset}$ above in which *c* is equal to or greater than 0.6 classify more accurately than when $F2_{onset} + F2_{mid} + F3_{onset}$ is used. However, the improvement is never larger than 1.1 percentage points. Earlier it was shown that $F2_R$ under the four normalization conditions was less accurate than $F2_{onset} + F2_{mid}$, by 0.2 to 0.8 percentage points. Over all the conditions examined, $F2_R$ has exceeded the performance of $F2_{onset} + F2_{mid}$ on two occasions (on the unnormalized data and on the present data in which $F3_{onset}$ is included) whereas $F2_{onset} + F2_{mid}$ was slightly stronger in the remainder of the conditions. But because the difference in classification accuracy between the two has never exceeded 1.1 percentage points, it seems wise to regard the two as having approximately the same ability to distinguish place of articulation.

5.4.4 Mean Values of $F2_R$ as an Indicator of the Locus Frequency

We now investigate in greater detail what $F2_R$ is doing to the front-vowel and back-vowel contexts in an effort to find a deeper principle for setting the value of *c*. It was noted in Chapter 2 that the 1950s research indicated that the concept of a locus frequency worked best for /d/, less well for /b/, and least well for /g/. Let us now probe the results for $F2_R$ in further detail by observing how increasing the value of *c* affects the *mean* $F2_R$ values for front vowels vis-à-vis back vowels. The aim is to identify the value of *c* for which the mean value of back-vowel $F2_R$ converges with that of front-vowel $F2_R$. This is noted for (1) alveolars, and (2) bilabials. (The same shall also be investigated in velars, though as we noted in Chapter 2 velars are not expected to show similar loci before front and back vowels and hence will presumably not yield an intersection of their front-vowel and back-vowel loci, for any value of *c*).

This investigation will be performed for male and female data separately, in case the two sexes articulate the consonants or vowels differently enough to result in them having

different values of *c* at which front and back-vowel contexts converge (one reason why this might be expected is given on the following page, from Fant (1975)). We begin with /b/. Here are the results:



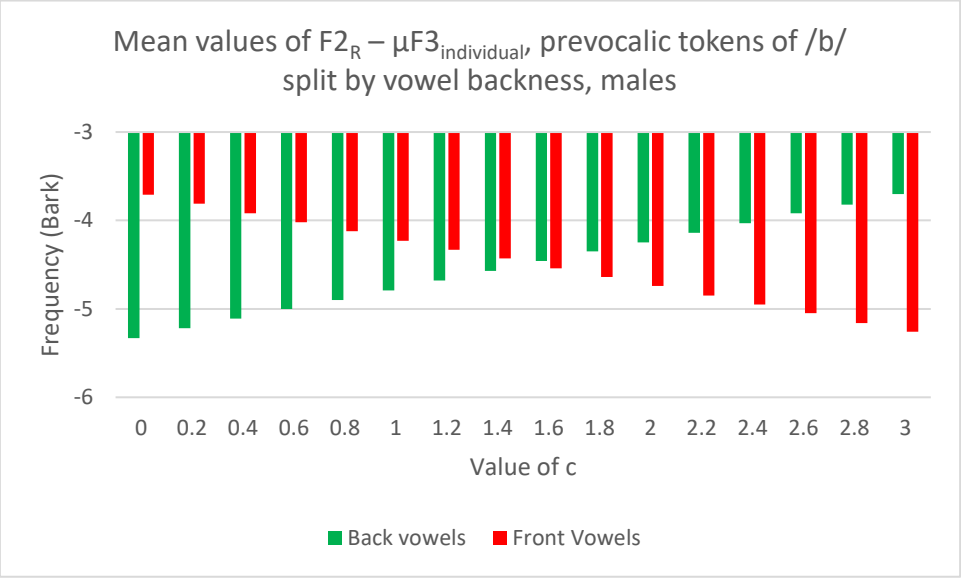


Figure 5.19: Mean frequency of $F2_R$ – $\mu F3_{individual}$ (in Bark) for prevocalic tokens of /b/ separated by vowel context, female and males speakers.
Front vowel context shown in red, back vowel contexts shown in green. Female N = 286, male N = 308.

The convergence of front-vowel and back-vowel $F2_R$ can be seen to occur at a different value of *c* for males and females: 1.6 for males, 2.4 for females. The two sexes have strikingly similar mean values for front vowels (-3.71 versus -3.70 Bark for $F2_{onset}$ – $\mu F3_{individual}$ and -3.19 and -3.13 Bark for $F2_{mid}$ – $\mu F3_{individual}$). Thus the cause of this difference of convergence is due to a difference between males' and females' back vowels: for mean $F2_{onset}$ – $\mu F3_{individual}$ values are -5.33 Bark for males, -5.82 Bark for females, and there is a similar though smaller discrepancy

in their respective $F2_{mid} - \mu F3_{individual}$ values, namely -5.87 and -6.14 Bark. At both vowel onset and midpoint, then, female speakers appear to be using a backer realization than males.

Fant (1975), who amalgamated male and female mean formant frequencies for six European languages, found that the mean F2 frequencies of male and female speakers tended to be closer together for back vowels than front vowels. If this were also true of the present data set, it would lead one to expect the mean $F2_R - \mu F3_{individual}$ to be larger for female speakers than male speakers, which is indeed observed. This could go some way to accounting for the male-female difference in the value of *c* at which mean $F2_R$ converges for front- and back-vowel contexts. However, if this really is the case, the male-female discrepancy should also be expected to appear in the results for /b g/, which, as will be shown, does not occur. Therefore it is probably wiser not to read too much into this gender difference, especially given the small sample sizes involved with these data subsets.

Here are the results for /d/. The results for /d/ are more important than for the other two places of articulation since it was established in Section 2.2.1 that the F2 transitions of alveolars are the place of articulation that proved the most amenable to the locus theory.

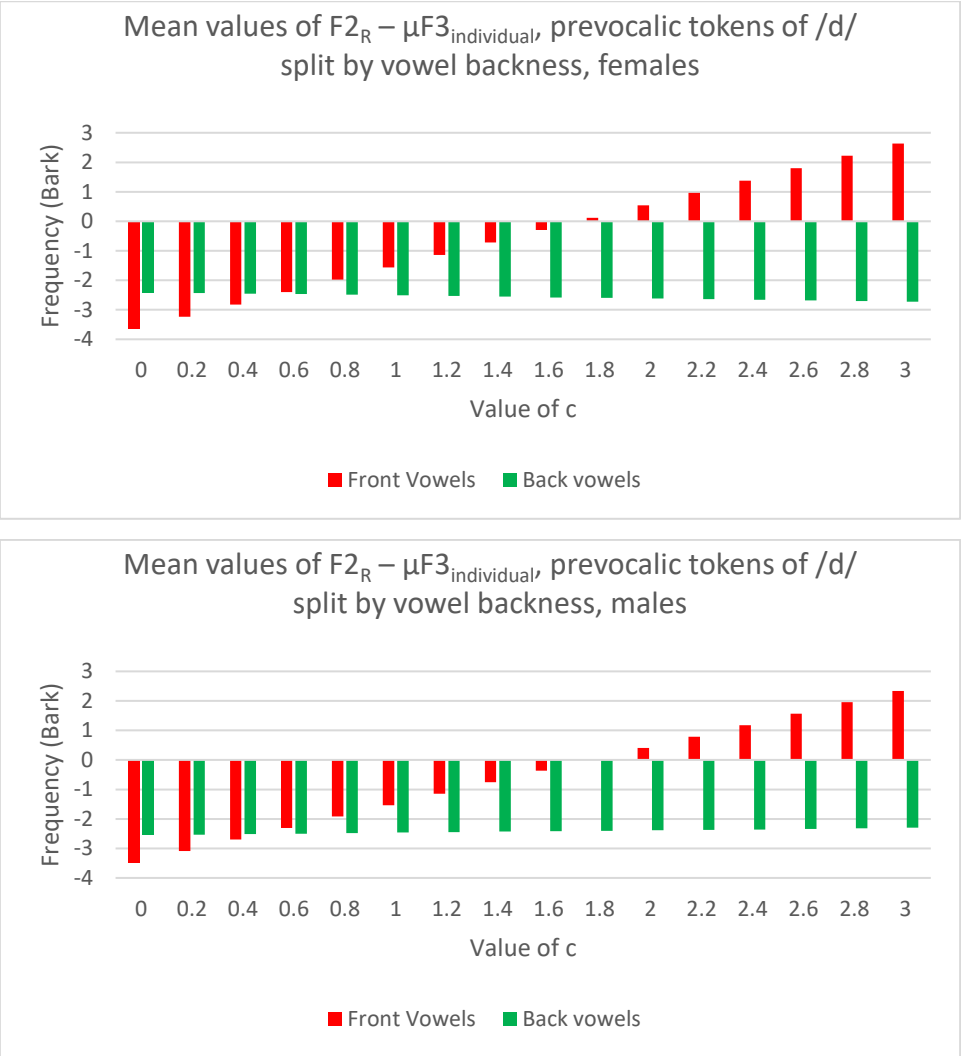


Figure 5.20: Mean frequency of $F2_R$ – $\mu F3_{individual}$ (in Bark) for prevocalic tokens of /d/ separated by vowel context, male and female speakers.
Front vowel context shown in red, back vowel contexts shown in green. Male N = 329, female N = 319.

The results for /d/ in males and females are strikingly similar to each other, unlike /b/. In particular, the value of *c* for which front-vowel $F2_R$ and back-vowel $F2_R$ converge (0.6) is the same in both sexes. Another striking feature of the two figures is that the mean front-vowel frequencies for values of *c* equal to or greater than 1.8 are *positive*. That is, the mean $F2_R$ – $\mu F3_{individual}$ value in front-vowel context for *c* values greater than 1.8 is higher in frequency than the speakers' mean F3! It need hardly be said that such values are absurd, since by definition F2 is lower in frequency than F3. Thus, a large value of *c* yields theoretically bizarre $F2_{locus}$ frequencies.

These positive $F2_R$ – $\mu F3_{individual}$ values would be greater than approximately 2,531 Hz for male speakers and greater than 2,906 Hz for female speakers (these are the mean male and female $F3_{mid}$ frequencies in the dataset). It was noted in Chapter 2 that Delattre et al. (1955) estimated the $F2_{locus}$ for /d/ to be around 1,800 Hz (for male speakers). Thus the idea that *c* has

a value of 1.8 or greater is highly implausible. To a lesser extent, the idea that *c* would have a value greater than 0.6 is also dubious, since such values of *c* result in mean alveolar $F2_R$ being greater in value for back vowels than front vowels. This could be thought of as a kind of 'overcompensation for coarticulation' since the original aim of the locus theory was to *unite* the alveolar frequencies across different vowel contexts, as illustrated in Figure 5.3.

On the other hand, there is the pragmatic consideration of what value of *c* yields the greatest classification accuracy. As we saw in the examination of different methods of normalization, this value is between 1 and 1.6. Nevertheless the results when $F3_{onset}$ was added to the classification was that any variant of $F2_R$ in which *c* was equal to or greater than 0.6 yielded a classification accuracy (78.6%) that was higher than that of $F3_{onset}$, $F2_{onset}$, and $F2_{mid}$ combined (78.4%). Thus, setting the value of *c* to 0.6 results in a classification accuracy that surpasses that of $F2_{onset}$ and $F2_{mid}$ combined as though they were separate features, while avoiding $F2_{locus}$ values that are out of touch with what has been indicated in the literature.

In Section 8.7 the excessively high value of back-vowel $F2_R$ in /d/ will be revisited and it will be shown that the problem is related to the use of $F2_{mid}$ in the $F2_R$ formula. Replacing $F2_{mid}$ with an $F2_{mean}$ that is calculated on a syllable-length chunk of F2 frequencies surrounding $F2_{onset}$ could lead to an $F2_R$ formula that generates $F2_R$ frequencies more in touch with the 1950s $F2_{locus}$ theory.

Here are the results for /g/:

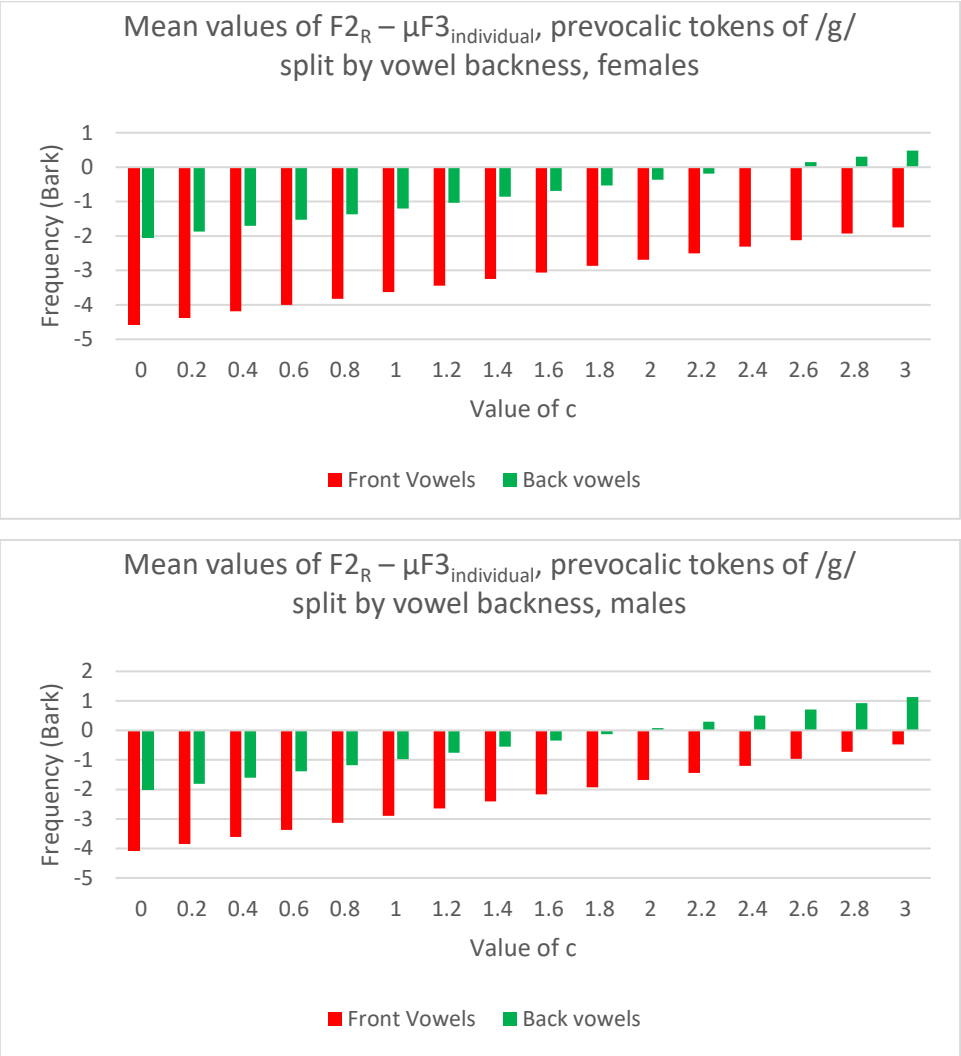


Figure 5.21: Mean frequency of $F2_R - \mu F3_{individual}$ (in Bark) for prevocalic tokens of /g/ separated by vowel context, female and male speakers.

Front vowel context shown in red, back vowel contexts shown in green. Female N = 148, male N = 145.

The $F2_R$ values for back vowels and front vowels do not converge, at least for the values of *c* examined. However, the gap between the two $F2_R$ values shrinks as *c* is increased to 3. Theoretically the two $F2_R$ values would converge if the values for *c* were set enormously high, to perhaps 15. However, we have already seen that values of *c* equal to or greater than 1.8 yield strange results for alveolars, let alone values greater than 3. Thus setting *c* to such values would not yield a realistic locus for the other places of articulation.

But in any event, as noted in Section 2.2.1 there is a solid theoretical reason for regarding /g/ as having more than one locus: from a phonetic point of view it contains more than one place of articulation.

5.4.5 VC Formant Transitions

The results up until now have pertained to CV transitions, that is, the formant transitions that occur between the plosive and a *following* vowel. This section pertains to the formant transitions that *precede* a plosive, VC transitions. $F2_R$ is computed in the same manner as for the CV transitions except that $F2_{offset}$ is used in place of $F2_{onset}$.

The total number of tokens (N = 972) is smaller than for CV (N = 1,535). Recall two facts noted in Chapter 2 about VC transitions: (1) they have been found to have steeper locus-equation slopes than CV transitions (Section 2.2.3); (2) listeners' accuracy at identifying place of articulation when given only the VC transition is haphazard (Section 2.3.2). These two considerations led to a prediction that VC transitions should less reliably encode place than CV transitions. Here are the results of a comparison of the classification accuracy of the two transitions:

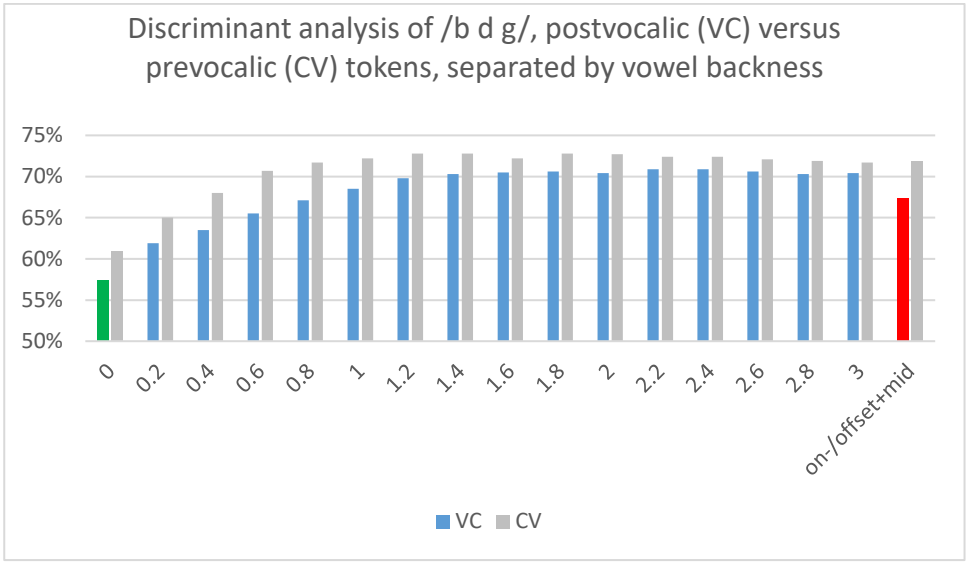


Figure 5.22: Comparison of vowel-consonant (VC) and consonant-vowel (CV) $F2_R$ in terms of its ability to distinguish the place of articulation of /b d g/ for values of *c* between 0 and 3.
Separate classifications for front vowels and back vowels. Normalized by $\mu F3_{individual}$. N = 972 for VC, N = 1,535 for CV.

The classification accuracy of $F2_R$ in VC transitions is weaker than that in CV transitions across all conditions, and this is also true when $F2_{offset}$ and $F2_{mid}$ are used as classifiers (67.4% versus 71.9%). The mean classification accuracy of the VC transitions over the 17 conditions in Figure 5.22 is 68.0% as against 70.8% for the CV transitions. This superiority of CV transitions over VC transitions for classifying place of articulation, though modest, is in accord with previous findings (e.g. Modarresi et al., 2004; Suchato, 2004). Another major difference with VC context is that the value of *c* needed to optimize the performance of $F2_R$ is far higher than for the CV context (the best classification accuracy for CV, 72.8%, is for $F2_R$1.2 and 1.4, whereas for VC

the best classification accuracy, 70.9%, is for $F2_R$2.2 and 2.4). This suggests that the $F2_{offset}$ frequencies for each of the places of articulation in VC contexts tend to be more similar to each other than the $F2_{onset}$ frequencies of CV contexts, and hence need a greater value of *c* to be separated sufficiently. This is unsurprising since, as noted in 2.3.2, F2 transitions in VC transitions have been found to have steeper locus-equation slopes than CV transitions (Al-Tamimi, 2004).

To illustrate this hypothesis, here are the mean values for $F2_{offset} - \mu F3_{individual}$ separated by backness:

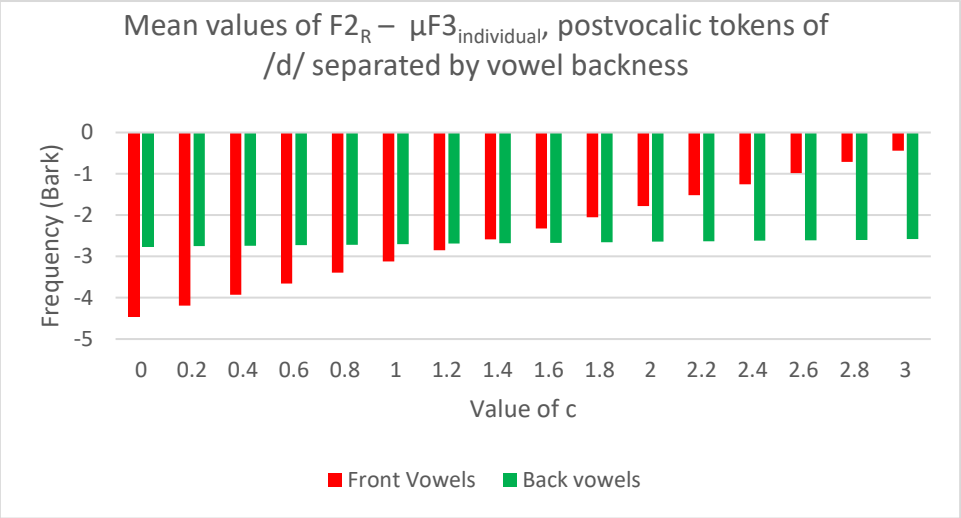


Figure 5.23: Mean values of VC $F2_R - \mu F3_{individual}$ for /d/ with values of *c* between 0 and 3, separated by vowel backness.
N = 972.

The value of *c* at which the mean $F2_R$ value for back-vowel /d/ context converges with that for front-vowel /d/ context is 1.4, which is far larger than the value shown earlier for /d/'s CV transitions, namely 0.6. This suggests that the VC F2 transition (for alveolar place at least) involves less frequency change from vowel midpoint to vowel offset than the frequency change from vowel onset to vowel midpoint in CV context. Or, to put it in the language of locus equations, the slope of VC transitions is steeper than for CV transitions. (The articulatory reason for this is presumably that the tongue is less deactivated at the end of a vowel than it is at the onset of a vowel (Houde, 1967).)

Still another way of seeing this is to compare the mean $F2_{offset} - F2_{mid}$ values for back-vowel VC context with the mean $F2_{onset} - F2_{mid}$ values for back-vowel CV context:

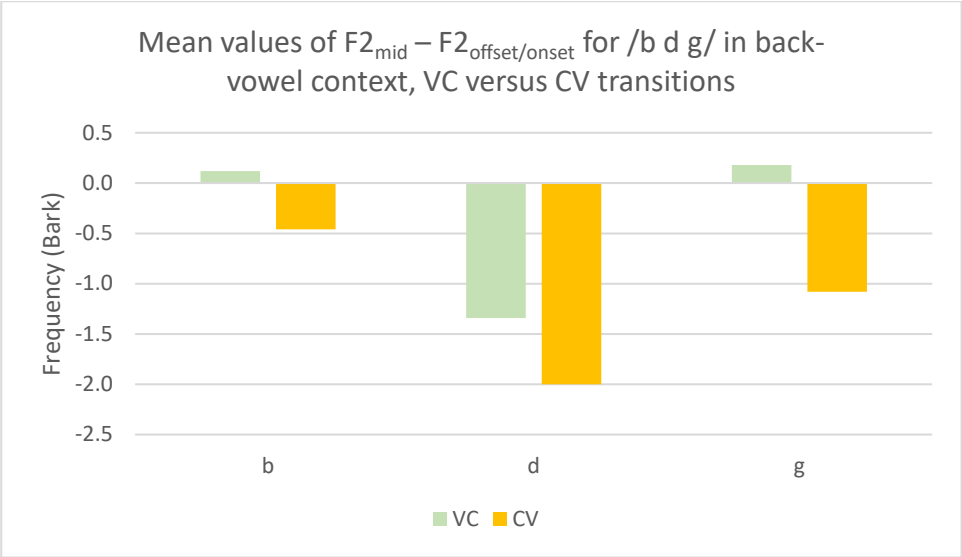


Figure 5.24: Comparison of VC and CV transitions in terms of the mean difference in frequency (in Bark) between $F2_{mid}$ and $F2_{onset/offset}$ for each of /b d g/ in back-vowel context.
VC N = 268, CV N = 429.

For all three of the places of articulation, the mean difference in frequency between $F2_{onset/offset}$ and $F2_{mid}$ is less for VC transitions than CV transitions. In other words the formant transitions in VC context involve less frequency change than those in CV context. Another notable feature is that the mean formant transitions for velars and bilabials in back-vowel VC context are almost identical in extent and direction, whereas in CV context they are not. This is yet further evidence suggesting that the contrastivity of VC transitions for signalling place of articulation is less than that of CV transitions.

Turning now to the results for front-vowel context:

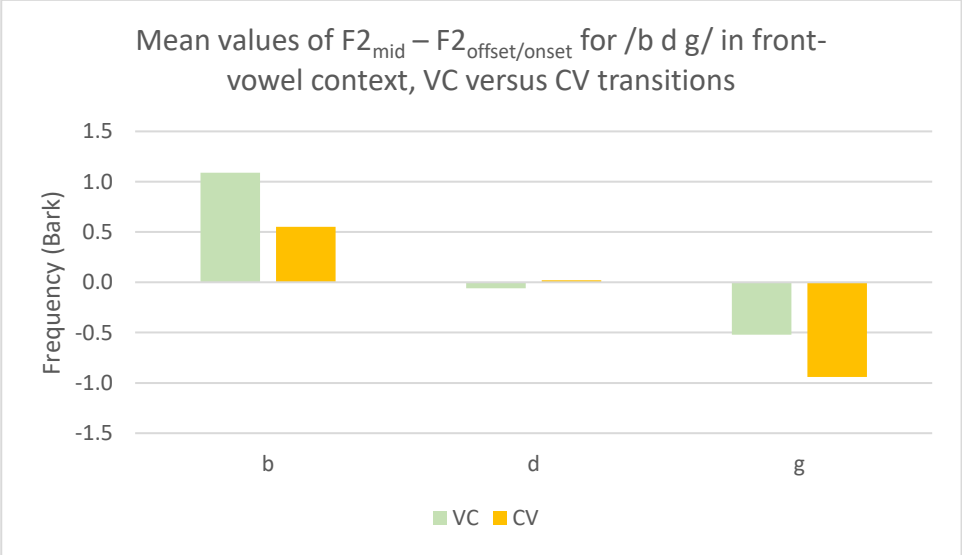


Figure 5.25: Comparison of VC and CV transitions in terms of the mean difference in frequency (in Bark) between $F2_{mid}$ and $F2_{onset/offset}$ for each of /b d g/ in front-vowel context.
VC N = 704, CV N = 1,106.

For bilabials we again see that the formant transitions are smaller in extent in VC than in CV context. However, for alveolars we see no such difference, but this is incidental since the formant transitions happen to be flat (on average) in this context. For velars we see a reverse of the pattern, i.e. that VC transitions are larger in extent than CV transitions.

In conclusion, the formant transitions in VC context are smaller in extent than those for CV transitions in four out of five of the relevant backness-place combinations we have seen. This is likely to be a factor in why VC transitions classify place somewhat less accurately than CV transitions.

Here are the VC results for of $F3_R$:

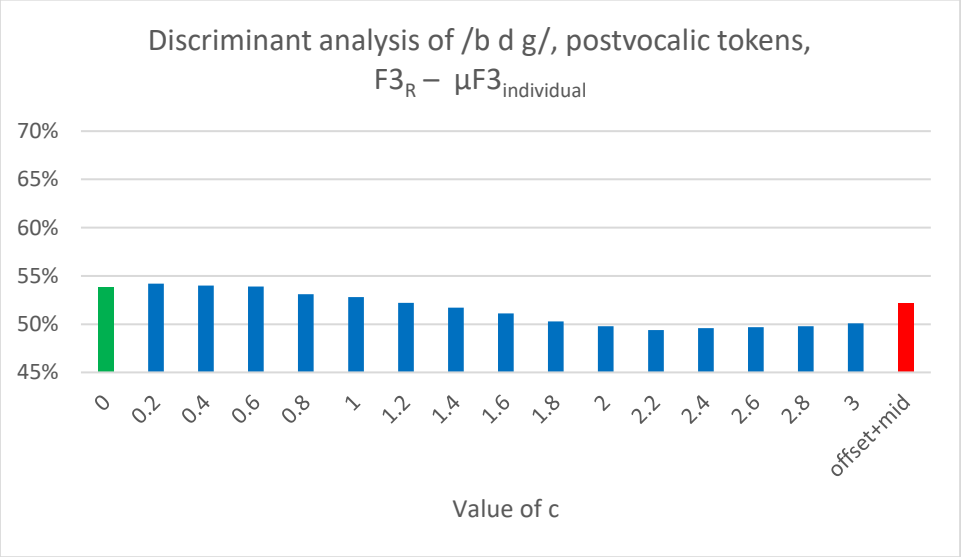


Figure 5.26: The performance of $F3_R$ on VC transitions, for values of *c* between 0 and 3. N = 974.

As with the CV tokens examined earlier, we see that $F3_R$ does *not* classify more accurately than $F3_{offset}$ ($F3_R0.2$, the best of the $F3_R$ candidates, classifies just 0.4 percentage points more accurately than $F3_{offset}$, too small an improvement to be meaningful). In the present case the result seems to be due to the fact that $F3_{mid}$ hurts the classification accuracy, as can be seen in Figure 5.26 from the red bar representing $F3_{onset}$ + $F3_{mid}$ (52.2% accurate, which is 1.6 percentage points less than $F2_{onset}$ on its own). This influence of $F3_{mid}$ is of course to be seen in all of the blue bars representing $F3_R$, since $F3_{mid}$ is one of the ingredients in $F3_R$.

The most important point about the above result is that it suggests that $F3_{offset}$ should be chosen as an attribute rather than $F3_R$ for VC transitions. This picture is the mirror image of the one we arrived at earlier in the chapter (5.4.2.3) regarding CV transitions. The overall conclusion is that including $F3_{mid}$, whether it is from V(C) or (C)V, does not add to the

classification accuracy of place of articulation. Rather, only $F3_{offset}$ (in VC) and $F3_{onset}$ (in CV) help.

A second question which Figure 5.26 answers is the degree to which F3 in VC context is better or worse at classifying place of articulation than F3 in CV context. The classification of $F3_{offset}$ above is 53.8% which, compared to $F3_{onset}$ in CV transitions, is 5.0 percentage points lower. This inferiority of the VC classification to the CV classification is somewhat smaller in magnitude to what we saw for F2 (in CV context, the peak classification accuracy of $F2_R$ was 72.8%, whereas in VC context the peak classification accuracy of $F2_R$ was 63.0%, a 9.8 percentage-point difference). Nevertheless the trend for both F2 and F3 is the same: VC context appears to contain less reliable information for distinguishing place of articulation than CV context.

One final question with regard to VC context is this: how much does the classification accuracy of $F2_R$ improve when $F3_{offset}$ is added to it?

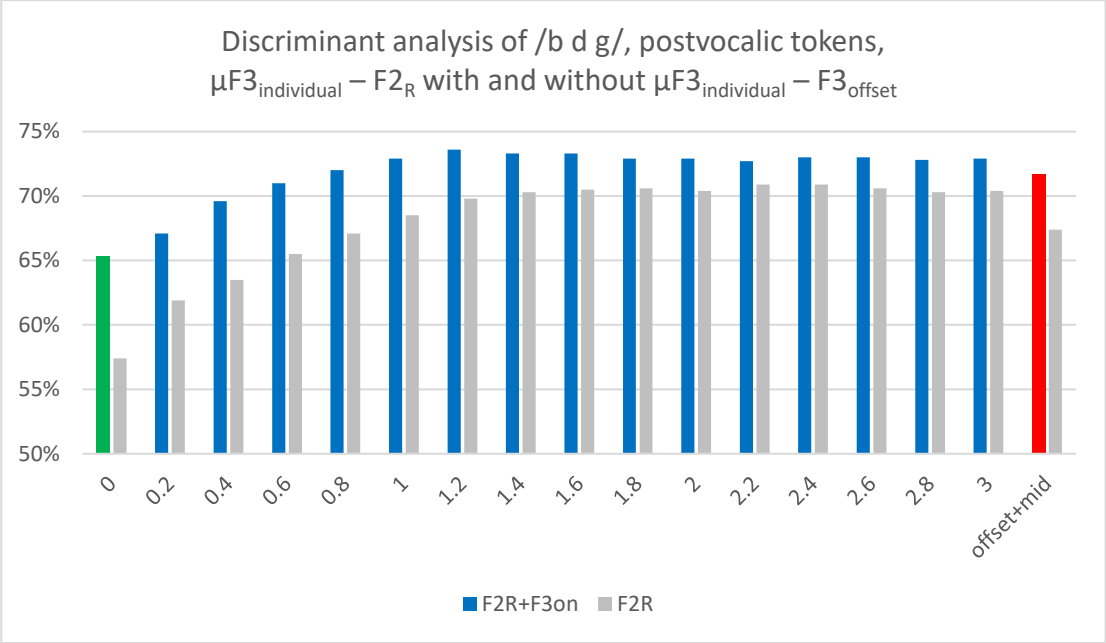


Figure 5.27: The performance of $F2_R$ in VC transitions with and without the inclusion of $F3_{offset}$. Classification run separately on front and back vowels. N = 974.

As with $F3_{onset}$ in CV context, the inclusion of $F3_{offset}$ with $F2_R$ results in a consistent boost in the classification accuracy of $F2_R$ across all values of *c*. The classification accuracy peaks at 73.8% when $c = 1.4$. The classification accuracy of the VC information is again less than that of the CV information under the same conditions: recall that the classification of $F2_R + F3_{onset}$ in CV context was over 79% at its highest, which is over 5 percentage points higher than the peak performance of $F2_R + F3_{offset}$ in VC context.

In terms of statistical significance, recall that in the CV transitions the inclusion of $F3_{onset}$ in the classification resulted in a highly statistically significant change in the classification ($p < 0.001$ in all 17 conditions tested). For the VC transitions, this $p < 0.001$ level of significance is only attained for the five leftmost pairs in Figure 5.27 above, i.e. when c = 0.8 or less. For c = 1.0 or 1.2, $p < 0.01$, while for c = 1.4 $p < 0.05$. For the onset + mid condition (the red bar above), the improvement is $p < 0.01$.

To summarize, the inclusion of F3 in the classification of VC context improves the classification accurarcy, though not to the same degree as CV context and not to the same degree of statistical significances in all cases. The more unreliable nature of VC formant information for identifying place of articulation is in line with previous studies using other methodologies such as listener responses (e.g. Malécot (1958) reviewed in 2.3.2).

5.4.6 Formant Distances

We now examine the accuracy at discriminating place of articulation by formant distances. The two distances examined are $F2 - F1_R$ and $F3 - F2_R$. As noted in Chapter 2, $F3_{onset} - F2_{onset}$ was investigated by Suchato (2004: 95) and was found to be approximately as strong an attribute as $F2_{onset}$ (and stronger than $F3_{onset}$) based on its Maximum-Likelihood classification error (which was 0.51 for both $F2_{onset}$ and $F3_{onset} - F2_{onset}$). The theoretical motivation behind the attribute was that the distance between F2 and F3 tends to be smaller in velars than in the other two places of articulation. However, as pointed out in Section 2.2.4, this proximity of the two formants as an indicator of velar place is only a tendency and there are two main exceptions to it: (1) before back vowels (e.g. /g/ in *golf*), the distance between F3 and F2 in velars tends to be large; (2) before front vowels, especially high front vowels, the distance between F3 and F2 can be small in bilabials (see Öhman (1966: 160) for an illustration). Taking Suchato's finding together with these theoretical considerations, it is something of an open question to what extent $F3_{onset} - F2_{onset}$ (and, by extension, $F3 - F2_R$) will be a performant attribute.

As regards $F2_{onset} - F1_{onset}$, this has been found by Al-Tamimi (2017) to be a strong predictor in distinguishing pharyngealized and non-pharyngealized /dˤ d/ in Arabic (pharyngealized /dˤ/ has higher $F1_{onset}$ and lower $F2_{onset}$ than /d/, hence a smaller $F2_{onset} - F1_{onset}$). In /b d g/, however, $F2_{onset} - F1_{onset}$ does not appear to have been examined in previous studies; hence it is necessary to first justify its use. Suchato (2004: 71) found that bilabials had a higher mean $F1_{onset}$ than the other two places of articulation (484 Hz as against 439 Hz for alveolars and 414 Hz for velars). The fact that $F1_{onset}$ was higher in bilabials is presumably due to the the consonant's constriction not requiring the tongue, which means that the tongue is free to start moving towards the following vowel target sooner than is the case with alveolars and

velars. Suchato mixed together the results for /p t k/ and /b d g/ so perhaps the magnitude of the difference between bilabials and non-bilabials would be larger with /b d g/ only, since $F1_{onset}$ occurs several tens of milliseconds after the release of the plosive in the voiceless series and so the F1 may have already risen to its position for the following vowel target.

Here are the mean $F1_{onset}$ and $F2_{onset}$ for /b d g/ in the present dataset:

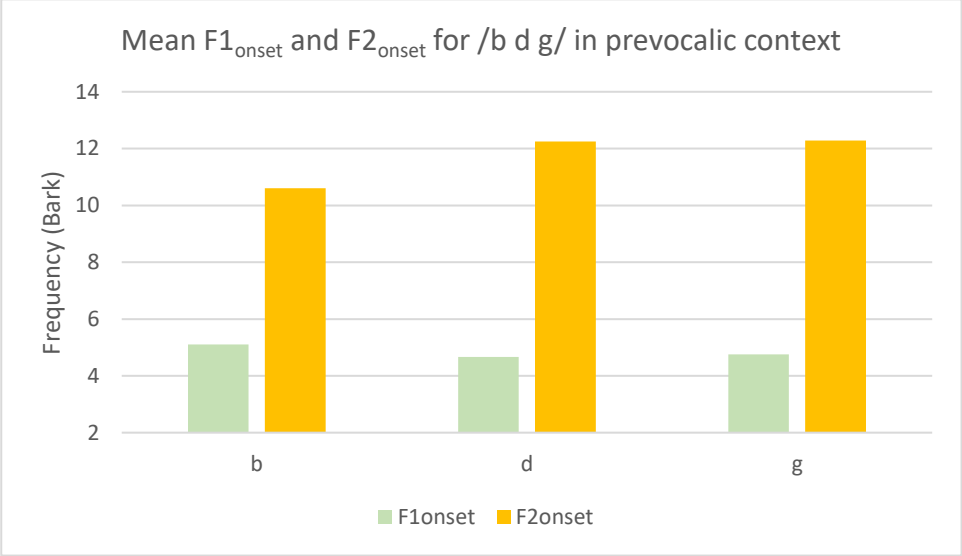


Figure 5.28: Mean $F1_{onset}$ and $F2_{onset}$ values for prevocalic /b d g/.
N = 1,535.

We see that /b/ has the highest mean $F1_{onset}$ (5.1 Bark as against 4.67 Bark for /d/ and 4.76 Bark for /g/) but the lowest mean $F2_{onset}$ (10.6 Bark as against 12.3 Bark for /d g/). This leads one to expect that its $F2_{onset} - F1_{onset}$ values will tend to be smaller than those of velars and alveolars. To what extent this works as an attribute in practice is, of course, is an empirical question, to be addressed shortly.

Before presenting the results, however, it is necessary to be clear as to how formant distances are implemented using the F_R concept. Let us illustrate with the example of $F3 - F2_R$ (analogous comments apply to $F2 - F1_R$).

Step 1 is to subtract $F2_{onset}$ from $F3_{onset}$:

(1) $Onset_{difference} = F3_{onset} - F2_{onset}$

Step 2 is to subtract $F3_{mid}$ from $F2_{mid}$:

(2) $Midpoint_{difference} = F3_{mid} - F2_{mid}$

Step 3 is to generate the F_R series for the same values of *c* as before (namely between 0 and 3 in increments of 0.2):

$$(3) \qquad F3 - F2_R = Onset_{difference} - ((Midpoint_{difference} - Onset_{difference}) \times c)$$

Another feature of F_R as applied to formant distances is that there is no normalization. This is because the normalization that was used in previous sections itself involves formant distances, i.e. the mean $F3_{mid}$ value of a given speaker or sex was subtracted from $F2_R$ or $F3_{onset/offset}$. However, F2 – F1 and F3 – F2 are *themselves* formant distances, which means that if the F2 and F1 or F3 and F2 values were normalized prior to being subtracted from each other, the same formant distance would occur as occurs without normalization.

One final matter to understand before examining the results is what the effect of F_R on formant distances is expected to be. Let us take the example of $F3 - F2_R$. As has been noted, velars consonants (especially before front vowels) involve 'velar pinch'. In acoustic terms, velar pinch means that the difference in frequency between $F3_{onset}$ and $F2_{onset}$ is smaller than the difference between $F3_{mid}$ and $F2_{mid}$. When $F3 - F2_R$ is applied to this set of four points as shown in (3) above, its effect will be to make the velar pinch even smaller than it actually is. In contrast, if the difference between $F3_{mid}$ and $F2_{mid}$ is about the same as that of $F3_{onset}$ and $F2_{onset}$ (as tends to occur with /b/ before back vowels and /d/ before front vowels, as shown in Section 5.4.5), then $F3 - F2_R$ will yield the same result as $F3_{onset} - F2_{onset}$. The net result, then, is that the difference between velars and non-velars is expected to be enhanced by $F3 - F2_R$ relative to $F3_{onset} - F2_{onset}$. We turn now to the results, beginning with $F2 - F1_R$:

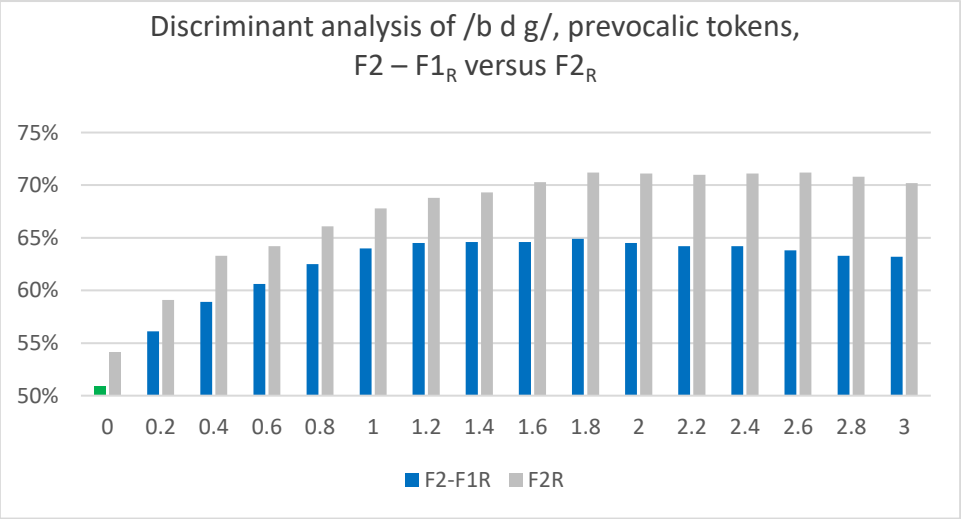


Figure 5.29: Classification accuracy of $F2 - F1_R$ relative to $F2_R$ over the same range of values of *c* that have been utilized throughout this chapter.

Classification run separately on front-vowel and back-vowel contexts and summed. No normalization. N = 1,535.

F2 – $F1_R$ classifies place of articulation notably worse than $F2_R$. This is true across all values of *c* examined. Over the 16 conditions in 5.29 above, F2 – $F1_R$ has a mean classification accuracy of 62.2% as against 67.5% for $F2_R$. The peak classification accuracy of F2 – $F1_R$ (for $c = 1.8$) is 64.9%, which is 6.3 percentage points less than that of $F2_R$ (for $c = 1.8$ and 2.6).

As for statistical significance, a comparison of the 16 pairs in Figure 5.29 above revealed that the worsening of the classification by subtracting F1 was statistically significant at the $p < 0.001$ level for all cases where $c = 1.2$ or greater. For $c = 0.8$ or 1.0, $p < 0.01$ and for $c = 0.2$ or 0.4 $p < 0.05$. For $c = 0$, the difference in the classification between $F2_R$ and F2 – $F1_R$ is not statistically significant. Overall, then, the worsening of the classification by the subtraction of F1 is a true effect in nearly all cases.

We turn now to the second formant-distance attribute, F3 – $F2_R$. Here are the mean $F2_{onset}$ and $F3_{onset}$ values for each place:

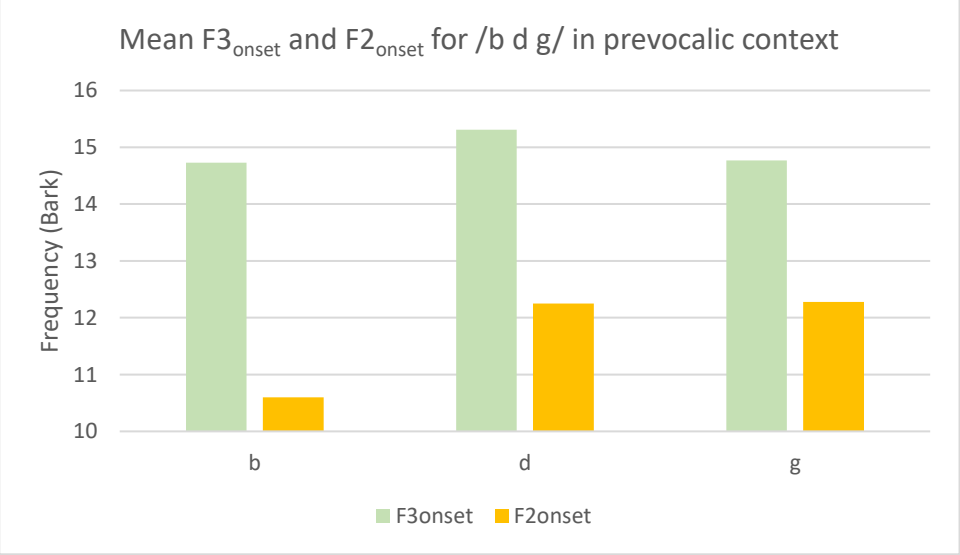


Figure 5.30: Mean $F2_{onset}$ and $F3_{onset}$ values for prevocalic /b d g/.
N = 1,535.

The difference in frequency between $F3_{onset}$ and $F2_{onset}$ is smaller for velars (2.49 Bark) than the other two places of articulation (3.06 Bark for /d/, 4.13 Bark for /b/). However, this data combines both front-vowel and back-vowel contexts, even though velar pinch is known to be more typical of front-vowel context. So here are the results for this latter context:

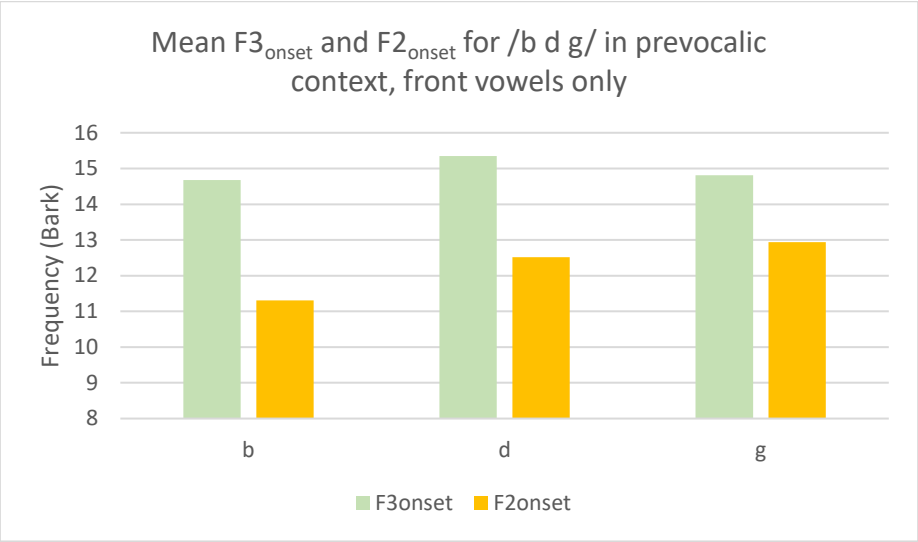


Figure 5.31: Mean $F2_{onset}$ and $F3_{onset}$ frequencies for pre-front-vowel /b d g/.
'Front vowel' is defined in the present study to include central vowels. N = 1,106.

The difference between mean $F3_{onset}$ and mean $F2_{onset}$ is now substantially smaller in velars (1.89 Bark) than for the other two places (2.83 Bark for /d/, 3.37 Bark for /b/). Of course, the real question is whether $F3 - F2_R$ is a worthwhile attribute, to which we now turn:

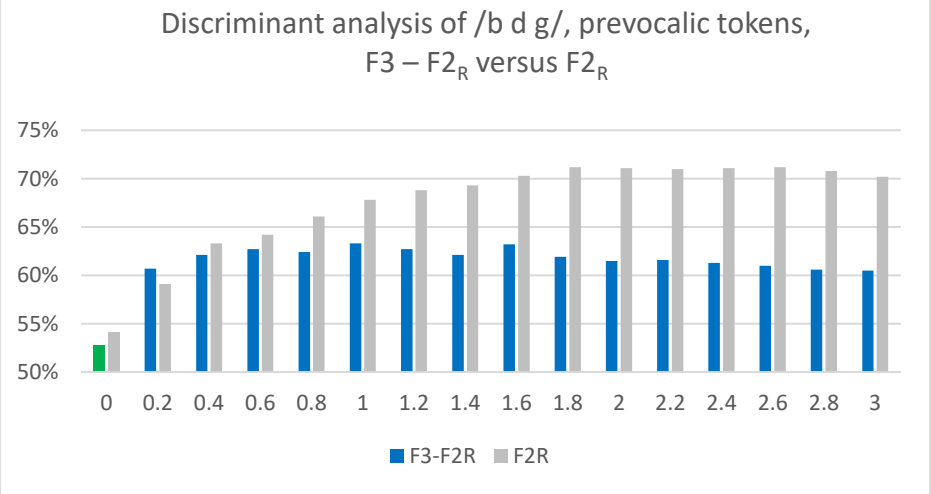


Figure 5.32: Classification accuracy of $F3 - F2_R$ relative to $F2_R$ over the same range of values of *c* that have been utilized throughout this chapter.
Classification run separately on front-vowel and back-vowel contexts and summed. No normalization. N = 1,535.

As with $F2 - F1_R$, the classification accuracy of $F3 - F2_R$ does not outstrip that of $F2_R$. The sole exception to this is when $c = 0.2$. For all other values of *c*, however, $F2_R$ performs better. Over the 16 conditions in Figure 5.32, the mean classification accuracy of $F3 - F2_R$ is 61.3% as against 67.5% for $F2_R$. This 61.3% is similar to the 62.2% found for $F2 - F1_R$ over the same set of conditions.

In terms of statistical significance, the difference in classification between the $F3 - F2_R$ variants and the $F2_R$ variants is statistically significant at the $p < 0.001$ level for all variants of $F2_R$ in which $c = 1.0$ or more. For $c = 0.8$, $p < 0.01$ and for $c = 0$, $p < 0.05$. For $c = 0.2$, 0.4, or 0.6, the difference between $F2_R$ and $F3 - F2_R$ is not statistically significant. Overall, then, the worsening of the classification under most of the conditions shown in Figure 5.32 is highly statistically significant.

Recall that $F3 - F2_R$ was expected to classify better before front vowels than back vowels, since this is the context in which velar pinch is most likely to occur. The results for this particular context, though not presented here, revealed that the performance of $F3 - F2_R$ did not outperform $F2_R$: the performance of $F3 - F2_R$ peaked at 69.6% (for $c = 1$) and for $F2_R$ peaked at 79.0% (for $c = 1.8$, 2, 2.4, and 2.6). Indeed, even the performance of $F2 - F1_R$ exceeded that of $F3 - F2_R$ in the front-vowel context (70.8% for $c = 1.8$).

In short, the performance of formant distances relative to that of F2 is inferior.

5.4.7 Schwa

Up until now, the classification results have been run on all vowel types except for schwa. This has been done because, relative to other vowels, schwa tends to be very short and to lack a clear vowel-quality target. As a result the vowel's formant frequencies tend to be influenced more readily by coarticulation from surrounding segments. Thus it is expected that the classification accuracy of all formant-based attributes, whether $F3_{onset}$ or $F2_R$ or $F2_{onset} + F2_{mid}$, will be reduced when schwa is included in the classification. Including schwa in the original classification would have left it unclear to what extent the classification accuracy was being distorted by the different coarticulatory pattern of schwa. This is one question to be dealt with in the present section.

The aims of this section are: (1) to quantify the decrease in classification resulting from including schwa; (2) to determine whether the value of *c* that is optimal for dealing with the schwa context is similar to that for the non-schwa contexts; and (3) determine whether it is $F2_{mid}$ or $F2_{onset}$ that differs the most in the classification accuracy in the schwa versus non-schwa conditions and, if so, suggest why such a discrepancy would exist.

Here are the results for $F2_R$ on schwa versus non-schwa vowels:

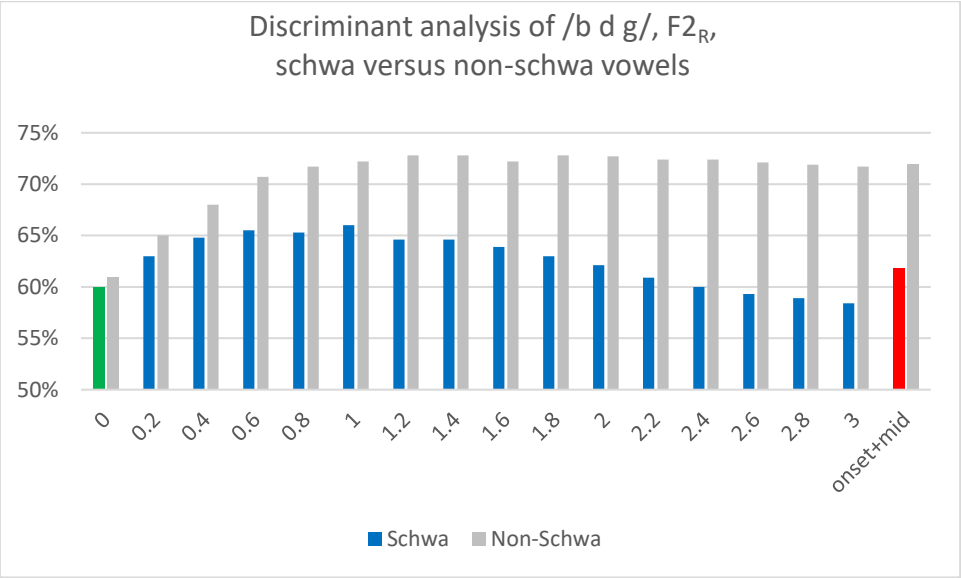


Figure 5.33: The performance of schwa on the $F2_R$ series relative to non-schwa vowels, for CV formant transitions.

Non-schwa vowels were split into front-vowel and back-vowel classifications, with the results combined. Normalization, as before, consists of subtracting $\mu F3_{individual}$. N = 435 (schwa), N = 1,535 (non-schwa vowels).

The peak performance of $F2_R$ for the schwa context (in c = 1) is 66.0%, whereas the peak performance on the non-schwa context (for c = 1.2, 1.4, 1.8) is 72.8%, some 6.8 percentage points higher.

For $F2_{onset}$ on its own (see the two leftmost bars), the classification accuracy for the schwa and non-schwa tokens is very similar: 60.0% versus 60.9%. However, when $F2_{mid}$ is added to the classification (see the two rightmost bars), the classification accuracy for schwa versus non-schwa is strikingly different: 61.8% versus 71.9%. In other words, adding $F2_{mid}$ has enhanced the classification accuracy of $F2_{onset}$ by just 1.8 percentage points for schwa but by 11.0 percentage points for the non-schwa vowels. This weakness of $F2_{mid}$ for the schwa tokens relative to the non-schwa tokens can again be seen in the 15 values of $F2_R$: as the value of c increases from 0.2 to 1, the classification accuracy increases by just 3.0 percentage points for schwa but by 7.2 percentage points for non-schwa tokens. In summary, the difficulty of classifying place of articulation in pre-schwa plosives appears to be specifically due to $F2_{mid}$, not $F2_{onset}$.

This reduced role of $F2_{mid}$ in schwa tokens suggests that the difficulty in correctly classifying place of articulation in pre-schwa voiced plosives using formant information is related to the lack of a well-defined articulatory target for schwa, which perhaps results in $F2_{mid}$ being more variable and/or being heavily influenced by the segment *after* the schwa. However, this latter theory is difficult to examine using the present dataset since in most cases there is no information on the identity of the post-schwa segment (since it is not adjacent to the plosive).

In any event, here are the schwa results for VC context.

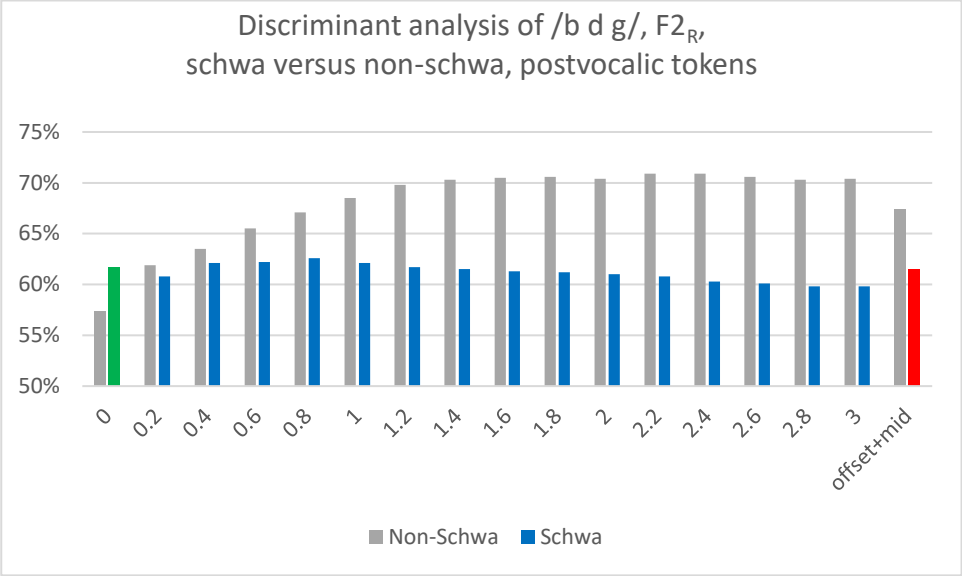


Figure 5.34: The performance of schwa on the $F2_R$ series relative to non-schwa vowels, for VC formant transitions.

Non-schwa vowels have been split into front-vowel and back-vowel classifications with their results combined. Normalized by $\mu F3_{individual}$. N = 536 (schwa), N = 972 (non-schwa vowels).

Surprisingly, $F2_{offset}$ classifies place of articulation in the schwa cases somewhat better than the non-schwa cases, namely 61.7% as against 57.4%. This is different from the result for CV context above, in which the $F2_{onset}$ classification accuracy for schwa cases was 0.9 percentage points less than those for non-schwa cases.

However, one major similarity of the VC context to the CV context is that the result when $F2_{mid}$ is added: the classification accuracy for the schwa cases scarcely improves at all (from 61.7% to 62.6%, for c = 0.8) whereas the non-schwa improves considerably (from 60.9% to 70.9%, for c = 2.2, 2.4). This again suggests that $F2_{mid}$ is qualitatively different in schwa than other vowels. Presumably it is related to the fact that schwa lacks a well-defined quality or articulatory target, though perhaps an additional factor is that schwa is shorter than other vowels, which on its own would lead one to expect $F2_{mid}$ being more similar to $F2_{onset}$ than for other vowels.

5.4.8 Adding Time to F_R

In introducing the F_R concept in Section 5.3 above, c was treated as arbitrary in nature. Since then the classification accuracy has been explored over a wide range of values for c, and under each condition the values of c that yielded the highest classification accuracy has been noted.

In this section the investigation of c continues by incorporating time into the F_R formula. This new type of F_R formula is termed F_RTime.

But why add time? Recall that F_R is simply a variant of F_{onset} that has been modified by F_{mid}: the larger the difference in frequency between F_{onset} and F_{mid}, the larger the modification of F_R away from the value of F_{onset}.

However, F_{onset} and F_{mid} do not just differ in *frequency*, they also differ in *time*. After all, some vowels are short in duration, some vowels are long, and there are countless intermediate durations. This means that the difference in time between F_{onset} and F_{mid} is an uncontrolled variable in the existing formula. Might this be important and, if so, why? If we assume that there is an upper speed limit on how quickly the articulators can move, then a short distance in time between F_{onset} and F_{mid} will tend to result in less of a frequency change between the two points than when the distance in time between F_{onset} and F_{mid} is larger. To compensate for this effect, one would need a formula that makes F_R more dissimilar from F_{onset} whenever the distance in time between F_{onset} and F_{mid} is smaller. In other words the role of time in the F_R formula is inverse to frequency: the greater the difference in frequency between F_{onset} and F_{mid}, the greater the difference in frequency between F_{onset} and F_R (as before); conversely, the *smaller* the difference in *time* between F_{onset} and F_{mid}, the greater the difference in frequency between F_{onset} and F_R.

Thus the new formula is as follows:

$$FreqF_R = FreqF_{onset} - \mathbf{(}((FreqF_{mid} - FreqF_{onset}) \times \mathbf{(}1 / ((TimeF_{mid} - TimeF_{onset}) / c)\mathbf{)}\mathbf{)}$$

'$FreqF_R$' refers to the frequency of F_R, which is the output of the formula, i.e. the acoustic attribute used for distinguishing place of articulation. '$FreqF_{onset}$' is the frequency of F_R at vowel onset, and '$FreqF_{mid}$' is the equivalent measure at vowel midpoint; '$TimeF_{mid}$' is the time at which $FreqF_{mid}$ occurs, and '$TimeF_{onset}$' is the equivalent measure for $FreqF_{onset}$. Thus '$(1 \div ((TimeF_{mid} - TimeF_{onset})) / c^T))$' simply expresses the inverse relationship between time and frequency described above.

Note also that the constant c has not been removed but rather has been rebranded as c^T to reflect its new function: it is used to scale the influence of the time information relative to the frequency information: the larger the value of c^T, the larger the influence of the time difference on the frequency difference. The value of c^T corresponds to milliseconds, and nine values have been tested: 10, 20, 30, 40, 50, 60, 70, 80, and 90. So for example if c^T is set to 30, this means that any $F_{mid} - F_{onset}$ difference of 30 milliseconds will be weighted 1 (since 30 ms divided by the 30 ms value of c^T equals 1), and any $F_{mid} - F_{onset}$ difference less than 30 milliseconds will be weighted greater than 1 (e.g. if the distance between $F_{midpoint} - F_{onset}$ is 15 ms the weight will be 2, since 1 / (15/30) = 2), and any $F_{mid} - F_{onset}$ difference *greater* than 30

milliseconds will be weighted *less* than 1. When this weight is multiplied by the $F_{mid} - F_{onset}$ frequency difference, the result is the amount of F change that will be subtracted from F_{onset} to yield F_RTime.

Given that formant transitions seem to change most rapidly immediately after the release of a consonant and then change more slowly approaching the vowel midpoint (see Stevens, 1998: 341, 356, and 366 for diagrams for each of /b d g/), one might wonder if changing the measurement of time from linear to logarithmic units would result in formant transitions that change at a more even speed over time. If so, it could be the case that the use of logarithmic rather than linear units of time would improve the classification accuracy of the formula, in that it would weight longer distances in time between F_{onset} and F_{mid} more similarly than smaller distances.

Here is the classification accuracy of $F2_R$Time – $\mu F3_{individual}$, comparing time with linear and logarithmic units:

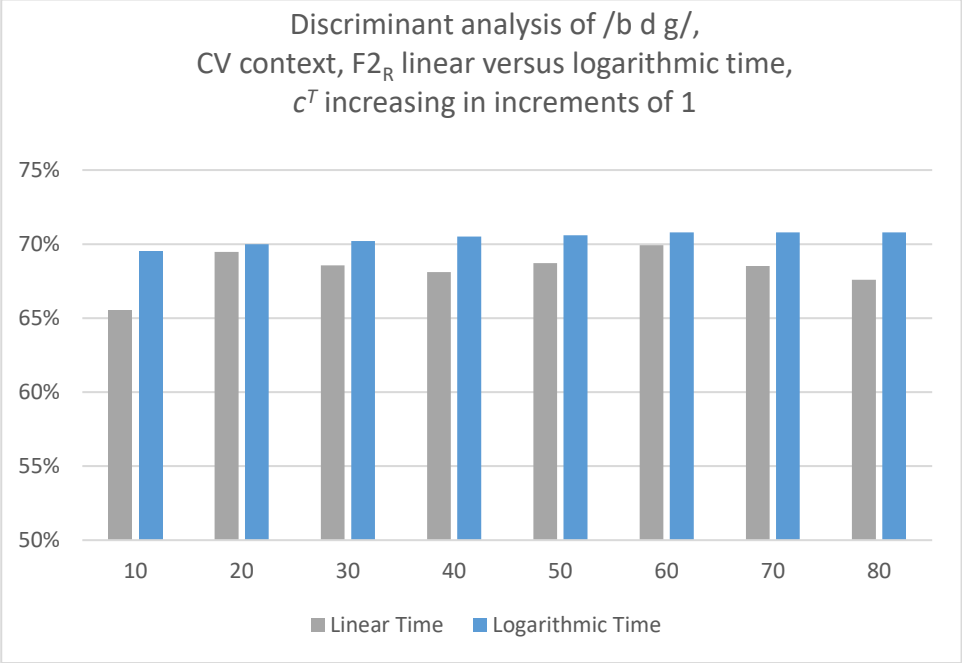


Figure 5.35: Classification accuracy of two kinds of $F2_R$ incorporating time.
Classifications for schwa, front vowels, and back vowels have been run separately and summed. The value of c^T is increased in increments of 10ms from 10ms to 80ms. Normalized by $\mu F3_{individual}$. N = 1,969 (schwa N = 434; front vowel N = 1,106; back vowel N = 429).

The classification accuracy of the logarithmic variant of the formula is slightly higher than that of the linear one (70.9% as against 69.9% at their respective maxima), and this is true over all values of *c* examined. Furthermore, the results for log time are less dependant on the exact value of *c* than those of linear time: that is, the results stay more consistently high.

Nevertheless the classification accuracy of even the log-time formula does not amount to an improvement over that of the original $F2_R$ formula. For the log-time formula the peak classification, as we have noted above, is 70.9%, whereas the optimal classification of the original non-time $F2_R$ formula is 71.0% (for c = 1.2 or 1.4).

In summary, for all its sophistication, incorporating the time difference between $F2_{onset}$ and $F2_{mid}$ has not improved the ability of $F2_R$ to identify place of articulation. Admittedly, incorporating time into the formula has required making some theoretical assumptions and it remains possible that with a different implementation or a different set of assumptions, time could be incorporated into the formula to yield a greater benefit.

5.4.9 The Role of *c* in F_R

Thus the redesignating of c from weighting the influence of $F2_{difference}$ to weighting the influence of time did not amount to an improvement. This raises the question: what would happen to the $F2_R$ formula if c were omitted entirely? The resulting formula is, of course, equivalent to c = 1. To answer the question we simply note the results that have been presented for c = 1. This classification accuracy is 70.9%, which is almost identical to the 71.0% for the optimal classification value over all the values of c examined. (And both of these results are, incidentally, higher than those when $F2_{onset}$ and $F2_{mid}$ are used as classifiers, 69.6%.)

The main objection to abolishing c from $F2_R$ is theoretical in nature rather than empirical. Recall from Section 5.4.4 that the value of c for which the mean $F2_R$ values in alveolars for back vowels and front vowels converged (in both male and female speech) was 0.6. (See Figure 5.20 on page 183.)

For c = 1 it can be seen that the mean $F2_R$ values before back vowels are almost 1 Bark larger than those before front vowels. This is strange in that the 1950s locus theory reviewed in Chapter 2 and again at the beginning of the present chapter indicates that the $F2_{locus}$ frequency for alveolars – more than any other place of articulation – should be the same before front vowels and back vowels. With c = 1 we end up in the curious situation where the alveolar $F2_R$ locus is *higher* before back vowels than front vowels, the reverse of the situation for the other two places of articulation.

However, there is more than one way of thinking about $F2_R$. If the purpose of $F2_R$ is not seen as a means of retracing the 1950s-style $F2_{locus}$ but rather as simply being an $F2_{onset}$ enhancement mechanism, then the fact that the mean /d/ $F2_R$ for back vowels is higher than that for front vowels is less of a problem. Fixing c at 0.6 might indeed result in the front- and back-vowel mean /d/ $F2_R$ values being the same (Section 5.4.4), but it results in a classification accuracy (69.5%) that is 1.4 percentage points less than when c is omitted from the formula

(i.e. $c = 1$). However, an even stronger argument in favour of not setting c to 0.6 is the fact that no value of c is capable of purging the coarticulatory effect of the vowel context entirely, which was shown particularly vividly for /g/ in Section 5.4.4 above (recall that it was estimated that c would have to be set to an absurdly high value, approximately 15, for the mean $F2_R$ value before front vowels and back vowels to converge).

If it is simply not possible for the influence of the vowel context to be removed from the $F2_R$ values entirely (as argued by Al-Tamimi, 2007), then that warrants having a separate $F2_R$ value for different vowel contexts, as was done in the present study with the separation of front vowels, back vowels, and schwa. Under these conditions, the classification accuracy was 72.8% when c was set to 1.2 or 1.4 and a mere 0.6 percentage points less when c was omitted from the $F2_R$ formula (i.e. $c = 1$). (When $F2_{onset}$ and $F2_{mid}$ were used as classifiers the accuracy was below this figure, 71.9%.) Thus it does not particularly matter whether one omits c from the $F2_R$ formula when vowels are separated by backness or schwa in the sense that the resulting classification will be about the same as when $F2_{onset}$ and $F2_{mid}$ are used as classifiers instead.

Still, one might wonder to what extent separating the classification by vowel backness is tipping the exercise in favour of $F2_R$ over $F2_{onset}$ and $F2_{mid}$. When there is no separation by backness, the classification accuracy of $F2_{onset}$ and $F2_{mid}$ is indeed better than that of $F2_R$: 67.2% as against 66.5%. However, none of the values of c manage to outperform $F2_{onset}$ and $F2_{mid}$: the best result, for $c = 1.6$, is 66.9%. To summarize, omitting c from the $F2_R$ formula hurts the classification accuracy by a mere 0.4 to 0.6 percentage points. As was found in 5.4.1, such small differences in classification accuracy are generally not statistically significant.

In conclusion, the simplest $F2_R$ formula – in which there is no c constant – yields a classification accuracy that is close to its maximum observed accuracy, especially when vowels are separated by backness or schwa. Thus collapsing the two attributes $F2_{onset}$ and $F2_{mid}$ into a single attribute, $F2_R$, and omitting the constant c is feasible, with little sacrifice in accuracy.

Nevertheless, the discrepancy between $F2_R$ and the 1950s $F2_{locus}$ theory is important to keep in mind, and in Section 8.7 this issue will be revisited with an alternative version of the $F2_R$ formula being sketched in which $F2_{mid}$ is replaced by an average of the F2 values occurring within 250 ms of $F2_{onset}$. Such a formula seems to have the potential to yield $F2_R$ values that are more in line with what would be expected from the 1950s locus theory and achieves this without the arbitrary constant c.

5.4.10 The Final Picture

In this section, we present the results for VC and CV contexts, though this time showing the overall classification accuracy for all vowels, not just those that are not schwa as we did in

earlier sections. The classification accuracy is also examined for those cases in which there is both a preceding and following vowel. The attributes chosen are: the variant of $F2_R$ in which there is no *c* constant (which will be referred to simply as '$F2_R$') and $F3_{onset}$ (for CV context) or $F3_{offset}$ for (VC context).

We begin with the tokens that have both a preceding and following vowel, VCV (N = 799). The classification has not been separated by vowel backness or schwa due to the small number of tokens.

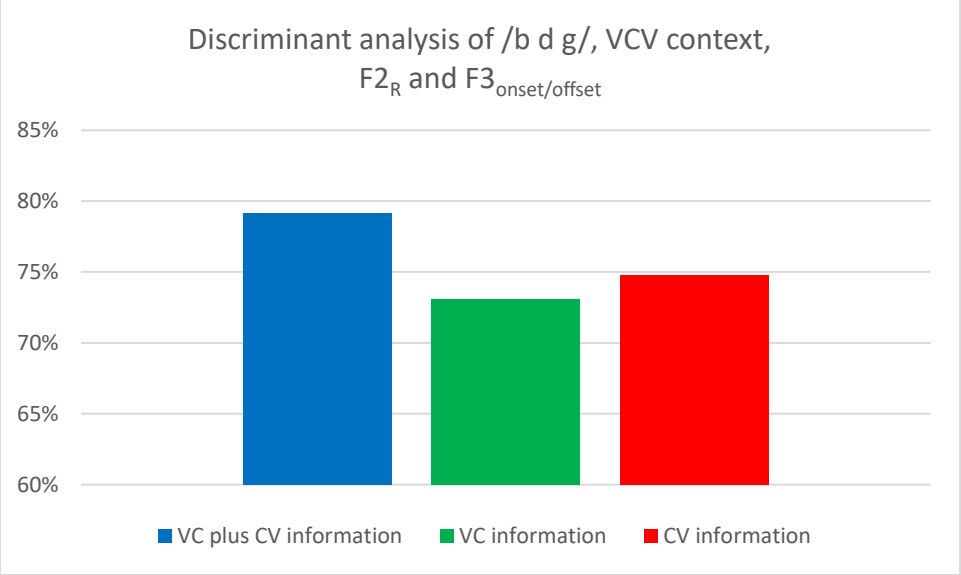


Figure 5.36: Combined classification accuracy of $F2_R$ and $F3_{onset/offset}$ on intervocalic voiced plosives under three conditions: VC information alone, CV information alone, and combined CV and CV information. Normalized by $\mu F3_{individual}$. N = 799. Note that $F2_R$ does not contain the constant *c*, which is the same as setting *c* = 1.

The contribution of the VC information is slightly less than that of the CV information (which was already established in 5.4.5). Also, having formant transitions from both sides of the plosive boosts the classification accuracy, by 4.2 percentage points relative to having CV transitions alone and 6.1 percentage points relative to having VC transitions alone. The combined classification accuracy of 79.2% is among the highest percentages in this chapter (no doubt it would have been even higher if the sample size had been large enough to run a separate classification for back vowels, non-back vowels, and schwa).

In terms of statistical significance, the difference in classification between the blue condition (VC + CV) and the green condition (VC) in Figure 5.37 is $p < 0.001$, while the difference between the blue and red (CV) conditions is also significant ($p < 0.01$). However the difference between the green (VC) and red (CV) conditions is not statistically significant. This is surprising in that it was shown in Section 5.4.5 that the VC transitions classify less accurately on average than the CV transition. It should be borne in mind, however, that the dataset used in

the present section is a subset of the data used in 5.4.5, since the present section is concerned with the classification of VCV tokens, which are necessarily smaller in number than real-life speech than VC and CV tokens.

The following bar chart compares the results from above with the equivalent results when $F2_{onset/offset}$ and $F2_{mid}$ are used instead of $F2_R$:

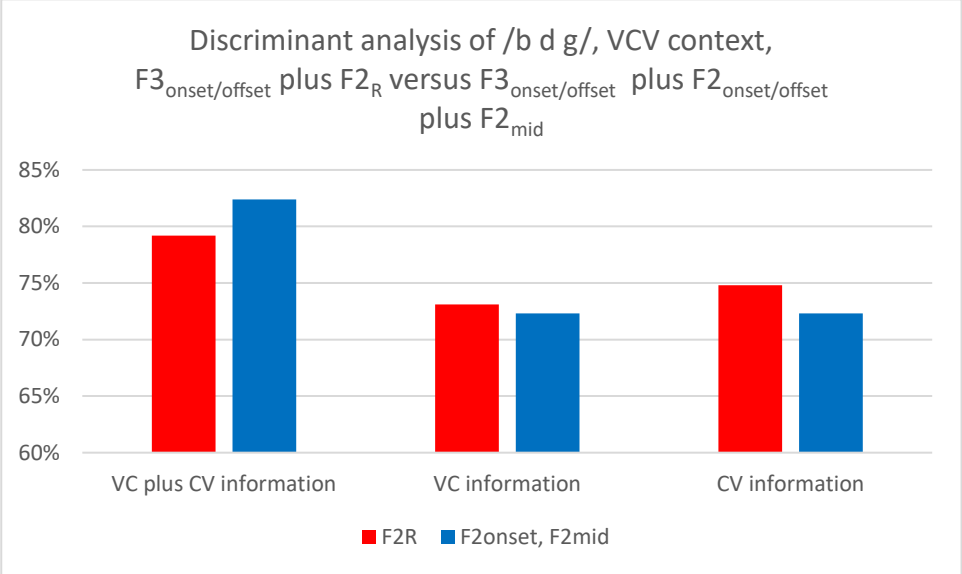


Figure 5.37: Comparison of $F2_{onset}$ and $F2_{mid}$ as classifiers relative to $F2_R$ on voiced intervocalic plosives. $F3_{onset/offset}$ included in all classifications. Normalized by $\mu F3_{individual}$. N = 799.

The classification accuracy in two out of three of the conditions is better when $F2_R$ is replaced by $F2_{onset}$ and $F2_{mid}$; indeed, the result for VC plus CV information is some 3.2 percentage points better. The difference is statistically significant in the CV and CV-plus-VC pairs ($p < 0.01$), but not in the VC condition. However, given the relatively small sample involved in the present data subset (less than half the N = 1,535 dataset used for most of this chapter), it seems wise to take these results with caution.

Here are the results for those CV tokens that lack a preceding vowel (i.e. #CV or CCV):

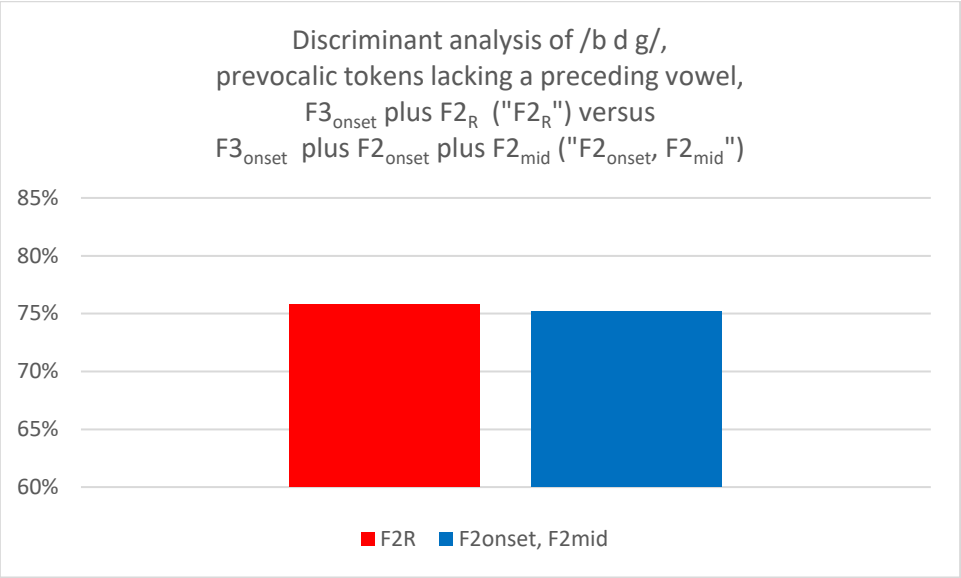


Figure 5.38: Classification accuracy of $F2_R$ versus $F2_{onset}$ and $F2_{mid}$ on those prevocalic voiced plosives that are not preceded by a vowel.

$F3_{onset/offset}$ included in all classifications. Normalized by $\mu F3_{individual}$. N = 1,146.

The results are very similar, although in this case it is $F2_R$ that is slightly stronger. This difference is not, however, statistically significant. And now the results for those VC tokens that lack a following vowel (i.e. VC# or VCC):

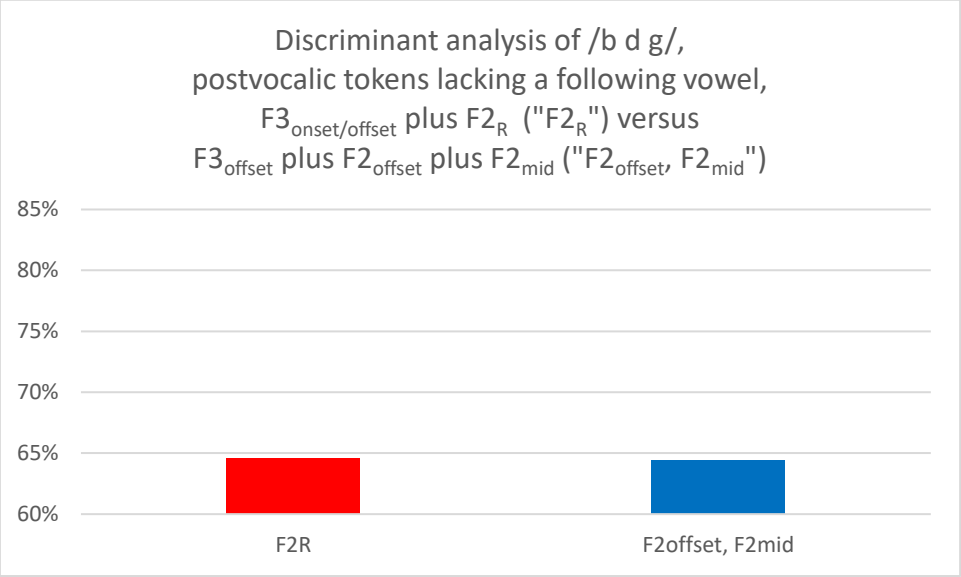


Figure 5.39: Classification accuracy of $F2_R$ relative to $F2_{offset}$ and $F2_{mid}$ on those VC tokens that lack a following vowel.

Normalized by $\mu F3_{individual}$. N = 714.

The classification of $F2_R$ (64.6%) is almost identical to that of $F2_{onset}$ and $F2_{mid}$ (64.4%). Both figures are substantially lower than for those cases involving a following vowel. The difference between the two is not statistically significant.

And finally, here are results for the above three contexts combined together:

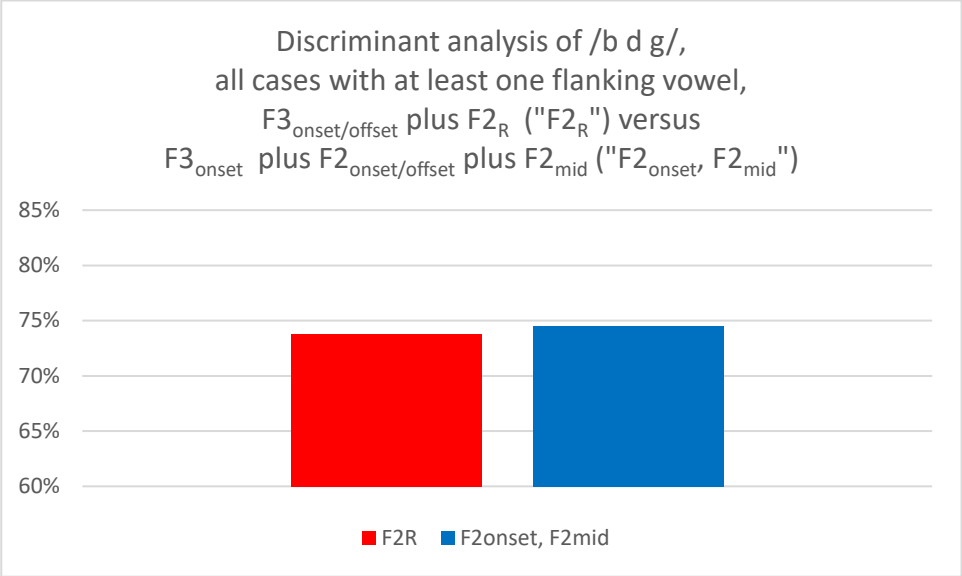


Figure 5.40: Classification accuracy of $F2_R$ + $F3_{onset/offset}$ relative to $F2_{offset}$ +$F2_{mid}$ + $F3_{onset}$ on plosive tokens in the dataset that contain at least one flanking vowel. Normalized by $\mu F3_{individual}$. N = 2,659.

The classification accuracy is extremely similar: 73.8% for $F2_R$ as against 74.5% for $F2_{onset}$ and $F2_{mid}$ (not statistically significant). The two figures, slightly below 75%, are somewhat lower than expected. This seems to be largely due to the VC cases lacking a following vowel, whose classification accuracy as we have seen is 64.4% for $F2_R$ and 64.6% for $F2_{onset}$ and $F2_{mid}$. When these tokens are taken out, the overall classification accuracy rises to 77.2% for $F2_R$ and 78.1% for $F2_{onset/offset}$ and $F2_{mid}$.

It was noted earlier in this chapter that the classification accuracy in VC context tends to be somewhat lower than in CV context. However, this comparison was done with schwa excluded. It turns out that, in the present data set at least, schwa is more likely to occur in VC context (36.7% of tokens) than in CV context (22.1% of tokens). It is likely that this is a factor in why the classification accuracy of the non-prevocalic VC tokens is over 10% less than for the other contexts.

5.5 Discussion

No matter how phonetic contexts are lumped and split, the classification of place of articulation using formant-transition information alone does not normally exceed 80%. This is in line with the findings of previous studies: Hasegawa-Johnson (1996: 25) notes that the strongest study he reviewed (Sussman et al., 1991) had a classification accuracy of 77% but that most other studies had classification accuracies of around 65% to 70%. It should be cautioned, however, that those results and the present study's results are for English; in contrast, for a language in

which ‘voiced’ plosives are actively prevoiced, the results for the formant transitions might be different (e.g. Al-Tamimi, 2007).

The fact that the present study’s 73.8% to 74.5% result is somewhat higher than other studies is likely to be due at least in part to the normalization technique of $\mu F3_{individual}$, though it is unlikely that this is the sole factor, since the four normalization techniques reviewed only boosted the classification accuracy between 1.5 and 2.7 percentage points. A larger boost was achieved when the classification task was split for front vowels, back vowels, and schwa. (Recall from 5.4.3 that separating front from back improved the classification accuracy by an average 5.8 percentage points over the 17 conditions examined.)

Errors in formant tracking were not corrected as it was felt that this would introduce a subjective, difficult-to-accurately-replicate element into the results (see Section 8.2 for discussion). Furthermore, the study’s materials were recorded in ideal acoustic conditions, i.e. in an anechoic booth. Despite this view, a classification was run in which formant tracking errors *were* corrected, to see to what extent the classification accuracy would improve. All formant tracking errors in CV /b d g/ context were corrected (N = 1,936), which involved changing 242 values of $F2_{onset}$, 242 values of $F2_{mid}$, and 440 values of $F3_{onset}$. When a discriminant analysis was run consisting of $F2_{onset}$, $F2_{mid}$, and $F3_{onset}$, the result was 68.8% accuracy prior to correction and 69.5% afterwards, an improvement of just 0.7 percentage points. As expected, correcting formant frequencies made little difference to the classification accuracy.

5.6 Summary

Here is a summary of the present chapter’s main findings:

- $F2_R$, at its peak performance (for values of *c* between 1 and 1.8), distinguishes the place of articulation of /b d g/ as well as $F2_{onset} + F2_{mid}$.
- $F2_R$ does not work on /p t k/, since $F2_{onset} + F2_{mid}$ itself does not work on /p t k/. This is presumably due to the fact that $F2_{onset}$ is located further away from the release of the plosive in /p t k/ relative to /b d g/.
- Including $F3_{mid}$ in the classification does increase accuracy beyond what $F3_{onset}$ achieves on its own (at least when both have been normalized by $\mu F3_{individual}$). This is unlike the pattern for $F2_{onset}$ and $F2_{mid}$ in which including $F2_{mid}$ aids accuracy.
- Normalization in which the mean $F2_{mid}$ or $F3_{mid}$ of a particular speaker or sex is subtracted from $F2_R$ improves the accuracy of $F2_R$, though only by between 1.5 and 2.7 percentage points. Normalization that uses mean F3 rather than mean F2

improves the classification accuracy more, though the difference between the two is slight (0.6 percentage points on average).

- Separating front vowels from back vowels improves the classification accuracy of $F2_R$ much more than normalization does (the average improvement being 5.8 percentage points). Including $F3_{onset}$ leads to an even greater increase in accuracy (7.4 percentage points).
- VC transitions are less reliable at distinguishing place of articulation than CV transitions, the average being 68.0% as against 70.8% for the CV transitions.
- The formant distances $F2_R$ – F1 and F3 – $F2_R$ do not classify place of articulation as accurately as $F2_R$, averaging over the 16 conditions 62.2% and 61.3% respectively as against 67.5% for $F2_R$.
- In schwa $F2_{onset}$ classifies place as well as the $F2_{onset}$ of non-schwa vowels. However, unlike non-schwa vowels the addition of $F2_{mid}$ to the classification only improves its accuracy slightly.
- The incorporation of time into the $F2_R$ formula did not substantially strengthen or weaken the ability of the formula to distinguish /b d g/. This was true whether time was scaled linearly or logarithmically, though logarithmic did appear to be more consistent than linear.
- For VCV context, having both the VC and CV transition improved the classification accuracy beyond what either VC or CV could deliver on their own.

Chapter 6: The Burst

In this chapter, we examine the information for distinguishing plosives' place of articulation found in the release burst. This begins with a bird's-eye view of the burst by examining its shape using spectral slices (Section 6.1). Following this mixed-effects modelling is used to derive a more refined picture of the burst (6.2), examining the extent to which each plosive varies as a function of vowel backness and stress. 27 attributes are then introduced that extract information from the release burst (6.3). Many of these attributes are equivalent in that some are derived from the Hz-dB spectrum, some from the Bark-phon spectrum, and some from the Bark-sone spectrum. The attributes will be presented in four groups: (1) spectral moments; (2) other attributes derived from the entire spectrum; (3) attributes derived from the high-frequency part of the spectrum; and (4) attributes derived from the mid-frequency part of the spectrum. The aim is to compare the classification accuracy of the attributes within and across each group (Section 6.4). In addition to providing information about how best to represent and abstract information about the burst, the performance of the attributes as a group can be used to identify which frequency region of the burst is most important for distinguishing place.

Following this the question is addressed of to what extent separating the burst tokens by voicing aids classification accuracy (6.4.2). The same question is also posed regarding the classification of prevocalic and non-prevocalic stops (6.4.3). In 6.4.4 Aim 2 of this thesis is addressed, namely to quantify the improvement in attribute accuracy yielded by two kinds of normalization (already introduced in 2.3.8) that normalize the burst values by individual speaker. In the three subsequent sections (6.4.5-6.4.7), the performance of compound attributes is explored, that is, attributes whose output is the result of combining two attributes, namely spectral tilt (6.4.5), frequency-based normalization (6.4.6) and amplitude-based normalization (6.4.7).

The chapter finishes (6.5) by comparing (using the random-forest statistic) the combined performance of five attributes derived from the Hz-dB spectrum with the same five attributes derived from the Bark-phon and Bark-sone spectra. This is a second means of investigating whether any one spectral representation yields an improvement in attribute performance over the others. Furthermore, the statistic allows us to see which of the attributes contribute most to the classification accuracy. These findings will be discussed in light of the theoretical conclusions from the literature review.

6.1 Visualizing the Burst

The dimensions of a spectrogram are time along the horizontal axis and frequency along the vertical axis. This allows the formant patterns to be seen easily, since formants involve changes in frequency. However, such a display is less good at representing the burst. This is because the burst – relative to the formants – is an ephemeral event, which means that having the horizontal axis represent time is unhelpful. Instead, placing frequency on the horizontal axis and amplitude on the vertical axis yields a more informative picture of the burst. Thus spectral slices will be used extensively in this initial exploration of the burst. The spectral representation chosen for exposition is the Bark-sone spectrum in that it is a closer approximation to human perception than the other two spectra.

This bird's-eye tour of the burst begins using a relatively crude method: averaging together the burst spectra of every /p/, every /t/, every /k/, every /b/, every /d/, and every /g/ in the dataset.

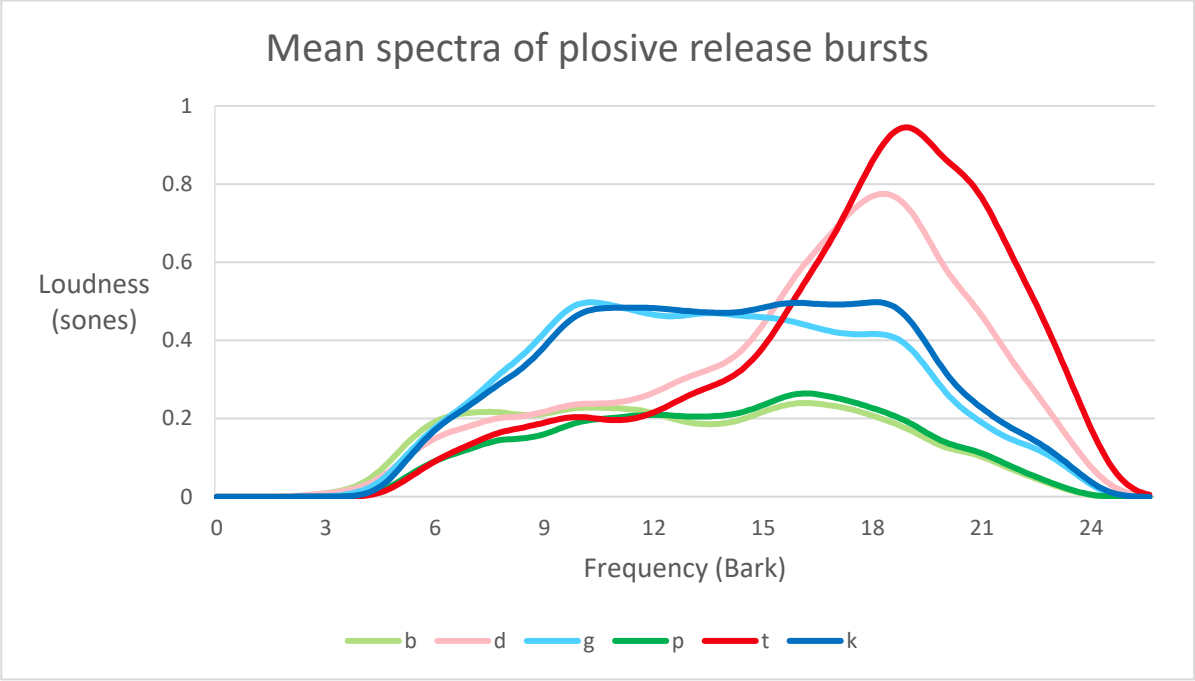


Figure 6.1: Mean spectral envelopes of the six plosive release bursts.
/b/ N = 825, /d/ N = 1,251, /g/ N = 565, /p/ N = 665, /t/ N = 1,178, /k/ N = 987.

As expected based on the previous studies examined in Chapter 2 (e.g. Liberman et al., 1952), the alveolar spectra have a peak in the high-frequency region whereas the bilabial spectra do not appear to have any prominent peak. In terms of voicing, the /b/ and /p/ spectra appear to be strikingly similar to each other, the only difference being that /b/ is slightly louder on average below 10 Bark. With the velars the voiced and voiceless spectra again look strikingly similar. With the alveolars, however, /t/ has a notably louder peak than /d/. Also, /d/ appears to have a

somewhat lower centre of gravity than /t/, since it has *greater* loudness than /t/ in the mid-frequency part of the spectral envelope (i.e. below 15 Bark) and *lower* loudness in the high-frequency region. Thirdly /d/ appears to have a slightly lower mean peak frequency than /t/. As noted in Chapter 2, Zue (1976) found the same pattern (3,300 Hz for /d/, 3,660 Hz for /t/).

At first blush, the velar spectra appear to be intermediate in shape between the bilabial and alveolar types, having a kind of plateau shape. However, it is possible that the lack of a clear peak in the velar spectrum is an artefact of averaging: recall from Section 2.3.1 that the peak in the velar spectrum has been found to vary much more with vowel backness than the peak of the alveolar spectrum. Indeed, when the velar spectra for pre-[o] and pre-[i] contexts are examined on their own, we see that the velar spectrum does indeed tend to have a sharp peak – not the plateau shape of Figure 6.1:

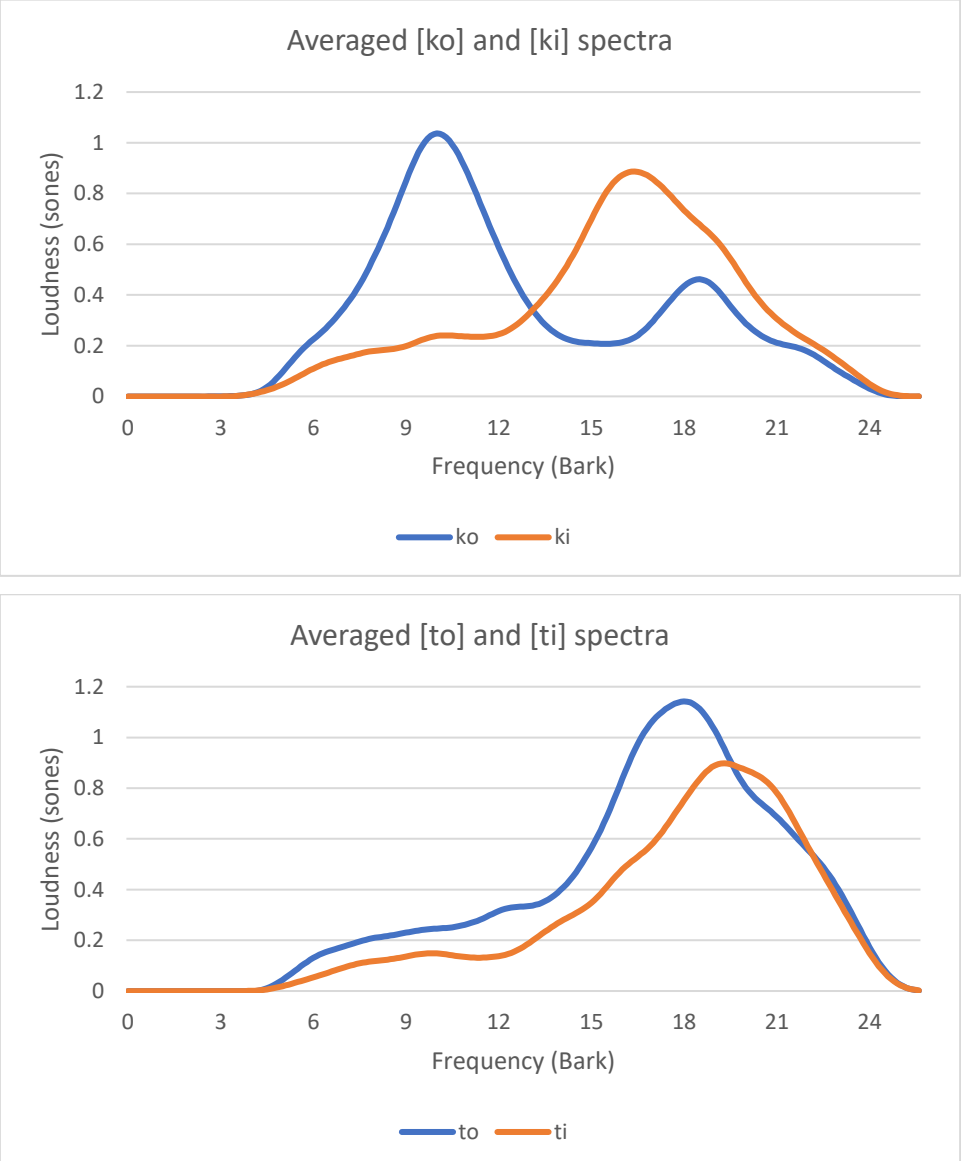


Figure 6.2: (a) Averaged pre-[o] /k/ spectrum (in blue, N = 18); averaged pre-[i] /k/ spectrum (in orange, N = 51); (b) averaged pre-[o] /t/ spectrum (in blue, N = 20), averaged pre-[i] /t/ spectrum (in orange, N = 74).

The /k/ peak before [o] is 10.0 Bark, whereas before [i] it is 16.4 Bark, a difference of over 6 Bark. In contrast, the /t/ peaks before [o] and [i] differ by only 1.2 Bark (19.4 Bark before [i] versus 18.2 Bark before [o]). Thus the amount of variation in the peak frequency of /k/ due to the following vowel's backness appears to be approximately five times larger than that for /t/. This means that when alveolar and velar spectra are averaged without regard to vowel backness as was done in Figure 6.1, the spectra of velars end up looking much flatter than those of alveolars, as we see simply by averaging the two contexts given in Figures 6.2 (a) and (b):

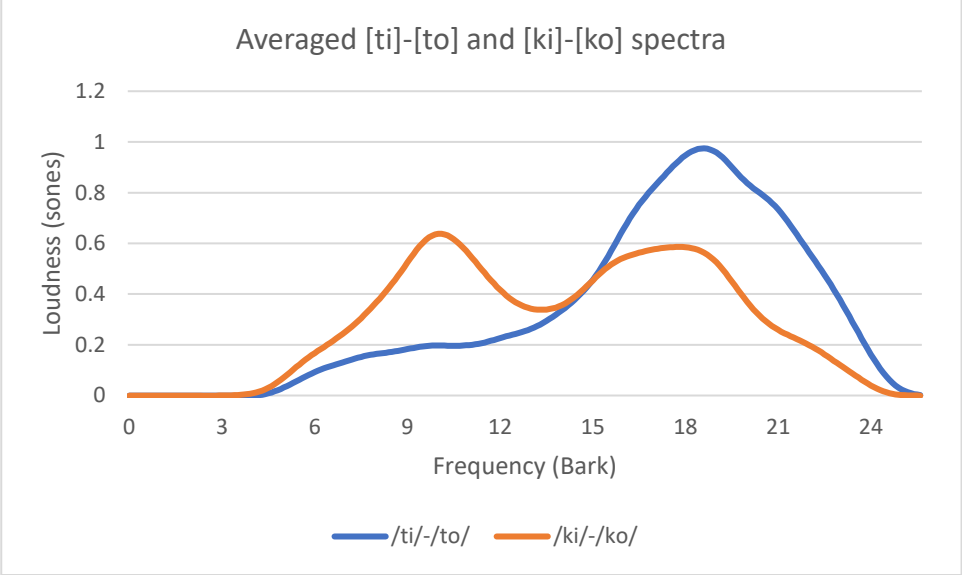


Figure 6.3: The spectra in Figure 6.2 (a) and (b) averaged by vowel backness.
Note how averaging has produced the artefact of flattening the spectrum of /k/ far more than that of /t/.

In conclusion, both velar and alveolar bursts tend to have a clear peak in their spectra, unlike bilabial spectra.

Do bilabial spectra vary much with backness? Here are the averaged spectra for [pi] and [po]:

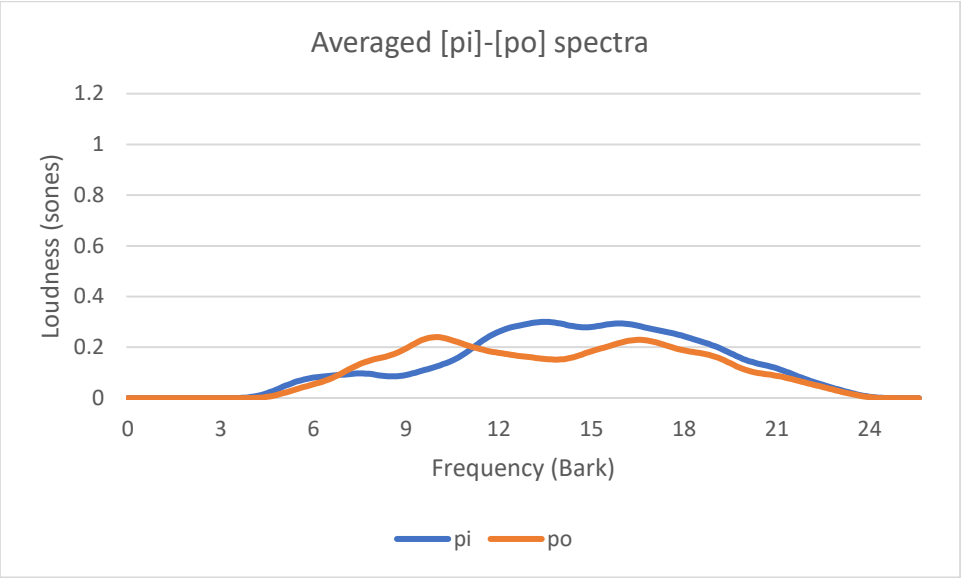


Figure 6.4: Averaged /p/ spectra for [i] context (in blue) and [o] context (in orange). [pi] N = 36, [po] N = 17.

The most striking feature of the bilabial spectra is how flat they appear compared to the other places of articulation. They also have a slightly higher centre of gravity before [i] than before [o], though the difference is modest compared to what we have seen for /k/: the difference in mean centre of gravity of [pi] and [po] is 1.23 Bark (14.90 versus 13.67 Bark), whereas the difference in mean centre of gravity of [ki] and [ko] is 2.80 Bark (15.72 versus 12.92 Bark), over twice as large. The variation between [ti] and [to] contexts (17.82 and 16.86 Bark respectively) is the smallest of the three places of articulation, 0.96 Bark.

There appears to be considerable overlap between the velar and bilabial spectra in terms of their centre of gravity (similar to the findings in Figure 2.19 of Section 2.3.1, from Liberman et al., 1952). But perhaps the more important trait of the bilabial spectrum in comparison with the other two places of articulation lies in the vertical plain: the maximum loudness in its spectrum is much less than that of the other two places of articulation. The average maximum loudness for /p/ is just 0.35 sones, as against 1.11 and 0.88 sones for /t/ and /k/ respectively.

The averaged spectra that introduced the present section were intended to give a general sense of what the variation in the data looks like. The spectra were generated by averaging together a large number of individual spectra. It was noted that such diagrams need to be taken with caution, since they do not represent well the effect that backness has on the spectrum, which is large in the case of velars. Another issue with the averaged spectra is the lack of statistical independence of spectra that come from the same speaker or the same word. These complicating factors will be dealt with using a mixed-effects model (Section 6.3). But first, the burst attributes to be used in this chapter will be introduced.

6.2 Profile of the Burst Attributes

Having presented a bird's-eye view of the burst, the burst attributes to be tested in the present chapter are introduced. These can be divided in two ways: (1) according to the spectrum type (Hz-dB, Bark-phon, and Bark-sone) or (2) according to the type of acoustic attribute. Section 4.5.2 discussed (1) extensively. Hence in our current discussion we introduce the different types of acoustic attribute.

There are four kinds of acoustic attribute: (1) spectral moments (which are derived from the entire burst spectrum); (2) other attributes derived from (close to) the entire spectrum; (3) attributes derived from the high-frequency part of the spectrum; and (4) attributes derived from the mid-frequency part of the spectrum.

6.2.1 Spectral Moments

As outlined in Section 2.3.1, spectral moments quantify the shape of a set of points. The moments used in the present study are centre of gravity (CoG) and standard deviation (SD). As noted in 2.3.1.8, skewness and kurtosis have been argued to give almost no extra information beyond what is yielded by CoG and SD (Wheeler, 2004: 56); in the pilot it was found that skewness and kurtosis had the highest (i.e. poorest) Wilks's Lambda scores out of the burst attributes tested. Thus they will not be brought forward to this part of the study.

CoG is given by the following formula:

$$CoG = \frac{\sum af}{\sum a}$$

This acoustic attribute consists of multiplying the amplitude of each spectral component (a) by its frequency (f), summing all these products, and dividing this sum by the sum of the said components' amplitudes. Note that Praat allows the moments to be calculated on either the absolute spectrum or the power spectrum; Suchato (2004: 40) used the power spectrum on the grounds that it increases the prominence of the spectrum's peaks, which he believes to be particularly informative for place of articulation. This squaring was performed in the present study for the Hz-dB spectrum and, as we saw in the literature review (and as will be found in the results for Sections 7.2 and 7.3), there is indeed evidence that the spectral peaks are especially important for classifying place of articulation. CoG is expected to yield relatively high values on average for alveolars, lower values for velars, with bilabials either equal or lower than velars (the behaviour of bilabials is more difficult to predict due to their relatively flat spectra).

The second spectral moment, standard deviation, is given by the following formula:

$$SDFreq = \sqrt{\frac{\sum a[(f - CoG)^2]}{\sum a}}$$

The function of standard deviation is in effect to quantify the spectrum's flatness. Theoretically a sinusoid would have the lowest SD score possible, whereas white noise would have the highest score. For our purposes this attribute is expected to aid the distinguishing of bilabial and non-bilabial spectra, since as was noted in Section 6.1 bilabial spectra appear to be relatively flat whereas non-bilabial spectra appear to be relatively pointy.

As described in 2.3.1.8, a second variety of standard deviation will be investigated in this study, one in which the frequency domain is bypassed in favour of the amplitude domain:

$$SDAmp = \sqrt{\frac{\sum[(a - \mu)^2]}{n}}$$

The spectral moments are calculated on the entire spectrum from 0 to 10,000 Hz. In this chapter a convention will be applied whereby the name of the spectrum from which the attribute is derived will be appended to the end of the attribute's name. For example, the Centre of Gravity derived from the Hz-dB spectrum will be termed 'CoGdB', while that derived from the Bark-phon spectrum is 'CoGPhon', and that derived from the Bark-sone spectrum is 'CoGSone'. This convention is followed for all attributes.

6.2.2 Peak Attributes

In Section 2.3.1 and elsewhere in the literature review it was noted that the peak of the burst has been identified by both acoustic studies (e.g. Zue, 1976) and perceptual studies (e.g. Li et al., 2010) as being important in the classification of place of articulation; indeed, Li et al. (2010) showed that when the peak is above the noise floor, listeners can usually identify the plosive's place of articulation correctly even if the rest of the burst is beneath the noise floor (Section 2.3.5).

There are two aspects of the peak that can be examined. The first is its frequency. In the literature review it was noted that previous studies have found bilabial spectra to lack a well-defined peak (Zue, 1976), which was again noted in the data of the present study in Section 6.1 above. This leaves velars and alveolars as being the two places of articulation with a well-defined peak. Whilst conducting the pilot study it was observed that the lowest peak frequency of a velar consonant was ca. 800 Hz, found in the nonce utterance /hɔː ˈgɔː/. Based on this it was decided to define the peak frequency as the frequency of the spectral component with the greatest amplitude between 750 and 8,000 Hz. There are two variants of this attribute:

1. AllPeakHz
2. AllPeakBark

One might wonder why there are not three variants of this formula. After all, there are three spectra: Hz-dB, Bark-phon, and Bark-sone. Recall from Section 4.5.2, however, that the conversion from phons to sones (unlike the conversion from dB to phons) does not change which frequency component in the spectrum has the greatest amplitude. What this means for our present purposes is that the peak frequency on a phon spectrum and the peak frequency on a sone spectrum will give exactly the same values and classification accuracy. 'AllPeakBark', then, can be thought of as being derived from either the Bark-phon or Bark-sone spectrum. The attribute is expected to yield relatively high values for alveolars, lower ones for velars, with bilabials yielding difficult-to-predict values due to their tendency to lack a clear peak.

The maximum amplitude from 750 from 8,000 Hz can be measured not just in the frequency domain but also the amplitude domain. This yields the following atttributes:

1. AllPeakdB
2. AllPeakPhon
3. AllPeakSone

6.2.3 High-Frequency Attributes

The high-frequency region runs from 3,500 Hz to 8,000 Hz on the Hz-dB spectrum and from 16.7 to 22.4 Bark on the Bark-phon and Bark-sone spectra. All these attributes measure amplitude rather than frequency. There is a batch of three attributes where the amplitude of the component in this region with the greatest amplitude (the peak) is measured:

1. HiPeakdB
2. HiPeakPhon
3. HiPeakSone

There is a second batch of three attributes in which the amplitude measured is the total amplitude of the frequency region:

1. HiTotaldB
2. HiTotalPhon
3. HiTotalSone

6.2.4 Mid-Frequency Attributes

The fourth and final set of attributes pertain to the mid-frequency region. These attributes are exactly parallel to the six from the high-frequency region that have just been presented. The

mid-frequency region runs from 1,250 to 3,000 Hz on the Hz-dB spectrum and from 9.9 to 15.6 Bark on the Bark-phon and Bark-sone spectra.

These are the six:

1. MidPeakdB
2. MidPeakPhon
3. MidPeakSone
4. MidTotaldB
5. MidTotalPhon
6. MidTotalSone

6.3 Modelling Burst Variation

In this section the non-independence of tokens belonging to the same speaker and/or the same word are dealt with using linear mixed-effects modelling. This was run in the R statistical programming language (R Core Team, 2018) using the `lmer` function in the `lme4` package (Bates et al., 2015).

The mixed-effects modelling will be restricted to four of this chapter's acoustic attributes (which constitute the dependent variable in the models): the centre of gravity (CoGSone), the frequency of the burst peak (AllPeakSone), the standard deviation of the spectral components' magnitudes (SDSoneAmp), and the loudness of the burst peak (AllPeakSone). The first two attributes measure frequency, the second two measure amplitude. CoGSone and SDSoneAmp are computed using all the components in the spectrum, whereas AllPeakBark and AllPeakSone merely pick out the one component with the greatest loudness and record its frequency and loudness respectively.

In the literature review several pieces of evidence were reviewed (e.g. Li et al., 2010, Cvengros, 2011) that suggest that it is the burst peak that contains the most important information for place of articulation for listeners, and this is why AllPeakBark and AllPeakSone have been chosen. CoGSone and SDSoneAmp are expected to give similar information for place of articulation as AllPeakBark and AllPeakSone but because they are computed on the entire spectrum, it is something of an open question to what extent their behaviour will differ from that of AllPeakBark and AllPeakSone respectively.

The fixed effect is $F2_{mid}$ in the models presented in 6.3.1 and vowel stress in 6.3.2. It should be borne in mind that $F2_{mid}$ is not being used in such models as a dimension of place of articulation, but rather as an index of vowel backness. In all models, the random effects are speaker (N = 20) and word (N = 140). Given that a separate mixed effects model was run for each plosive – which has the side-effect of shrinking the sample size of the models – it was

decided not to use random slopes in the model (in 5.1.2 it was noted that the mixed-effects model for /g/ when random slopes were included did not converge). As per the mixed effects models presented in Chapter 5, `REML` was set to `TRUE`, while for determing whether the difference between two mixed-effects models was statistically significant, a chi-squared test was run with `REML` set to `FALSE`. This test of statistical significance involves constructing one mixed-effects models in which the variable of interest is present and comparing it with another in which it is absent. This follows the procedures outlined by Winter (2013: 12-13).

6.3.1 Modelling Burst Variation Due to Vowel Backness

We begin by examining how the centre of gravity of each consonant varies as a function of vowel backness; from inspecting Figures 6.2 and 6.4 it is expected that the centre of gravity will be higher before front vowels than back vowels. The magnitude of this variation for each of the consonants is now estimated.

Many studies that deal with the effect of vowel backness on plosives (e.g. Suchato, 2004: 108 as well as Chapter 5 of the present study) treat it as a categorical variable. While this may be justified for certain purposes, in reality backness varies along a continuum, as indicated by the existence of central vowels and by the IPA's provision of centralization (¨), advancement (˖) and retraction (-) diacritics. Thus in this section vowel backness will be represented as a continuous variable, operationalized using the $F2_{mid}$ of the following vowel.

Note that $F2_{mid}$ has been Lobanov-normalized for each individual speaker to facilitate comparison across speakers. On this z-scored scale a central vowel has a value of approximately 0 ± 1, a front vowel has a value between ca. 1 and 2.7, and a back vowel a value of approximately -1 to -2.7.

Centre of Gravity (CoG) has been z-scored, but unlike $F2_{mid}$ it has not been z-scored by individual speaker, but rather by the entire dataset. This is because there is a lot of overlap between the speakers of each sex in mean CoG, unlike the speakers' mean $F2_{mid}$, for which there is zero overlap between the 10 male and 10 female speakers (the lowest mean $F2_{mid}$ among the female speakers, f03, is 11.54 Bark, whereas the highest mean $F2_{mid}$ among the male speakers, m04, is 11.27 Bark; N = 140 and 160 respectively). For example f03 has one of the *lowest* mean CoG values out of the 20 speakers for alveolars (16.1 Bark) even though she is a female speaker. Conversely m09 has one of the *highest* CoG scores for the said consonants (17.8 Bark) even though he is a male speaker. In contrast the mean $F2_{mid}$ value for m09 is lower than for f03, which is expected given the difference in vocal tract length between male and female speakers. Thus inspection of the mean burst frequencies for the individual speakers

suggested that its centre of gravity does not appear to vary neatly with vocal tract length in the way that $F2_{mid}$ does, which is why the z-scoring was not done by individual speaker.

The mean CoG is 15.1 Bark, which is equivalent to a z-scored CoG of 0. A score of 1 corresponds to a CoG of 17.2 Bark with a score of 2 corresponding to 19.3 Bark.

Variation in CoGSone (Bark) as a Function of $F2_{mid}$

Consonant	Intercept [in Bark units]	Slope [in Bark units]
/p/	14.2 (0.08)	0.17 (0.04)
/t/	19.4 (0.05)	0.17 (0.01)
/k/	14.4 (0.04)	0.71 (0.03)
/b/	13.2 (0.08)	0.36 (0.05)
/d/	16.0 (0.05)	0.21 (0.03)
/g/	14.5 (0.05)	0.83 (0.04)

Table 6.1 (a): Linear mixed-effects models for each of prevocalic /p t k b d g/ showing the variation in the burst's centre of gravity (CoGSone) as a function of the following vowel's $F2_{mid}$ (N = 3,550).
The fixed effects are $F2_{mid}$, the random effects are speaker and word. The values in parentheses are the associated standard errors.

Variation in CoGSone (Z-Scored) as a Function of $F2_{mid}$

Consonant	Intercept [Z-scored units]	Slope [Z-scored units]
/p/	-0.44 (0.05)	0.08 (0.03)
/t/	1.05 (0.08)	0.08 (0.02)
/k/	-0.34 (0.05)	0.34 (0.03)
/b/	-0.89 (0.06)	0.17 (0.04)
/d/	0.42 (0.08)	0.10 (0.03)
/g/	-0.28 (0.06)	0.39 (0.04)

Table 6.1 (b): The same model as in Table 6.1 (a) except that CoGSone has been z-scored (using all the burst tokens in the dataset N = 5,471), with $F2_{mid}$ z-scored using each individual speakers' non-schwa vowel tokens (N = 1,535).
The values in parentheses are the associated standard errors.

The results in Tables 6.1 (a) and (b) reflect what was observed in the averaged spectra of Figure 6.1 closely. For example, the Bark-frequency slope for /g/ (0.39) is nearly four times steeper than the slope for /d/ (0.10), and this difference is mirrored in /k/ and /t/: the slope for /k/ (0.34) is over four times steeper than the slope for /t/ (0.08). This is similar to the 'five times as large' estimate that was established using the naked eye when the averaged spectra for [ki ko] and [ti to] were compared in Figures 6.2 (a) and (b).

From inspecting Figure 6.1 it was noted that the averaged spectra for /k/ and /g/ looked strikingly similar to each other: this is reflected in Table 6.1 where the intercept values (-0.34 and -0.28 respectively) and the slope values (0.34 and 0.39) of these two consonants are strikingly similar. This is unlike /t/ and /d/, whose intercepts are 1.05 and 0.42 respectively, a difference of 0.63 z-scored units – approximately ten times the size of the /k/-/g/ difference. This difference between /t/ and /d/ was again noted in Figure 6.1: recall that the peak in the mean /d/ spectrum was less loud than the peak in the mean /t/ spectrum, and the loudness in the non-peak (below 15 Bark) region was greater in /d/ than in /t/. Both factors would contribute to making the CoG lower in /d/ than in /t/.

As regards the bilabials, their intercepts are the lowest of the three places of articulation and their slopes vary little, which suggests that the bilabials' CoG tends to be low regardless of the following vowel. As noted in Chapter 2, this balance of energy towards lower frequencies in many bilabial bursts has been reported in previous research (Li et al., 2010).

The results in Table 6.1 showed the results for CoGSone. Table 6.2 juxtaposes these results with the results for AllPeakBark under identical conditions:

Cons.	Intercept (CoGSone)	Intercept (AllP.Bark)	Slope (CoGSone)	Slope (AllP.Bark)	P-Value (CoGS.)	P-Value (A.P.B.)
/p/	-0.44 (0.05)	-0.49 (0.08)	0.08 (0.03)	0.08 (0.04)	< 0.01	0.055
/t/	1.05 (0.08)	0.93 (0.05)	0.08 (0.02)	0.06 (0.01)	< 0.001	< 0.001
/k/	-0.34 (0.05)	-0.41 (0.04)	0.34 (0.03)	0.37 (0.03)	< 0.001	< 0.001
/b/	-0.89 (0.06)	-0.97 (0.08)	0.17 (0.04)	0.19 (0.05)	< 0.001	< 0.001
/d/	0.42 (0.08)	0.63 (0.05)	0.10 (0.03)	0.07 (0.02)	< 0.01	< 0.01
/g/	-0.28 (0.06)	-0.35 (0.05)	0.39 (0.04)	0.43 (0.04)	< 0.001	< 0.001

Table 6.2: The results of Table 6.1 (b) for CoGSone are reproduced but with the results for AllPeakBark under identical conditions juxtaposed for comparison. Note the striking similarity between the two attributes. All values are represented on a z-scored Bark scale.
The values in parentheses are the associated standard errors.

The results for AllPeakBark have been placed next to those for CoGSone to underscore the similarity between the two attributes. The only case in which the two attributes differ by more than 0.2 z-scored units is in the intercept for /d/, which is 0.21 units higher for AllPeakBark than CoGSone. The averaged spectra in Figure 6.1 show why: centre of gravity, because it is computed on the entire spectrum, includes the louder mid-frequency region of /d/ (relative to /t/) as well as its quieter high-frequency peak (relative to /t/). Thus /d/'s centre of gravity (= CoGSone) tends to be lower than its peak (= AllPeakBark), since the only thing that the latter attribute measures is the frequency of the peak, not the rest of the spectrum.

One caveat about the results for CoGSone and AllPeakBark is that only the variation attributable to vowel backness has been modelled. It could be the case that the places of

articulation differ in terms of how much non-backness-related variation is found in AllPeakBark and CoGSone. Recall from Figures 6.1 and 6.4 that bilabial spectra tend to be relatively flat, that is, they tend to lack a well-defined peak, especially when compared with alveolars and velars. This raises the question of to what extent it is meaningful to measure the 'peak' of a bilabial spectrum. Recall that the attribute that measures the peak, AllPeakBark, defines the peak as the spectral component that happens to have the greatest amplitude. But if a burst is relatively flat, there will be quite a few components in the spectrum with an amplitude similar to that of the so-called peak.

This leads us to expect that bilabial spectra will show more random variability in their peak frequency; 'random' means that there will be variation in the frequency of the peak that cannot be attributed to vowel backness and which remains after controlling for this fixed effect and for the random effects of speaker and word. In the model of AllPeakBark shown in Table 6.1 above, the standard errors for the intercepts of /p/ and /b/ were 0.083 and 0.075 respectively, which is higher than the standard error for the other four consonants (0.052 for /t/, 0.054 for /d/, 0.044 for /k/, and 0.051 for /g/). Regarding the standard error for the slope, this was again higher for /p/ and /b/ than for the other places of articulation: 0.043 and 0.054 as against 0.014 for /t/, 0.025 for /d/, 0.031 for /k/, and 0.043 for /g/. Furthermore, the only case of a result not reaching $p < 0.05$ in Table 6.2 is for a bilabial, namely the variation in AllPeakBark for /p/. In summary there seems to be noticeably more statistical noise in the frequency of the bilabial 'peak' than in the peak of the other places of articulation, which is unsurprising given that it was seen in Figures 6.1 and 6.4 that bilabial spectra are so flat that the peak is a less meaningful concept than for the other two places of articulation.

Having modelled two frequency-based attributes, two amplitude-based attributes are now modelled, namely the standard deviation of the burst components' loudnesses (SDSoneAmp) and the loudness of the burst peak (AllPeakSone). These attributes have been z-scored by individual speaker (i.e. Lobanov-normalized). This is because the variation in burst amplitude, as mentioned in Section 4.2.2, varies considerably between individual speakers (some speakers talked louder than others). Because the z-scoring has been done by individual speaker, it is not possible to relate the numbers in Table 6.3 in a straightforward manner to sone values as was done for Bark values and z-scored CoG in our discussion of Table 6.1. However, it is sufficient to note that high z-scored values indicate higher relative loudness than low z-scores.

Cons.	Intercept (SDSoneAmp)	Intercept (AllPeakSone)	Slope (SDSoneAmp)	Slope (AllPeakSone)	P-Value (SDS.A.)	P-Value (A.P.S.)
/p/	-1.00 (0.04)	-1.03 (0.04)	0.03 (0.02)	0.02 (0.02)	0.13	0.37
/t/	0.75 (0.07)	0.71 (0.07)	0.04 (0.05)	-0.01 (0.04)	0.45	0.78
/k/	0.27 (0.06)	0.35 (0.06)	0.06 (0.04)	0.02 (0.04)	0.10	0.56
/b/	-0.88 (0.05)	-0.91 (0.05)	0.05 (0.03)	0.04 (0.03)	0.08	0.14
/d/	0.37 (0.05)	0.41 (0.05)	-0.02 (0.04)	-0.07 (0.04)	0.05	0.09
/g/	0.15 (0.09)	0.25 (0.10)	0.11 (0.06)	0.06 (0.07)	0.06	0.35

Table 6.3: Linear mixed-effects models for each of the six plosives, run under identical conditions to those described in Table 6.2 except that this time the results pertain to the attributes SDSoneAmp and AllPeakSone (both Z-scored).

The values in parentheses are the associated standard errors.

As with the results in Table 6.2 that compared CoGSone and AllPeakBark, the slopes and intercepts for SDSoneAmp and AllPeakSone are strikingly similar. The slopes for all six consonants are unremarkable in that they are all close to 0, which means that vowel backness has almost no effect on the burst's loudness. This is not surprising since there was no theoretical reason to expect otherwise. This is also reflected in the tests of statistical significance, which in no case reach $p < 0.05$. Contrast this to the results for vowel backness in Table 6.2 in which vowel backness affects the burst at $p < 0.05$ in all cases (except for the AllPeakBark values of /p/).

Instead, the most interesting information in Table 6.3 is the variation in the intercept. The intercepts for the bilabials are the lowest of the three places of articulation, being -1.03 for /p/ and -0.91 for /b/. This dovetails with the observation from Figure 6.1 at the beginning of the chapter that bilabials tend to have the quietest release bursts of the three places. The intercepts for the other two places of articulation are both above 0, with alveolars somewhat higher than velars. This means that the peak in the alveolar burst tends to be louder than that in the velars. Nevertheless the difference in loudness between these two places is far smaller than that between alveolars or velars and bilabials. Thus SD and MaxAmp's function seems to be primarily to distinguish bilabials from non-bilabials.

6.3.2 Modelling Burst Variation Due to Vowel Stress

In this section the same data subset is utilized as in the previous section but this time the change examined is the variation in the burst as a function of following-vowel stress. Unlike $F2_{mid}$ variation, vowel stress will not be treated as a continuous variable. Instead the data will be split categorically, with stressed vowels in one group, unstressed vowels in the other. As before we begin with the results for the frequency-domain attributes, CoGSone and AllPeakBark:

Cons.	Intercept (CoGSone)	Intercept (AllPeakBark)	Slope (CoGSone)	Slope (AllPeakBark)	P-Value (CoGS.)	P-Value (A.P.B.)
/p/	-0.38 (0.26)	-0.45 (0.46)	-0.08 (0.07)	-0.04 (0.13)	0.24	0.74
/t/	1.11 (0.21)	0.82 (0.14)	-0.11 (0.05)	0.02 (0.04)	< 0.05	0.64
/k/	-0.34 (0.40)	-0.41 (0.40)	-0.00 (0.11)	-0.01 (0.11)	0.97	0.96
/b/	-0.85 (0.48)	-0.81 (0.55)	-0.17 (0.13)	-0.31 (0.15)	0.20	< 0.05
/d/	0.41 (0.29)	0.59 (0.22)	0.08 (0.08)	0.12 (0.06)	0.34	0.06
/g/	[failure to converge]	-0.37 (0.68)		0.07 (0.18)		0.66

Table 6.4: Linear mixed-effects models for each of the six plosives, run under identical conditions to those described in Table 6.1 except that the fixed effect is vowel stress rather than vowel F2 frequency.
Note that the intercept indicates the estimated value (of CoGSone and AllPeakSone) for the unstressed-vowel condition, with the slope indicating how much this value changes when the vowel is stressed. The values in parentheses are the associated standard errors.

The slopes for most of the consonants are relatively small, which suggests that stress has little effect on the peak and centre of gravity. The only case where the slope is relatively large is the AllPeakBark result for /b/. However, the magnitude of the effect even in this case is relatively small (-0.31 of a standard deviation as one moves from stressed to unstressed). More importantly, the results do not reach statistical significance in 10 out of the 12 cases. In summary it appears that, to a first approximation, CoGSone and AllPeakBark do not vary as a function of stress in a statistically significant manner.

Here are the results for how the amplitude attributes SDSoneAmp and AllPeakSone vary as a function of stress.[7]

Cons.	Intercept (SDSoneAmp)	Intercept (AllPeakSone)	Slope (SDSoneAmp)	Slope (AllPeakSone)	P-Value (SDS.A.)	P-Value (A.P.S.)
/p/	-1.02 (0.19)	-1.08 (0.20)	0.04 (0.05)	0.09 (0.05)	0.44	0.11
/t/	0.82 (0.43)	0.73 (0.39)	-0.02 (0.12)	0.07 (0.11)	0.89	0.54
/k/	0.02 (0.36)	0.03 (0.37)	0.44 (0.10)	0.60 (0.10)	< 0.001	< 0.001
/b/	-1.08 (0.26)	-1.13 (0.27)	0.23 (0.07)	0.26 (0.07)	< 0.001	< 0.001
/d/	0.85 (0.33)	0.22 (0.32)	0.28 (0.10)	0.42 (0.09)	< 0.01	< 0.001
/g/	-0.18 (0.66)	-0.18 (0.76)	0.42 (0.17)	0.55 (0.20)	< 0.05	< 0.01

Table 6.5: Linear mixed-effects models for each of the six plosives, run under identical conditions to those described in Table 6.1 except that the fixed effect is vowel stress rather than vowel F2 frequency.
Note that the intercept indicates the estimated value (of SDSoneAmp and AllPeakSone) for the unstressed-vowel condition, with the slope indicating how much this value changes when the vowel is stressed. The values in parentheses are the associated standard errors.

[7] The slopes for AllPeakSone are always larger than those for SDSoneAmp since AllPeakSone consists of recording a single amplitude from the burst and is hence more capable of varying than SDSoneAmp, which consists of using all the amplitudes in the burst and is hence less free to vary in its values.

The same comments can be applied to the intercept values of Table 6.5 as were made regarding the values in Table 6.3: bilabials have the lowest values, alveolars the highest values, with velars intermediate. The more interesting information is contained in the slopes: the mean values for the three voiceless consonants across the two attributes is 0.20, whereas the slope for the voiced series is almost twice as large, 0.36. Indeed, if velars are set aside for a moment the true magnitude of this voicing effect seems to be even larger: /p/ and /t/ have very small slopes (averaging 0.07 and 0.03), which indicates that their burst amplitudes are almost unaffected by vowel stress (indeed note the lack of statistical significance). In contrast, /b/ and /d/ are much more affected by stress, with slopes that are almost five times larger (averaging 0.25 and 0.35) than those for /p/ and /t/. The direction of this effect is positive, that is, /b/ and /d/ have higher-amplitude bursts when they are stressed.

Why might this be the case? One possible factor is that because stressed vowels are louder than unstressed vowels, there is more intra-oral airflow. This means that the build-up of air pressure behind the plosive occlusion is larger in stressed syllables than unstressed syllables. A possible secondary factor is that the closure duration of voiced plosives before stressed vowels is on average longer (55.2 ms, N = 1,103) than that before unstressed vowels (40.4 ms, N = 857). The combined result of these two factors is that the build-up of air pressure would be expected to be larger before stressed vowels than before unstressed vowels. A larger build-up in air pressure means there is a greater disparity in the air pressure behind the constriction relative to that in front of the constriction at the moment of the plosive's release. This in turn would result in a greater volume velocity through the constriction during the plosive's release since (as was noted in Chapter 2) volume velocity is one of the conditions that favours turbulence (the other being a small constriction size; Clark et al., 2007). This would lead one to expect the release burst to be longer and louder in stressed position than in unstressed position. SDSoneAmp and AllPeakSone both measure an aspect of the burst's loudness and show this effect of stress.

However, this explanation is incomplete: it would predict that voiceless plosives should also show an effect of stress. Why, then (leaving aside velars for the moment), is this not the case? Voiceless plosives involve an abduction of the vocal folds, which allows the build-up of air pressure behind the plosive occlusion to occur more rapidly even when the airflow is reduced from the following vowel being unstressed. Also, the mean closure duration of voiceless plosives – unlike the voiced series – is approximately the same for the stressed and unstressed contexts (53.9 ms for the former, 54.7 ms for the latter; N = 1,050 and 903 respectively).

This is the picture with the bilabials and alveolars. The velars, in contrast, are heavily affected by vowel stress, regardless of the voicing (0.44 slope for /k/, 0.42 for /g/). There seems

to be no obvious theoretical reason to expect this result. Nevertheless previous work has indeed found greater context-dependent variation in velar bursts' amplitudes than for other places of articulation: Zue (1976: 126) noted that the amplitude of velars was more affected than that of alveolars by the presence of a following consonant. He found that /k g/ were attenuated by 8.2 and 8.7 dB in plosive + sonorant sequences relative to plosive + vowel sequences, whereas /t d/ were attenuated by just 2.3 and 4.0 dB respectively. Zue does not offer an explanation – aerodynamic or otherwise – for this difference.

6.4 Attribute Comparison

The present study's dataset contains tokens of all six English plosives. This raises the question of how best to split the data. The obvious way to do so is by voicing, that is, to run a separate classification for /b d g/ and /p t k/. However, as noted in Section 6.1 the difference betweeen the spectra of the voiced and voiceless series are fairly modest, being most noticeable for /t d/ and smaller for /p b k g/. However, it is also possible to split the data according to whether the plosive is followed by a vowel. Rather than prejudge the question of how best to split the tokens, this section begins by analysing the data with no splitting, in which the aim of the classifer is to tell whether the plosive is bilabial, alveolar, or velar regardless of whether the plosive in question is voiced or voiceless, prevocalic or non-prevocalic.[8]

6.4.1 Attribute Performance on Entire Dataset

6.4.1.1 Spectral Moments

The first batch of acoustic attributes is the spectral moments. Here is their performance on the discriminant analysis:

[8] In the remainder of this chapter, the total number of tokens in the burst-attribute comparison is 5,185 rather than 5,471. This reflects the exclusion of the data from one of the speakers (m08) due to accidental application of pre-emphasis to his burst spectra. Unfortunately this error was detected too late in the data analysis for his data to be re-extracted without pre-emphasis, hence his exclusion. This error does not affect his formant, Bark-phon, and Bark-sone attributes, hence his inclusion in Chapters 5 and 7 and in Sections 6.1 and 6.2.

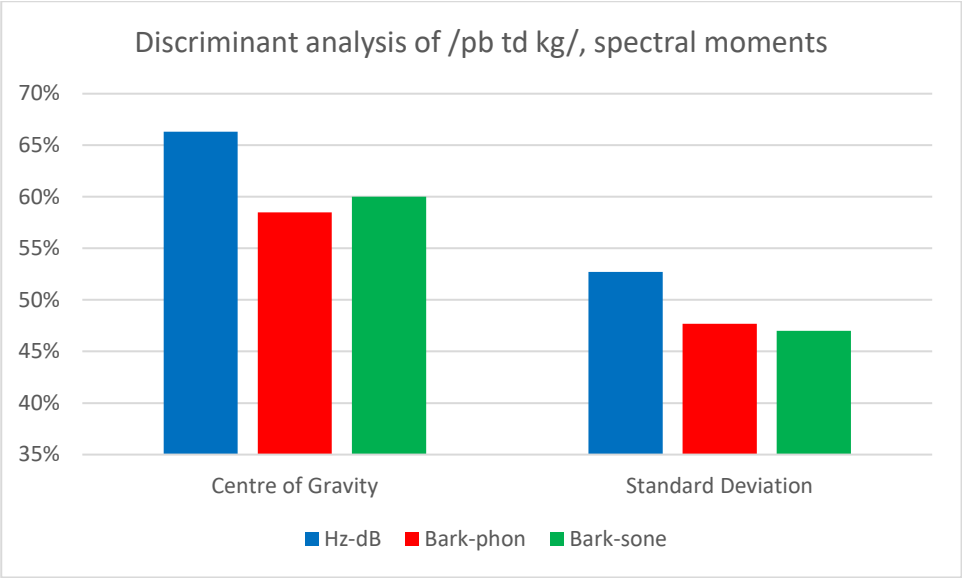


Figure 6.5: Discriminant analysis classification accuracy of the spectral moments Centre of Gravity and Standard Deviation.

Note: all tokens in the dataset with a release burst are entered into a single classification. N = 5,185.

The strongest attribute is Centre of Gravity with a mean classification accuracy over the three spectra of 61.6%. Standard Deviation is much weaker with a mean accuracy of 49.1%. The accuracy of the two moments on the Hz-dB spectrum average 58.9% as against 53.1% on the Bark-phon and 53.8% on the Bark-sone spectra. The final verdict on which of the spectral representations classifies the most accurately will, of course, have to be deferred until all attributes and their combination have been examined.

The pairwise difference in classification between the three kinds of CoG is in all three cases highly significant ($p < 0.001$). For the three kinds of SD, the difference in classification between the dB and Phon variants is significant, as is the difference between the dB and Sone variants ($p < 0.001$ in both cases). The difference in classification between the Phon and Sone variants is not, however, statistically significant.

The statistic *Wilks's Lambda* can be used to determine whether the means of groups (in our case, places of articulation) differ on a discriminant function (Cramer and Howitt, 2004: 181). At one extreme a score of 0 indicates that the means of groups differ, whereas a score of 1 indicates that the means of the groups are the same and so do not differ (ibid.). For our purposes the lower the value, the better the acoustic attribute's performance.

Wilks's Lambda			
Attribute	**Hz-dB**	**Bark-phon**	**Bark-sone**
Centre of Gravity	.524	.701	.703
Standard Deviation	.792	.993	.986

Table 6.6: Performance of the spectral moments on the Wilks's Lambda statistic.

The lower the score the stronger the attribute's group means are at discriminating place of articulation. N = 5,185.

The performance of Centre of Gravity is again highest, with the Hz-dB variant again higher than the Bark-phon and Bark-sone variants. Standard Deviation is considerably weaker; indeed the performance of the Bark-phon and Bark-sone variants is exceptionally poor, both being close to 1 (.993 and .986 respectively). This indicates that the bilabial, alveolar, and velar group means do not differ meaningfully for these attributes.

Recall that the formula for computing standard deviation is:

(1)

$$SD = \sqrt{\frac{\sum a[(f - CoG)^2]}{\sum a}}$$

The formula quantifies (in hertz) how far each spectral component is from the centre of gravity, and then divides this sum by the sum of the spectral components' amplitudes. Standard deviation effectively quantifies how flat a spectrum is.

The formula for standard deviation given above computes standard deviation in the frequency domain. However, given that standard deviation effectively measures variation in *amplitude* (since a flat spectrum is one whose component amplitudes do not vary much), a simplified version of the above formula can be made that involves amplitude rather than frequency:

$$SDAmp = \sqrt{\frac{\sum[(a - \mu)^2]}{n}}$$

In the above formula x represents the mean amplitude of the spectrum, $\bar{x}$ represents the amplitude of each spectral component, Σ represents the sum of these (squared) amplitude deviations, and n represents the number of components in the spectrum.

It is an open question whether this formula can classify place of articulation better than the frequency-domain formula above. Here are the results for this simplified SDPhon and SDSone, which we term SDPhonAmp and SDSoneAmp to reflect the fact that they have been computed from the amplitude alone:

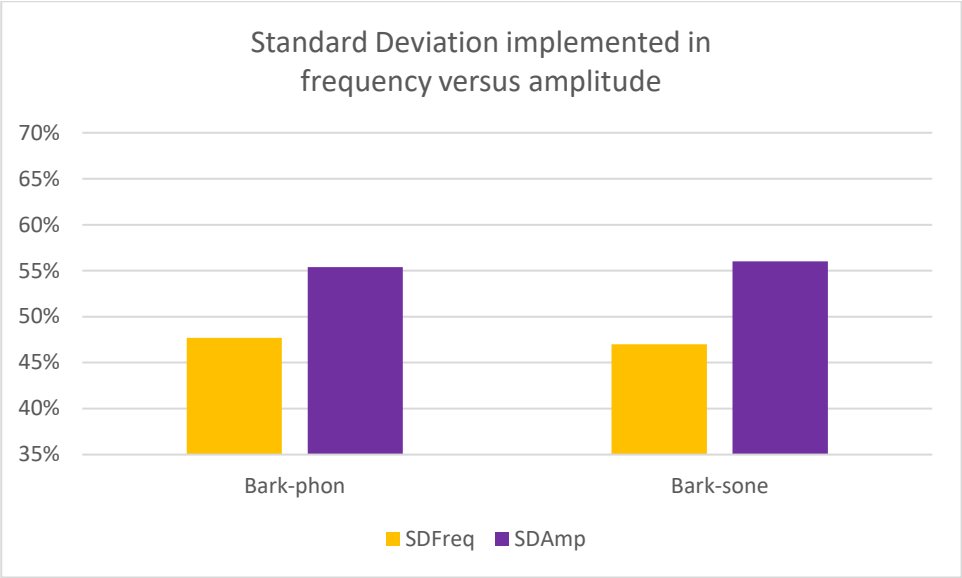


Figure 6.6: Discriminant analysis classification accuracy of Standard Deviation implemented using frequency (SDFreq) versus using amplitude (SDAmp), on both the phon and sone specrtra.
Note: all tokens in the dataset with a release burst are entered into a single classification. N = 5,185.

We see that the classification of the attributes has improved considerably, by 7.7 percentage points in the case of the Bark-phon variant and by 9.0 percentage points in the Bark-sone variant ($p < 0.001$ in both cases). Here is a comparison of the Wilks's Lambda results, which again shows that the performance of Standard Deviation is better when implemented with the amplitude alone:

	Wilks's Lambda	
Attribute	**Bark-phon**	**Bark-sone**
SDFreq	.993	.986
SDAmp	.654	.695

Table 6.7: Performance of the standard deviation on the Wilks's Lambda statistic when implemented in the frequency domain (SDFreq) versus the amplitude domain (SDAmp).
The lower the score the stronger the attribute's group means are at discriminating place of articulation. N = 5,185.

The reduced scores of the amplitude variants vis-à-vis their frequency-domain equivalents indicates that the alveolar, velar, and bilabial group means are more separate from each other in the amplitude-based variants than the frequency-based variants.

The improvement in classification accuracy yielded by implementing standard deviation using amplitude averages 8.35 percentage points over the two conditions. In contrast, the difference in classification accuracy of the phon and sone attributes over the same two conditions is just 0.1 percentage points. This suggests that the choice between these two spectra is unimportant relative to the type of standard deviation implemented.

6.4.1.2 All-Spectrum Attributes

We now turn away from the spectral moments to the other attributes derived from (nearly) the entire spectrum. The present set of attributes are obtained from the frequency region between 750 and 8,000 Hz, which is similar to the spectral moments in being derived from almost the entire spectrum, hence the word 'All'. There are five attributes in total, two based on frequency and three based on amplitude. We begin with the former. These are AllPeakHz (which measures the frequency of the burst's most intense spectral component on the Hz-dB spectrum) and AllPeakBark (which does the same on the Bark-phon spectrum). Earlier it was noted that AllPeakBark gives identical results on the Bark-phon and Bark-sone spectra, which means that the results of its performance below can be regarded as representing both spectra. The results are as follows:

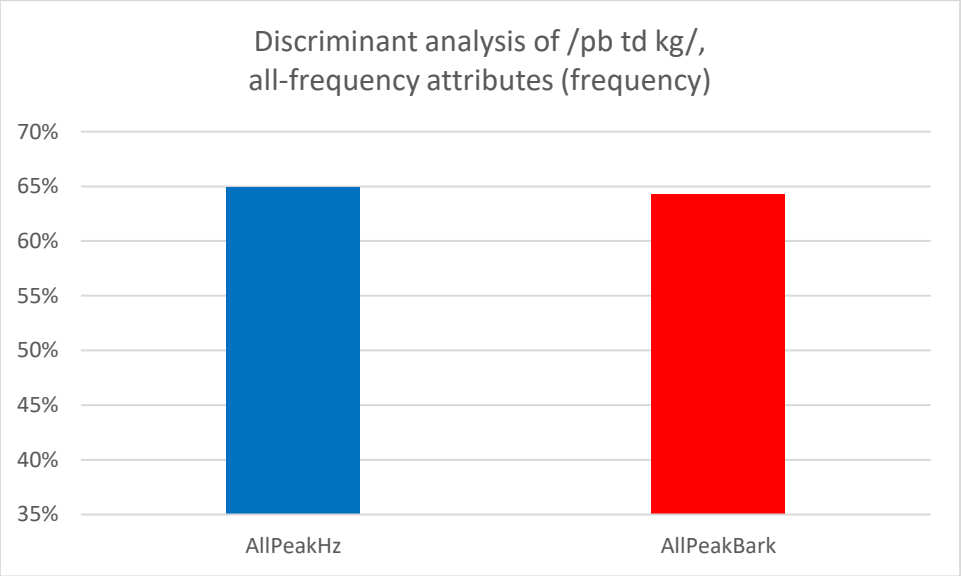


Figure 6.7: Discriminant analysis classification accuracy of the two all-spectrum frequency-of-peak attributes. Note: all tokens in the dataset with a release burst are entered into a single classification. N = 5,185.

The performance of the two attributes is extremely similar, with just 0.6 percentage points separating their accuracy (the difference is not statistically significant). This suggests that the choice between the Bark-phon, Bark-sone, and Hz-dB spectra is relatively inconsequential for this attribute. Neverthless it should be noted that the value of *Wilks's Lambda* is noticeably lower (i.e. better) for AllPeakBark than AllPeakHz:

Attribute	**Wilks's Lambda**
AllPeakHz	.611
AllPeakBark	.536

Table 6.8: Performance of the two frequency-domain 'peak' attributes on the Wilks's Lambda statistic. The lower the score the stronger the attribute's group means are at discriminating place of articulation. N = 5,185.

We turn now to the results for the three amplitude attributes. These consist of measuring the amplitude of the peak in the frequency region from 750 to 8,000 Hz (= 6.9 to 22.4 Bark).

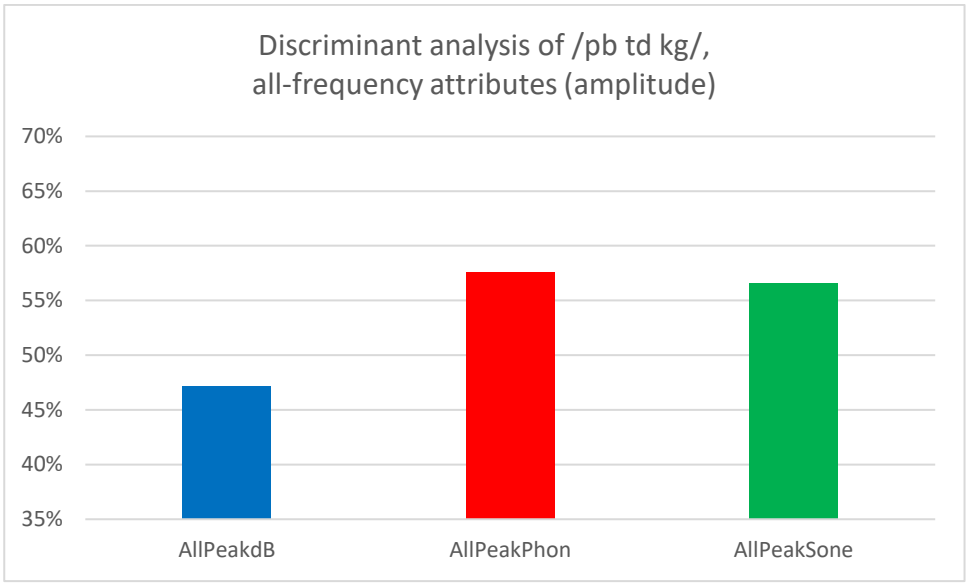


Figure 6.8: Discriminant analysis classification accuracy of the four all-spectrum amplitude attributes.
Note: all tokens in the dataset with a release burst are entered into a single classification. N = 5,185.

The difference in classification between all three variants is highly significant in all three pairwise comparisons ($p < 0.001$). The accuracy of all three attributes is considerably lower than that of the two frequency-domain attributes, which suggests that amplitude information for distinguishing the three places of articulation is less useful than the frequency information.

The results for Wilks's Lambda are as follows:

Attribute	**Wilks's Lambda**
AllPeakdB	.783
AllPeakPhon	.602
AllPeakSone	.685

Table 6.9: Performance of the three all-spectrum amplitude attributes on the Wilks's Lambda statistic.
The lower the score the stronger the attribute's group means are at discriminating place of articulation. N = 5,185.

The order of the attribute's performance is the same as in the discriminant analysis. One might wonder if using the total amplitude rather than the peak amplitude would have improved the performance of the above attributes. The following is a comparison of the two approaches:

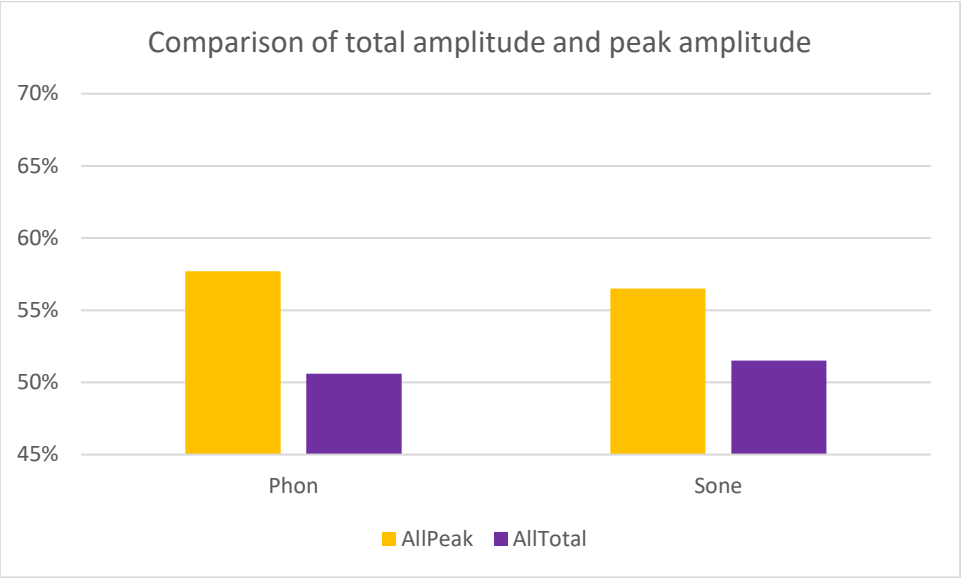


Figure 6.9: Discriminant analysis classification accuracy of AllPeak attributes versus AllTotal attributes.
Note: all tokens in the dataset with a release burst are entered into a single classification. N = 5,185.

The results indicate that summing the loudness of the entire burst and using it as an attribute is a less successful strategy than singling out the spectral component with the greatest amplitude, by 7.1 percentage points on the phon spectrum and by 5.0 percentage points on the sone spectrum. This suggests that the peaks in the spectrum are more informative than the entire spectrum, which is an unsurprising finding in light of the perceptual experiments presented in Chapter 2 (Li et al., 2010; Cvengros, 2011), which found the burst peak to be the most perceptually important part of the burst for listeners.

The difference in classification both between AllPeakPhon and AllTotalPhon and between AllPeakSone and AllTotalSone is statistically significant ($p < 0.001$), as is the difference between AllPeakPhon and AllPeakSone ($p < 0.01$), though the difference between AllTotalPhon and AllTotalSone is not.

Here are the results for Wilks's Lambda:

Wilks's Lambda		
Attribute	**Bark-phon**	**Bark-sone**
AllPeak	.602	.685
AllTotal	.761	.751

Table 6.10: Performance of the all-frequency amplitude attributes on the Wilks's Lambda statistic when the peak amplitude is picked out (AllPeak) relative to summing the entire set of amplitudes (AllTotal).
The lower the score the greater the difference between the attribute's group means. N = 5,185.

It is again found that the attributes that involve summing the total amplitude in the burst are weaker than those that pick out the single largest amplitude. The mean classification accuracy

of the two phon attributes is 54.15% and for the two sone attributes it is 54.0%, a difference of just 0.15 percentage points. In contrast, the mean classification of the two AllPeak attributes is 57.1% whereas the AllTotal ones are just 51.05%, a difference of 6.05 percentage points. Thus the choice between spectral peaks and spectral totals is much more important than the choice of spectrum (similar to the results for SD).

6.4.1.3 High-Frequency Attributes

The third batch of acoustic attributes pertain to the high-frequency region of the burst spectrum (i.e. from 3,500 to 8,000 dB or 16.7 to 22.4 Bark).

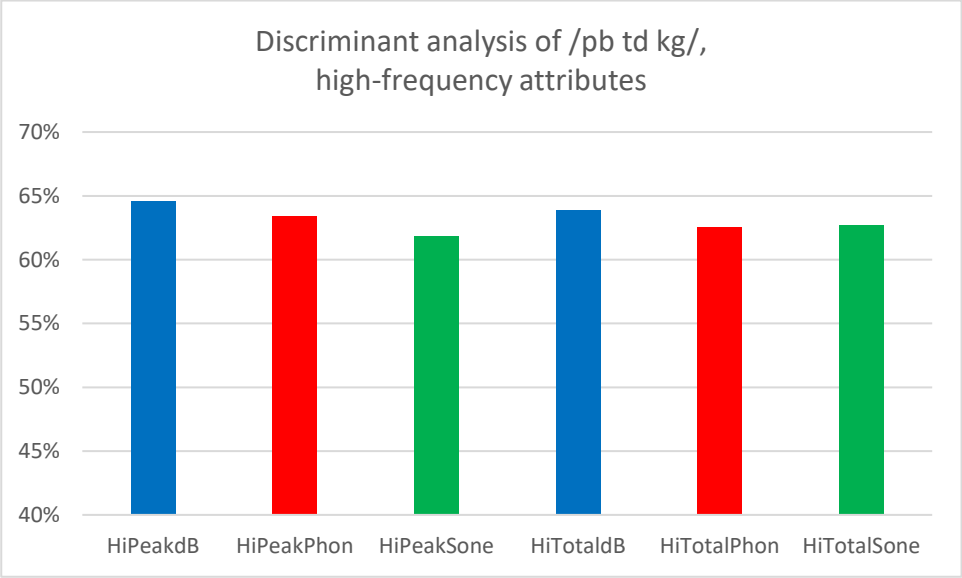


Figure 6.10: Discriminant analysis classification accuracy of the six high-frequency amplitude attributes. The first three attributes measure the peak amplitude whereas the second three measure the total amplitude. Note: all tokens in the dataset with a release burst are entered into a single classification. N = 5,185.

The performance of these attributes is relatively strong, with all six being above 60% accurate. The strongest one, HiPeakdB (64.9%), is almost as strong as CoGdB (66.3%) and is the strongest of all the amplitude attributes. The same attribute derived from the Bark-phon representation, HiPeakPhon, has a slightly lower accuracy of 63.4%, and the same attribute derived from the Bark-*sone* representation, HiPeakSone, has a still lower accuracy of 61.8%.

The difference in the HiPeakdB and HiPeakPhon classifications is statistically significant ($p < 0.01$), as is the difference between HiPeakPhon and HiPeakSone ($p < 0.001$) and between HiPeakdB and HiPeakSone ($p < 0.001$).

The performance of the attributes based on the peak amplitude versus the total amplitude is similar: HiTotalSone performs fractionally better than HiPeakSone (62.3% versus 61.9%, $p < 0.05$) but the performance of HiTotaldB is slightly weaker than that of HiPeakdB (63.9% versus 64.6%, not statistically significant). This is different from the results for the all-

frequency attributes in Figure 6.9 above in which the Peak attributes outperformed the Total attributes by 6 percentage points. For the high-frequency attributes, then, the choice between amplitude peaks and amplitude totals appears to be relatively inconsequential.

Although the choice of spectrum type was statistically significant for the HiPeak attributes, for the HiTotal attributes this is not the case: the difference between the HiTotaldB and HiTotalPhon classification is not statistically significant and the same is true of the difference between the HiTotalPhon and HiTotalSone. The difference between HiTotaldB and HiTotalSone *is* significant, though only at the $p < 0.05$ level. Thus the choice of spectrum has little effect on the HiTotal attributes.

The results for Wilks's Lambda are in accord with those for the discriminant analysis, with the minor exception that HiTotalPhon and HiTotalSone have swapped places:

Attribute	**Wilks's Lambda**
HiPeakdB	.537
HiPeakPhon	.551
HiPeakSone	.608
HiTotaldB	.546
HiTotalPhon	.559
HiTotalSone	.594

Table 6.11: Performance of the five high-frequency amplitude attributes on the Wilks's Lambda statistic. The lower the score the stronger the attribute's group means are at discriminating place of articulation. N = 5,185.

6.4.1.4 Mid-Frequency Attributes

The fourth and final set of burst attributes pertains to the mid-frequency region of the spectrum (that is, from 1,250 to 3,000 Hz or 9.9 to 15.6 Bark).

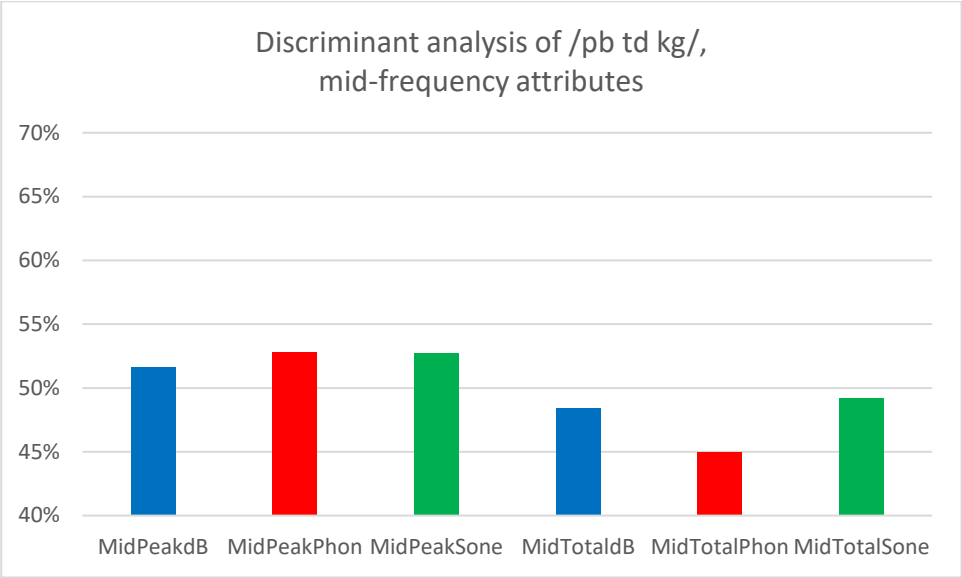


Figure 6.11: Discriminant analysis classification accuracy of the six mid-frequency amplitude attributes. The first three attributes measure the peak amplitude whereas the second three measure the total amplitude. Note: all tokens in the dataset with a release burst are entered into a single classification. N = 5,185.

The classification accuracy for the Bark-phon and Bark-sone attributes is the reverse of what was found for the high-frequency attributes. Nevertheless, the difference in accuracy is relatively small: for example, the classification accuracy of MidPeakdB is 51.6% whereas that of MidPeakPhon is 52.8% and MidPeakSone is 52.7%. The difference between the three is not statistically significant except in the case of MidPeakdB versus MidPeakPhon (though even in this case, the difference is only $p < 0.05$). Furthermore, by far the most notable fact about the mid-frequency attributes is their collective weakness relative to the high-frequency attributes: the mean classification of the six mid-frequency attributes is 50.0% whereas that of the six high-frequency attributes is 63.2%. This 13.2-percentage-point difference is far larger than the 1.1 or 1.2 percentage-point difference between MidPeakdB and MidPeakSone or MidPeakPhon. This strongly suggests that the high-frequency region of the burst is a richer source of information for plosives' place of articulation than the mid-frequency region and, furthermore, that this is a vastly more important factor in the classification accuracy of place of articulation than the choice of spectral representation.

One other difference between the mid-frequency and high-frequency results is that for the high-frequency attributes, the choice between amplitude *peak* and amplitude *total* had almost no effect on mean classification accuracy (0.3 percentage points), whereas for the mid-frequency region (as with the results for the all-frequency attributes in Figure 6.9) there is a more substantial difference in the accuracy of the two types of attribute, with the amplitude-total attributes being the weaker of the two ($p < 0.001$ in all three pairs, i.e. MidPeakdB versus MidTotaldB, MidPeakPhon versus MidTotalPhon, and MidPeakSone versus MidTotal). For

example, MidTotaldB classifies less than MidPeakdB by 3.2 percentage points, and the 3.0-percentage-point discrepancy between MidTotalSone and MidPeakSone goes in the same direction. The difference in classification between the three MidTotal attributes is statistically significant ($p < 0.001$ in the two pairwise cases). However, the difference in the MidTotaldB and MidTotalSone classifications is not statistically significant (which is unsurprising given that the difference in accuracy between the two, as seen in Figure 6.2 above, is just 0.8 percentage points).

Attribute	**Wilks's Lambda**
MidPeakdB	.730
MidPeakPhon	.713
MidPeakSone	.715
MidTotaldB	.816
MidTotalPhon	.843
MidTotalSone	.787

Table 6.12: Performance of the five mid-frequency amplitude attributes on the Wilks's Lambda statistic.
The lower the score the stronger the attribute's group means are at discriminating place of articulation. N = 5,185.

The values for mid-frequency attributes, ranging from .730 to .843, do not overlap with those for the high-frequency region, .537 to .651. This is another indication that the difference between the mid-frequency and high-frequency regions is much more important for the classification accuracy of place of articulation than the choice of spectral representation.

6.4.1.5 Summary of Results

Rank	Attribute	Accuracy (%)
1	CoGdB	66.3
2	AllPeakHz	64.9
3	HiPeakdB	64.6
4	AllPeakBark	64.3
5	HiTotaldB	63.9
6	HiPeakPhon	63.4
7	HiTotalPhon	63.0
8	HiTotalSone	62.7
9	HiPeakSone	61.8
10	CoGSone	60.0
11	CoGPhon	58.5
12	AllPeakPhon	57.7
13	AllPeakSone	56.5
14	SDSoneAmp	56.0
15	SDPhonAmp	55.4
16	MidPeakPhon	52.8
=17	SDdBFreq	52.7
=17	MidPeakSone	52.7
19	MidPeakdB	51.6
20	AllTotalSone	51.5
21	AllTotalPhon	50.6
22	MidTotalSone	49.2
23	MidTotaldB	48.4
24	SDPhonFreq	47.7
25	AllPeakdB	47.2
26	SDSoneFreq	47.0
27	MidTotalPhon	45.0

Table 6.13: Performance of the 27 attributes on the leave-one-out discriminant analysis statistic.

Note: these are the same results as presented through Section 6.4.1 above. N = 5,185.

And here are the results for Wilks's Lambda:

Rank	Attribute	Accuracy
1	CoGdB	0.524
2	AllPeakBark	0.536
3	HiPeakdB	0.537
4	HiTotaldB	0.546
5	HiPeakPhon	0.547
6	HiTotalPhon	0.559
7	HiTotalSone	0.594
8	AllPeakPhon	0.602
9	HiPeakSone	0.608
10	AllPeakHz	0.611
11	SDPhonAmp	0.652
12	AllPeakSone	0.687
13	SDSoneAmp	0.695
14	CoGPhon	0.701
15	CoGSone	0.703
16	MidPeakPhon	0.713
17	MidPeakSone	0.715
18	MidPeakdB	0.73
19	AllTotalPhon	0.761
20	AllTotalSone	0.769
21	AllPeakdB	0.783
22	MidTotalSone	0.787
23	SDdBFreq	0.792
24	MidTotaldB	0.816
25	MidTotalPhon	0.843
26	SDSoneFreq	0.986
27	SDPhonFreq	0.993

Table 6.14: Performance of the 27 attributes on the Wilks's Lambda statistic.

Note: these are the same results as presented throughout Section 6.4.1 above. N = 5,185.

As we can see, the attributes that performed relatively well on the discriminant analysis tended to also perform well on Wilks's Lambda, e.g. the strongest attribute on the discriminant

analysis, CoGdB, is also the strongest attribute on Wilks's Lambda. Other strong attributes are AllPeakHz, HiPeakdB, AllPeakBark, HiTotaldB, and AllPeakPhon.

6.4.2 Voiced and Voiceless Plosives

The results of the previous section showed the classification accuracy of place of articulation for all six phonemes. That is, /p b/, /t d/, and /k g/ were lumped into the one classification. The present section examines the classification accuracy for /b d g/ and /p t k/ separately. This will indicate whether the attributes that perform best at distinguishing the place of voiced plosives are also the best attributes for distinguishing the place of voiceless plosives. Also, we will see whether the classification accuracy of the attributes *on the whole* is stronger for /b d g/ or /p t k/. This is one method of evaluating the importance of the burst as a cue when the consonant is voiced relative to when it is voiceless.

We begin with the results for the voiced series. Note that the column 'Change from the All Condition' records how much the classification accuracy of an attribute has increased or decreased on the voiced series relative to its performance in 6.4.1 above on the entire dataset. That is, the classifier is no longer forced to classify /p/ and /b/, /t/ and /d/, and /k/ and /g/ under the same category. Here are the results:

Rank	Attribute	Classification Accuracy (%)	Change from the All Condition (% pts)	Change from the All Condition (Rank)
1	AllPeakBark	68.4	+4.1	+3
2	HiPeakdB	64.4	-0.2	+1
=3	CoGdB	63.6	-2.7	-2
=3	HiPeakPhon	63.6	+0.2	+3
5	HiPeakSone	62.5	+0.7	+2
6	HiTotalSone	61.5	-1.2	+1
=7	AllPeakHz	61.3	-3.6	-5
=7	HiTotaldB	61.3	-2.6	-2
9	HiTotalPhon	60.2	-2.8	-2
10	CoGSone	59.4	-0.6	0
11	CoGPhon	57.5	-1	0
12	AllPeakPhon	56.6	-1.1	0
13	AllPeakSone	55.2	-1.3	0
14	SDPhonAmp	54.9	-0.6	+1
=15	SDdBFreq	54.0	+1.3	+2
=15	SDSoneAmp	54.0	-2	-1
17	MidPeakSone	53.0	+0.3	0
18	AllTotalPhon	52.5	+1.9	+3
19	AllTotalSone	52.3	+0.8	+1
20	MidPeakPhon	51.6	-1.2	-4
21	SDPhonFreq	50.9	+3.2	+6
22	SDSoneFreq	49.5	+2.5	+5
23	MidPeakdB	49.1	-2.5	-4
24	MidTotalSone	48.9	-0.3	-2
25	AllPeakdB	46.6	-0.6	0
26	MidTotaldB	45.9	-2.5	-3
27	MidTotalPhon	43.6	-1.4	0

Table 6.15: Discriminant analysis classification accuracy of the 27 attributes for /b d g/. Change in classification accuracy relative to the /pb td kg/ condition is shown in the second-from-right column. Results in red are discussed in the text. N = 2,494.

For the most part, the attributes that performed well when /b d g/ were lumped in with /p t k/ (Table 6.13) also tend to perform well in the present /b d g/ data subset. The attribute whose performance improves the most is 'AllPeakBark' whose performance improves by 4.1 percentage points. In contrast 'AllPeakHz', which is a similar attribute but derived from the Hz-dB rather than the Bark-phon spectrum, shows a considerable decline of 3.6 percentage points. Recall that in the All condition AllPeakHz outperformed AllPeakBark by 0.6 percentage points, but in the present /b d g/ data subset, it is AllPeakBark that outperforms AllPeakHz, by 6.9 percentage points. SDPhonFreq and SDSoneFreq also show a relatively large change in classification accuracy (up 3.2 and 2.5 percentage points respectively) but this does not change the fact that they are relatively weak attributes. Other than these few exceptions, however, the data for /b d g/ repeat broadly the same picture as was established for the /pb td kg/ condition in 6.4.1 above. To a first approximation, then, it makes little difference to relative attribute performance whether /b d g/ are lumped in with /p t k/ or classified on their own.

Here are the results for /p t k/:

Rank	Attribute	Classification Accuracy (%)	Change from the All Condition (% pts)	Change from the All Condition (Rank)
1	CoGdB	69.9	+3.6	0
2	AllPeakHz	68.1	+3.2	0
3	HiTotaldB	66.9	+3	+2
4	HiPeakdB	66.7	+1.4	-1
5	AllPeakBark	66	+1.7	-1
6	HiTotalPhon	65.4	+2.4	+1
7	HiTotalSone	64.3	+1.5	+1
8	HiPeakPhon	63.9	+0.5	-2
9	CoGSone	63.9	+3.9	+1
10	HiPeakSone	61.2	-0.5	-1
11	AllPeakPhon	60.2	+2.5	+1
12	CoGPhon	60	+1.5	-1
13	AllPeakSone	59.7	+3.2	0
14	SDSoneAmp	58.3	+2.3	0
15	SDPhonAmp	56.9	+1.5	0
16	MidPeakPhon	55.1	+2.3	0
17	SDdBFreq	55.1	+2.4	0
18	MidPeakdB	55	+3.4	+1
19	MidPeakSone	53.2	+0.5	-2
20	AllTotalSone	51.8	+0.3	0
21	MidTotaldB	50.9	+2.5	+3
22	MidTotalSone	50.3	+1.1	0
23	AllTotalPhon	49.6	-1	-2
24	MidTotalPhon	47.8	+2.8	+3
25	AllPeakdB	47.6	+0.4	0
26	SDSoneFreq	44.8	-2.2	0
27	SDPhonFreq	44.6	-3.1	-3

Table 6.16: Discriminant analysis classification accuracy of the 27 attributes for /p t k/.

Change in classification accuracy relative to the /pb td kg/ condition is shown in the second-from-right column. Results in red are discussed in the text. N = 2,691.

CoGdB, which was the strongest attribute on the overall dataset but only fourth strongest on the voiced data subset, is once more the strongest attribute with a classification accuracy of just under 70%. AllPeakBark, which was strongest on the voiced series and fourth strongest in the All context, ranks only fifth on the voiceless series. Despite these reversals of performance, there is nevertheless a pronounced tendency for attributes that are strong on the voiced series to also be strong on the voiceless one.

The mean classification rate for the 27 attributes on /b d g/ is 55.6%, whereas for /p t k/ the rate is 57.7%. This might suggest that the information in the burst for the voiceless plosives is on average more reliable than that for the voiced series, but the difference between the two groups is nevertheless small.

The following table shows the classification accuracy when the results of the voiced and voiceless series are summed.

Rank	Attribute	Accuracy (%)	Improvement (%)
1	AllPeakBark	67.2	+2.9
2	CoGdB	66.75	+0.45
3	HiPeakdB	65.55	+0.95
4	AllPeakHz	64.7	-0.2
5	HiTotaldB	64.1	+0.2
6	HiPeakPhon	63.75	+0.35
7	HiTotalSone	62.9	+0.2
8	HiTotalPhon	62.8	-0.2
9	HiPeakSone	61.85	+0.05
10	CoGSone	61.65	+1.65
11	CoGPhon	58.75	+0.25
12	AllPeakPhon	58.4	+0.7
13	AllPeakSone	57.45	+0.95
14	SDSoneAmp	56.15	+0.15
15	SDPhonAmp	55.9	+0.5
16	SDdBFreq	54.55	+1.85
17	MidPeakPhon	53.35	+0.55
18	MidPeakSone	53.1	+0.4
19	AllTotalSone	52.05	+0.55
20	MidPeakdB	52.05	+0.45
21	AllTotalPhon	51.05	+0.45
22	MidTotalSone	49.6	+0.4
23	MidTotaldB	48.4	0
24	AllPeakdB	47.1	-0.1
25	MidTotalPhon	45.7	+0.7
26	SDPhonFreq	47.85	+0.15
27	SDSoneFreq	47.05	+0.05

Table 6.17: Discriminant analysis classification accuracy of the 27 attributes when the results for /b d g/ (Table 6.15) and /p t k/ (Table 6.16) are averaged together.

Change in classification accuracy relative to the /pb td kg/ condition (Table 6.13) is shown in the second-from-right column; this indicates how much better the attribute classifies place of articulation when the voicing category is known. N = 5,185 (/b d g/ = 2,494; /p t k/ = 2,691).

Broadly speaking, the ranking of the attributes is similar to what was found when a single classification was used on the voiced and voiceless categories (in Table 6.13). However, AllPeakBark is now the highest-ranked attribute whereas originally it was only ranked fourth, its improvement being 2.9 percentage points as opposed to the 0.45 percentage points for CoGdB, for example. This indicates that the AllPeakBark's performance improves more than other attributes when the voiced and voiceless tokens are classified separately. In Section 6.1 it was shown that the peak frequency of /d/ tended to be slighly lower on average than /t/, unlike what was found with /k g/ and /p b/. It is thus not difficult to imagine an attribute like AllPeakBark benefiting from the separation by voicing. However, AllPeakHz, which is also frequency-based, classifies slightly *worse* when /p t k/ and /b d g/ are separated, not better. Thus why AllPeakBark should benefit more from the separation by voicing than the other attributes is unclear.

As per 6.4.1, the mid-frequency attributes tend to perform relatively weakly: all six of them are found in the bottom half of the rankings. The high-frequency attributes, in contrast, are all found in the *top* half of the rankings. This striking difference between the attributes from the two frequency regions has remained remarkably consistent under all the conditions examined.

6.4.3 Prevocalic and Non-Prevocalic Plosives

Earlier it was noted that the 27 attributes have a mean classification accuracy of 55.6% for /b d g/ and 57.7% for /p t k/. Given that this difference is small, it was argued that the lumping together of voiced and voiceless into a single classification is relatively inconsequential.

We now turn to the following questions: (1) Does the classification accuracy vary depending on whether the consonant is followed by a vowel? (2) Is this effect larger or smaller than that for voicing? The answer to (1) is affirmative. The answer to (2) is that the mean classification accuracy of the 27 attributes on prevocalic tokens (N = 3,605) is 58.5%, whereas the mean classification accuracy on non-prevocalic tokens (N = 1,580) is 50.1%. This 8.4-percentage-point difference is far larger than the 2.1-percentage-point difference in the classification accuracy between voiced and voiceless plosives. Thus the acoustic change to the burst caused by the plosive being followed by a consonant (or pause) affects the information for distinguishing place of articulation far more than the acoustic change caused by the plosive being voiced. The mean classification accuracy of the 27 attributes when run on the entire dataset (N = 5,185) is 56.2%, which is more accurate than that of the non-prevocalic data subset (N = 1,580) by 6.1 percentage points. Only 284 of the 1,580 non-prevocalic tokens (18%) involve a following pause, so it is probably the effect of a following consonant that is driving

most of the degradation in accuracy. This tendency for the the acoustic information in the burst to be made less reliable by a following consonant may be one factor why many of the world's languages to varying degrees disfavour consonant clusters and/or prohibit consonants word-finally. Or to put it another way: all of the world's languages have plosive-vowel sequences whereas not all of them have plosive-consonant sequences (Ladefoged and Maddieson, 1996).

6.4.4 Self-Normalization

The classification results presented thus far have involved putting the raw acoustic measurements of the 19 speakers into a single classification without attempting to address differences in the acoustics between speakers. Such differences can arise from a variety of sources, e.g. one speaker (e.g. f02) may talk more loudly than another (e.g. f05), or a speaker (e.g. f03) may have a tendency to realize /t d/ as apical postalveolar, unlike most of the other speakers, or a speaker (e.g. m06) may have a greater tendency to prevoice /b d g/ than other speakers, with potential reduction of the burst's amplitude.

This section will present the results of two kinds of speaker-specific normalization (see 2.3.8 for further discussion). In one case, the speaker's mean value for an attribute is subtracted from the value of each token. As outlined in Section 2.3.8 of the literature review, this kind of normalization will be referred to as 'Norm':

$$\text{Norm} = x - \mu$$

where x represents an individual token in the dataset and μ represents the mean value for that attribute in the speech of an individual speaker.

In the second kind of normalization, the speaker's mean value for the attribute is again subtracted from the value of each token, but this time the further step is taken of dividing this by the speaker's *standard deviation* for that attribute. This is the same as Lobanov's (1971: 606) technique for normalizing the formant frequencies of vowels. We refer to this normalization as 'Lobanov-normalization' or 'Lobanov' for short:

$$\text{Lobanov} = \frac{x - \mu}{\sigma}$$

Here are the results for Norm:

Rank	Attribute	Accuracy (%)	Change from No Norm (% pts)	Change from No Norm (Rank)
1	CoGdB	67.1	+0.8	0
2	HiPeakdB	66.2	+1.6	+1
3	HiTotaldB	66.1	+2.2	+2
4	HiPeakPhon	65.0	+1.6	+2
5	HiTotalPhon	64.8	+1.8	+2
6	AllPeakHz	64.4	-0.5	-4
7	AllPeakBark	64.1	-0.2	-3
8	HiTotalSone	63.6	+0.9	0
9	HiPeakSone	63.3	+1.5	+1
10	CoGSone	60.3	+0.3	0
11	AllPeakPhon	59.9	+2.2	+1
12	CoGPhon	59.1	+0.6	-1
13	AllPeakSone	58.2	+1.7	0
14	SDSoneAmp	57.3	+1.3	0
15	SDPhonAmp	56.8	+1.4	0
16	MidPeakPhon	54.7	+1.9	0
17	MidPeakSone	54.5	+1.8	0
18	MidPeakdB	54.1	+2.5	+1
19	AllTotalSone	53.7	+2.2	+1
=20	AllTotalPhon	52.7	+2.1	+1
=20	SDdBFreq	52.7	0	-3
22	MidTotaldB	49.5	+1.1	+2
23	MidTotalSone	49.4	+0.2	-1
24	SDPhonFreq	47.7	0	+1
25	AllPeakdB	47.4	+0.2	0
26	MidTotalPhon	47.0	+2.0	+1
27	SDSoneFreq	39.5	-7.5	-1

Table 6.18: Discriminant analysis classification accuracy of the 27 attributes when the Norm normalization is applied relative to no normalization.

Change in classification accuracy relative to the no-normalization condition (Table 6.13) is shown in the second-from-right column, with the change in the relative rank of the attribute when normalized shown in the rightmost column. N = 5,185.

The overall improvement in classification accuracy is modest, and averages 0.9 percentage points for the 27 attributes. However, when SDPhonFreq and SDSoneFreq are excluded (which is justifiable given their high Wilks's Lambda scores), the improvement increases somewhat, to 1.2 percentage points. For three of the attributes the introduction of the speaker-specific normalization has resulted in a *decrease* in classification accuracy, and for two others the normalization has made no difference to the accuracy. Nevertheless, for 22 of the 27 attributes, the normalization has to some degree improved the classification accuracy, though for none does the improvement reach 3 percentage points.

In short, the effect of Norm on the classification accuracy is at best modest. Also, the improvements are not large enough to change the broad picture of which attributes are the strongest. Recall from Chapter 5 that the use of self-normalization of F2 (namely $F2_R - \mu F2_{individual}$) yielded an improvement in classification of 2.2 percentage points. Thus the improvement in classification accuracy yielded by normalization on the burst appears to be considerably smaller than the (itself modest) improvement in accuracy on the formants.

We turn now to the results for Lobanov normalization. Recall that this attribute normalizes the same as Norm except that the additional step of dividing by the standard deviation (of the speaker's attribute tokens) is added. The following table shows how much improvement the technique yields above and beyond what was already yielded by Norm. That is, the table shows how much improvement in classification is yielded for each attribute by dividing by the speaker's standard deviation σ:

Rank	Attribute	Accuracy (% pts)	Change from Norm Accuracy (% pts)	Change from Norm Rank
1	CoGdB	67.2	+0.1	0
2	HiPeakdB	66.6	+0.4	0
3	HiTotaldB	66.5	+0.4	0
4	HiPeakPhon	65.5	+0.5	0
5	HiTotalPhon	65.4	+0.6	0
6	HiTotalSone	64.6	+1.0	+2
7	AllPeakHz	64.4	0	-1
8	AllPeakBark	64.3	+0.2	-1
9	HiPeakSone	64.0	+0.7	0
10	CoGSone	60.4	+0.1	0
11	AllPeakPhon	60.3	+0.4	0
12	CoGPhon	59.1	0	0
13	AllPeakSone	58.7	+0.5	0
14	SDSoneAmp	58.3	+1.0	0
15	SDPhonAmp	57.2	+0.4	0
16	MidPeakPhon	54.5	-0.2	-1
=17	MidPeakdB	54.3	+0.2	0
=17	MidPeakSone	54.3	-0.2	-1
19	AllTotalSone	53.8	+0.1	-1
20	SDdBFreq	52.8	+0.1	+1
21	AllTotalPhon	52.7	0	-1
22	MidTotalSone	49.8	+0.4	+1
23	MidTotaldB	49.3	-0.2	-1
24	SDPhonFreq	48.7	+1.0	0
25	AllPeakdB	47.6	+0.2	0
26	MidTotalPhon	46.8	-0.2	0
27	SDSoneFreq	45.1	+5.6	0

Table 6.19: Discriminant analysis classification accuracy of the 27 attributes when the Lobanov normalization is compared to the Norm normalization.

Change in classification accuracy relative to the Norm condition (Table 6.18) is shown in the second-from-right column, with the change in the relative rank of the attribute when normalized shown in the rightmost column. N = 5,185.

With the exception of two attributes (HiTotalSone and SDSoneAmp, which improved by a percentage point), the addition of standard deviation to the normalization formula has not yielded a noticeable improvement in classification accuracy. Indeed, for three of the attributes it has resulted in a decline, and for three others it has made no difference. The average improvement in classification accuracy for the 27 Lobanov-normalized attributes over their Norm equivalents is just 0.49 percentage points, and if SDSoneFreq and SDPhonFreq are excluded this drops to 0.26 percentage points. This is less than a quarter of the size of the improvement between no normalization and Norm (1.2 percentage points). Thus by far the greatest amount of normalization improvement comes from subtracting the value of each data point from the speaker's mean value, not from taking the further step of dividing this figure by the speaker's standard deviation.

Nevertheless the difference in classification between the Lobanov-normalized and non-normalized variants is highly statistically significant ($p < 0.001$) for 14 of the 27 attributes, and is statistically significant at the $p < 0.01$ level for another two attributes (and reaches $p < 0.05$ for a further three attributes, though recall from 5.4.1 that the threshold for statistical significance for comparing two classifiers has been set at $p < 0.01$ in the present study). The attributes that show the greatest improvement are HiTotaldB (+2.6 percentage points, $p < 0.001$), AllPeakPhon (+2.6 percentage points, $p < 0.001$), MidPeakdB (+2.7 percentage points, $p < 0.001$), AllTotalSone (+2.3 percentage points, $p < 0.001$), and AllTotalPhon (+2.1 percentage points, $p < 0.001$)

In terms of attribute *groups*, the greatest improvement in mean classification accuracy is found for the six high-frequency attributes (+2.20 percentage points; $p < 0.001$ for all six attribute pairs), followed by five all-frequency amplitude attributes (+1.88 percentage points; $p < 0.001$ for four of the five) and the six mid-frequency attributes (+1.55 percentage points; $p < 0.001$ for two pairs, $p < 0.01$ for two other pairs, and not statistically significant for the other two pairs), The spectral moments' mean improvement in classification accuracy is 0.87 percentage points; however, in three of the six pairs the difference in classification before and after Lobanov normalization is not statistically significant, while in the remaining three cases the difference only reaches $p < 0.05$. Thus the tests of statistical significance suggest that the improvement to the classification accuracy of the spectral moment attributes is not significant.

Finally, the two frequency-based attributes (to wit, AllPeakHz and AllPeakBark) do not benefit from the Lobanova normalization, with a mean change in accuracy of -0.25 percentage points (neither result is statistically significant, i.e. it is best to treat these two cases as showing a lack of evidence of improvement rather than positive evidence of decline).

But in any event, perhaps the most important overall finding of this section is that neither the Norm nor the Lobanov normalization yields a dramatic improvement in the classification accuracy of attributes.

6.4.5 Spectral Tilt

What does spectral tilt consist of? Rather than using the high-frequency information as a cue on its own, spectral tilt measures the difference between the high-frequency information and the equivalent mid-frequency information. This has the potential to distinguish place of articulation better than either attribute can do on its own, since the absolute amplitude of a single peak or region is not just the product of place of articulation but of a host of other factors. For instance if one speaker speaks louder than another speaker, his or her values for HiPeakdB will be consistently higher than those of the other speaker. Spectral tilt, by gauging the amplitude of the high-frequency region relative to the mid-frequency region, may mitigate such non-place-of-articulation acoustic factors.

Previous research (e.g. Stevens et al., 1999; Suchato, 2004) has examined spectral tilt using the Hz-dB spectrum. The present study extends this research by comparing the performance of Hz-dB spectral tilt with that of Bark-phon and Bark-sone tilt. In addition, the performance of spectral tilt is compared when the input is the *peak* amplitude in a frequency region (e.g. HiPeakdB, HiPeakPhon, HiPeakSone) versus the *total* amplitude in a frequency region (e.g. HiTotaldB, HiTotalSone). Also, Suchato's Ahi-A23 tilt attribute will be replicated and compared with the current attributes. Finally, the performance of a new kind of spectral tilt attribute will be examined that calculates tilt using not just the amplitudes of the mid- and high-frequency peaks but also using the *frequency difference* between the peaks.

The attributes to be used in the different kinds of spectral tilt have already been introduced. They come in the following pairs:

1. TiltPeakdB = HiPeakdB – MidPeakdB
2. TiltPeakPhon = HiPeakPhon – MidPeakPhon
3. TiltPeakSone = HiPeakSone – MidPeakSone
4. TiltTotaldB = HiTotaldB – MidTotaldB
5. TiltTotalPhon = HiTotalPhon – MidTotalPhon
6. TiltTotalSone = HiTotalSone – MidTotalSone

For each attribute, Tilt consists of subtracting the latter member of each pair from the former.

In addition, this is Suchato's (2004: 44-45; 58-59) Ahi-A23dB attribute:

7. Ahi-A23dB = HiPeakdB – MidMeandB

MidMeandB, which has not been used in the attributes heretofore, is the mean amplitude of the mid-frequency region.

To put the results for the different kinds of tilt in context, we begin by recalling the results for the high-frequency attributes above (Figure 6.11), except that this time the /b d g/ and /p t k/ tokens have been run in separate classifications, with their results summed. This has been done because our inspection of the burst in Section 6.1 indicated that there were noticeable differences between /d/ and /t/: /d/ not only had a lower amplitude high-frequency region but also a noticeably higher-amplitude mid-frequency region. This suggests that the tilt attributes would run best with separate classifications for the voiced and voiceless series.

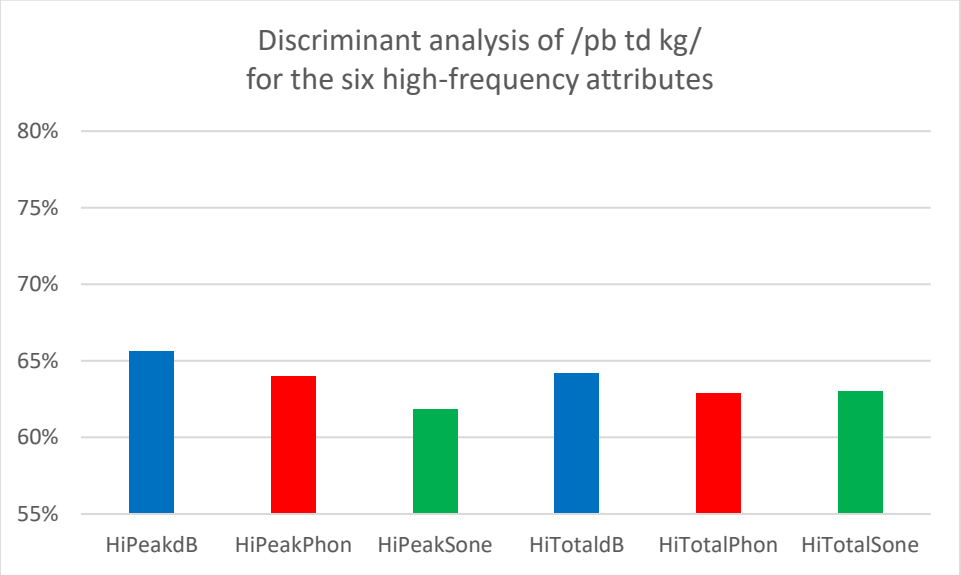


Figure 6.12: Discriminant analysis classification accuracy of the six high-frequency amplitude attributes. The first three attributes use the peak amplitude whereas the fourth and fifth measure the total amplitude. Note: /b d g/ and /p t k/ were run separately, with their results added together (and weighted according to their relative sample size). N = 5,185.

Thus the results in Figure 6.12 are similar to but somewhat higher than those in Figure 6.10 due to the separation by voicing. In terms of the 'peak' attributes, decibel outperforms phon (by 1.6 percentage points) and phon outperforms sone (by 2.2 percentage points). In terms of the 'total' attributes, decibel again outperforms phon (by 1.3 percentage points), with phon approximately the same as sone. The peak attributes outperform the total attributes, but only to a trivial degree (0.4 percentage points).

With this in mind, these are the results for the six kinds of spectral tilt:

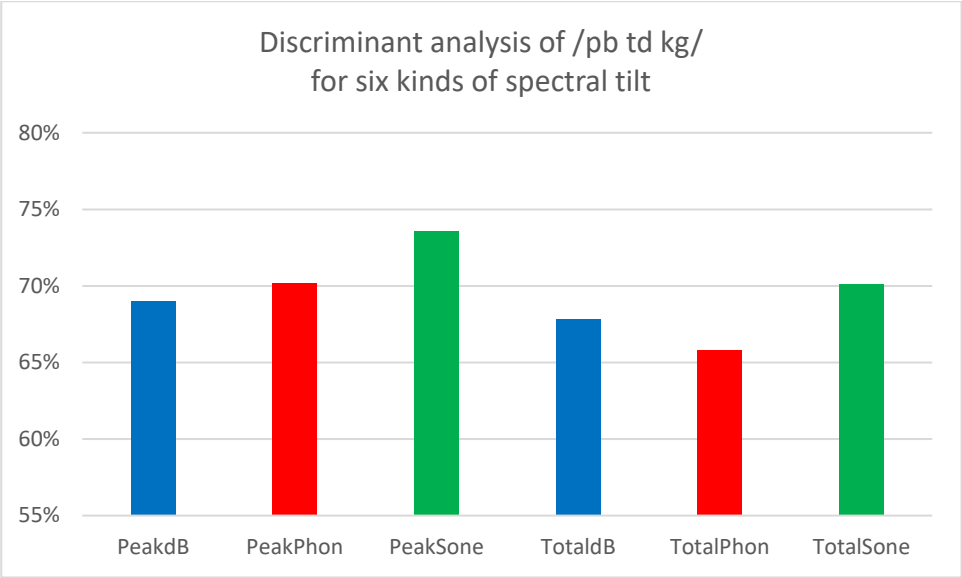


Figure 6.13: Discriminant analysis classification accuracy of the six kinds of spectral tilt.
The first three attributes use the peak amplitude whereas the final three use the total amplitude. Note: /b d g/ and /p t k/ were run separately, with their results added together (and weighted according to their relative sample size). N = 5,185.

The results for the 'peak' attributes are the reverse of what was found above: instead of Hz-dB being strongest it is now the weakest of the three, while sones has switched from being the weakest of the three to the strongest. Regarding the results for the 'total' attributes, the Bark-sone attribute has increased its margin over the Hz-dB attribute, from 0.7 percentage points to 2.0 percentage points.

Another way to think about these results is to ask: how much is the classification accuracy of the high-frequency attributes improved by subtracting the mid-frequency amplitude relative to *not* subtracting the mid-frequency amplitude? If looked at in this way, the results are much clearer:

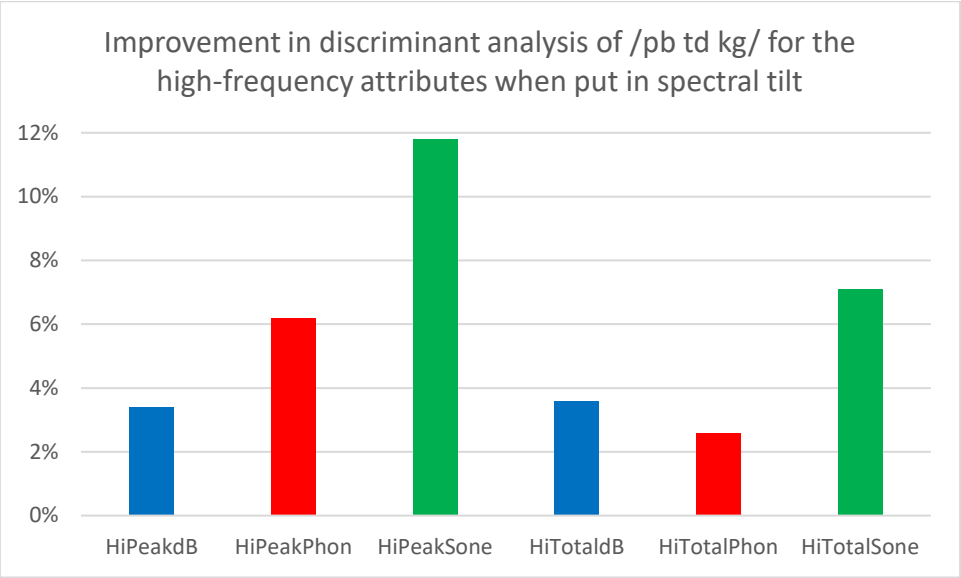


Figure 6.14: Increase in the discriminant analysis classification accuracy of the six kinds of spectral tilt relative to the equivalent high-frequency attributes.

The first three attributes use the peak amplitude whereas the second three use the total amplitude. Note: /b d g/ and /p t k/ were run separately, with their results added together (and weighted according to their relative sample size). N = 5,185.

The improvement for the two decibel attributes is 3.4 and 3.6 percentage points; for the phon attribute it is 6.2 and 2.6 percentage points; and for the two sone attributes it is 11.8 and 7.1 percentage points. It seems, then, that there is something about the scaling of amplitude on the sone scale that causes it to be superior for calculating spectral tilt.

Recall from Chapter 4 the scaling of the mean alveolar and bilabial spectra on the phon scale relative to the sone scale, reproduced below:

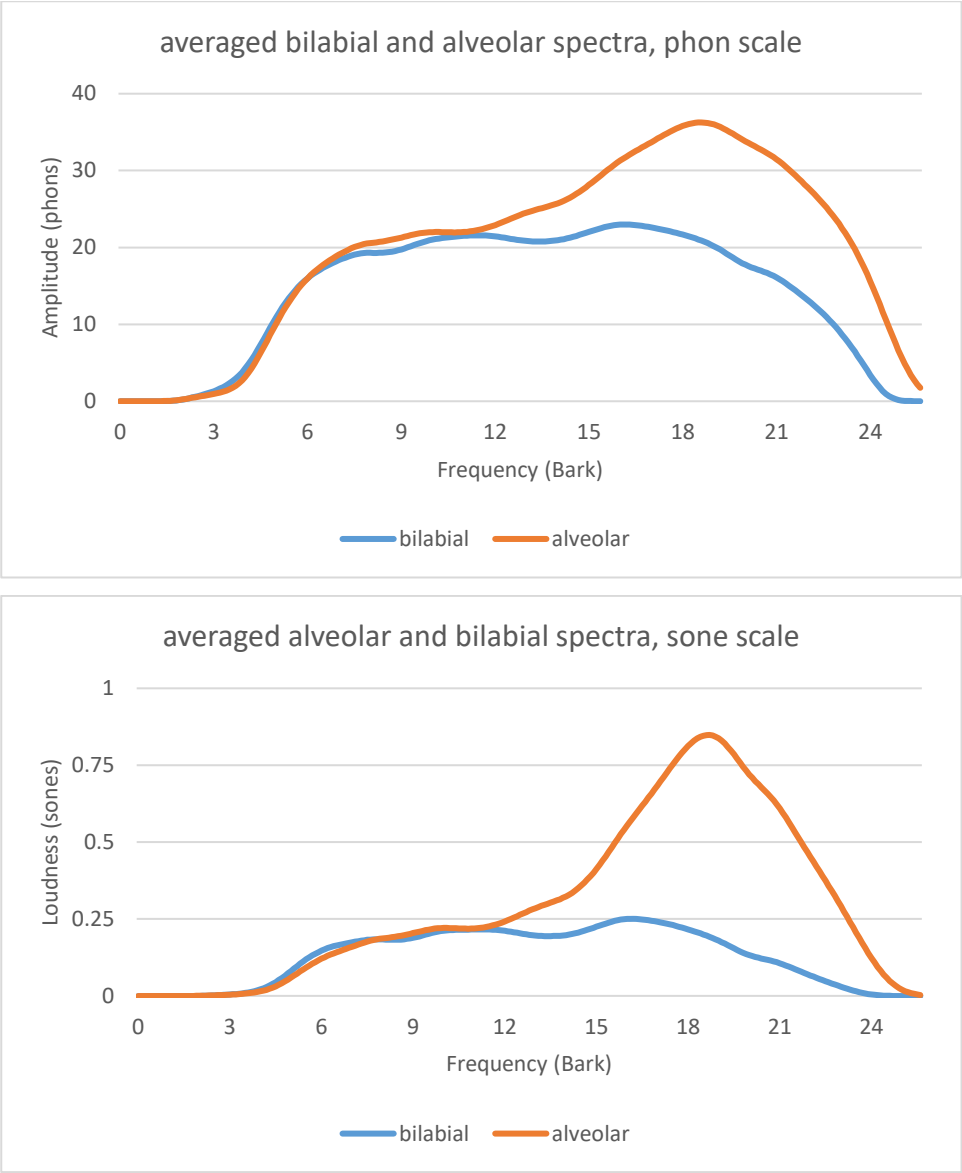


Figure 6.15: Comparison of the mean bilabial and alveolar burst spectra on the phon scale and sone scale. The sone scale makes the alveolar's high-frequency peak more prominent, and makes the bilabial's entire spectral envelope appear smaller, at least relative the alveolar one. Bilabial N = 1,490; alveolar N = 2,429.

The alveolar spectrum appears only moderately peakier than the bilabial spectrum on the phon scale, whereas on the sone scale the high frequency peak is far more prominent. Thus it is presumably the case that the sone scale enhances the spectral tilt of the alveolar spectrum (and probably also the velar spectrum – see Figure 6.2(a)) vis-à-vis the bilabial spectrum by making the spectral peak more prominent.

We now compare the present results with those for Suchato's (2004) attribute 'Ahi-A23 (dB)'. This attribute is very similar to the PeakdB tilt attribute. The only difference is that, rather than using the mid-frequency region's *peak* amplitude, Ahi-A23 instead uses the mid-frequency's *mean* amplitude (Suchato, 2004: 45). Its classification accuracy is 63.0%, which is

6.0 percentage points lower than that of TiltPeakdB and 2.8 percentage points lower than the weakest of the six tilt attributes presented in Figures 6.14 and 6.15.

Turning now to two further kinds of tilt attribute. In the 'peak' tilt attributes examined above, only the amplitude of the high- and mid-frequency peaks were utilized. The *frequency* of each peak in each frequency region was not noted, and one might wonder to what extent this introduces error into the calculation of spectral tilt, since when the precise distance in frequency between the two peaks is not noted, one is effectively assuming that the distance between the two peaks is the same in every token, which is of course not the case.

Thus we now examine a more elaborate version of the TiltPeakPhon and TiltPeakSone attributes: the frequency of the two peaks is entered into the calculation of the tilt. These attributes will be termed 'TiltPeakPhonPrecise' and 'TiltPeakSonePrecise', since the measure of tilt is more precise than for the PeakTilt attributes presented above. The distance between the two peaks is entered into the calculation of tilt by first subtracting the frequency of the high-frequency peak from the mid-frequency peak. This difference is then multiplied by the pre-existing tilt value:

$$\text{PeakTiltPrecise} = (\text{HiPeak[Bark]} - \text{MidPeak[Bark]}) \times \text{PeakTilt}$$

where 'PeakTilt' = (HiPeak[phon/sone] – MidPeak[phon/sone]).

The classification accuracy of TiltPeakPhonPrecise (under the same conditions as for TiltPeakPhon above) is 66.1%, which is 4.1 percentage points less than that of TiltPeakPhon. For TiltPeakSonePrecise the result is 68.7%, which is 4.9 percentage points less than that of TiltPeakSone. Thus we can say that including the exact location in frequency of the high-frequency and mid-frequency peaks *hurts* the classification accuracy of spectral tilt attributes. Sometimes, then, less information is better.

Finally, one might wonder to what extent the various tilt attributes are improved by Lobanov normalization. There are three logically possible ways of implementing the normalization: (1) the two input values are Lobanov-normalized first but the resulting tilt value is not; (2) the two input values are not Lobanov-normalized but the resulting tilt value is; (3) both the input values to the tilt and the tilt values themselves are Lobanov-normalized. We present the values for (1) and (2).

Here are the results for (1):

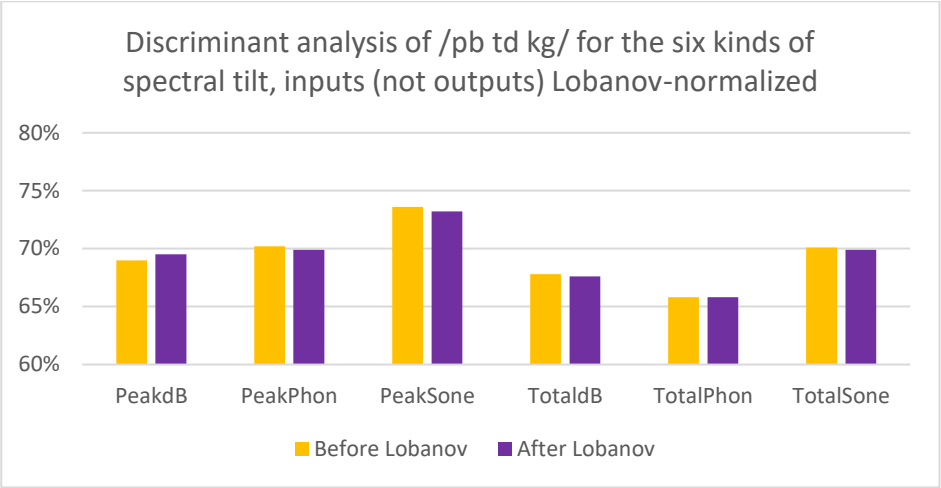


Figure 6.16: Discriminant analysis classification accuracy of the six kinds of spectral tilt, in which the inputs (i.e. high-frequency and mid-frequency amplitudes) have been Lobanov-normalized but the output (tilt) has not. The first three attributes use the peak amplitude whereas the second three use the total amplitude. Note: /b d g/ and /p t k/ were run separately, with their results added together (and weighted according to their relative sample size). N = 5,185.

The results look similar to those for the non-normalized equivalents, except that the difference in classification accuracy between HiPeakdB and HiPeakPhon has shrunk. However, the most significant change is that the classification accuracy for all but two of the attributes has worsened. Thus Lobanov-normalizing the *inputs* to the tilt does not help tilt attributes.

We turn now to the results for (2), in which it is the tilt's *output* that is Lobanov-normalized. Here are the results:

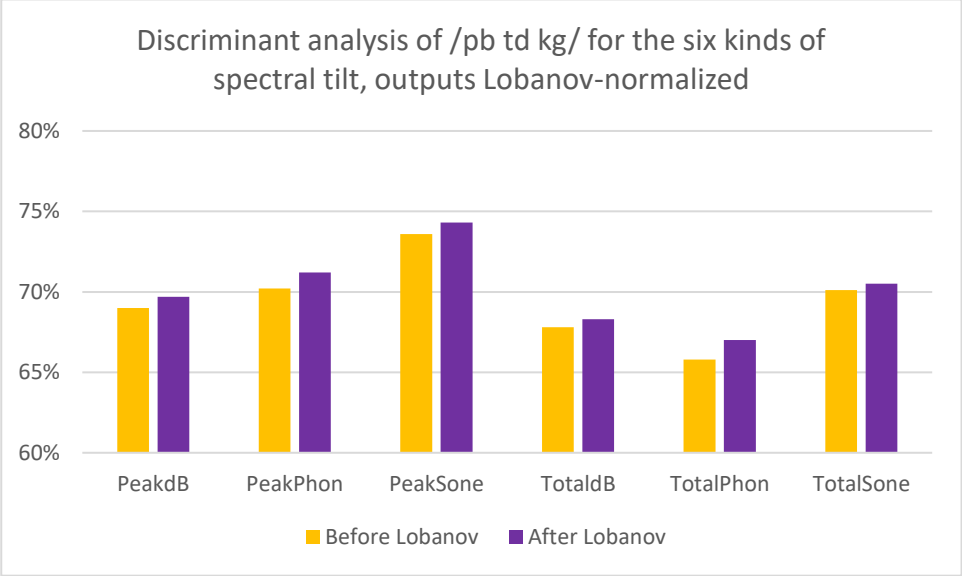


Figure 6.17: Discriminant analysis classification accuracy of the six kinds of spectral tilt, in which the inputs (i.e. high-frequency and mid-frequency amplitudes) have not been Lobanov-normalized but the output (tilt) has. The first three attributes use the peak amplitude whereas the second three use the total amplitude. Note: /b d g/ and /p t k/ were run separately, with their results added together (and weighted according to their relative sample size). N = 5,185.

At a glance we can see that the Lobanov normalization has not led to any switch in the relative performance of the six attributes. More noteworthy is that the classification accuracy of all six attributes has improved, unlike what happened with the first kind of Lobanov normalization in Figure 6.16. The classification accuracy has improved most for TotalPhon (+1.2 percentage points), least for TotalSone and TotaldB (+0.4 and +0.5 percentage points respectively), with PeakSone (+0.8 percentage points) and PeakdB and TotalSone (both +0.7 percentage points) intermediate. The classification accuracy of the best attribute, TiltPeakSone, has risen to 74.3%, which is 12.5 percentage points higher than HiPeakSone and 21.6 percentage points higher than MidPeakSone (the two attributes from which it was calculated).

It is also 5.2 percentage points more accurate than the strongest of the non-compound attributes, CoGdB. Another way to think about this is in terms of error rate: CoGdB classifies roughly 1 out of 3 tokens incorrectly (3.1 to be exact), whereas PeakSone classifies roughly 1 out of 4 tokens incorrectly (3.9 to be exact). Thus there are gains to be made in using compound attributes such as spectral tilt. Indeed five out of six of the tilt attributes presented here classify place of articulation more accurately than the best of the non-compound attributes.

The results for spectral tilt are the only case thus far in which the Bark-sone spectrum seems to be clearly superior to the other two spectra. What is curious about this finding is that when HiPeakSone and MidPeakSone were compared to their decibel and phon equivalents, they were consistently weaker in their classification accuracy than the attributes from the other two representations. However, when HiPeakSone and MidPeakSone were combined to form TiltPeakSone, the resulting attribute had greater classification accuracy than both the decibel and phon equivalents (and outperformed them by a greater margin than the other two representations had outperformed the sone attributes on the original singleton attributes). Thus the effect of the Bark-sone spectrum is subtle: if one's sole criterion for comparing the three spectral representations is singleton attributes, then the mean performance of Bark-sone attributes is slightly weaker. But if compound attributes such as spectral tilt are investigated, then the Bark-sone spectrum is stronger.

The contribution of the present work to the study of spectral tilt is that five further kinds of tilt have been examined. The previous work of Suchato (2004) used two kinds of tilt attributes, 'Ahi-A23(dB)' and 'Ehi-E23(dB)'. Our replication of 'Ahi-A23(dB)' showed that the attribute was weaker than the six attributes presented here. As for 'Ehi-E23(dB)', this has been termed 'TiltTotaldB' in the present work. As was shown in Figure 6.7 above, it is the second weakest of the present study's six tilt attributes. Thus four out of five of the new tilt attributes in the present work outperform Suchato's (2004) tilt attributes, which has been found

by investigating the Bark-phon and Bark-sone representations in addition to the more widely used Hz-dB representation.

6.4.6 Frequency Normalization

The normalization methods Norm and Lobanov, presented in 6.4.4, can be thought of as a kind of 'self-normalization': the data points for each attribute, instead of being considered in terms of their absolute values, are expressed relative to the attribute's mean and standard deviation for that attribute in the speech of a particular speaker. The present section investigates normalization methods in which the attribute is normalized by considering something *outside* of itself: the F2 frequency of the following vowel (henceforth $F2_{mid}$), or the mean F2 or F3 frequency for the specific speaker ($\mu F3_{individual}$ and $\mu F2_{individual}$). Because it is a token-intrinsic method of normalization, the use of the $F2_{mid}$ normalization is of course limited to those cases in which the plosive happens to be followed by a vowel as opposed to a consonant or pause. In contrast, the speaker-specific mean-formant normalization can in principle be applied to all plosives. For the sake of comparing it with the $F2_{mid}$ normalization, the initial presentation of the results will focus on prevocalic plosive tokens only.

What is the theoretical rationale for using $F2_{mid}$? In 6.2 we modelled the variation in the burst peak or centre of gravity and found that the peak or centre of gravity of all six plosives was higher before front vowels than back vowels. Given that front vowels have a higher F2 than back vowels, this means that subtracting $F2_{mid}$ from the burst peak or centre of gravity might reduce the amount of variation in the burst peak or centre of gravity caused by coarticulation with the following vowel.

The attributes that will be normalized using formant information come from the frequency domain (since formant information itself comes from the frequency domain). They are: AllPeakBark, AllPeakHz, CoGdB, CoGPhon, and CoGSone. As was shown in 6.4.1, these particular attributes are among the strongest of the 27 attributes that were examined, especially AllPeakBark, CoGdB, and AllPeakHz.

These are the results:

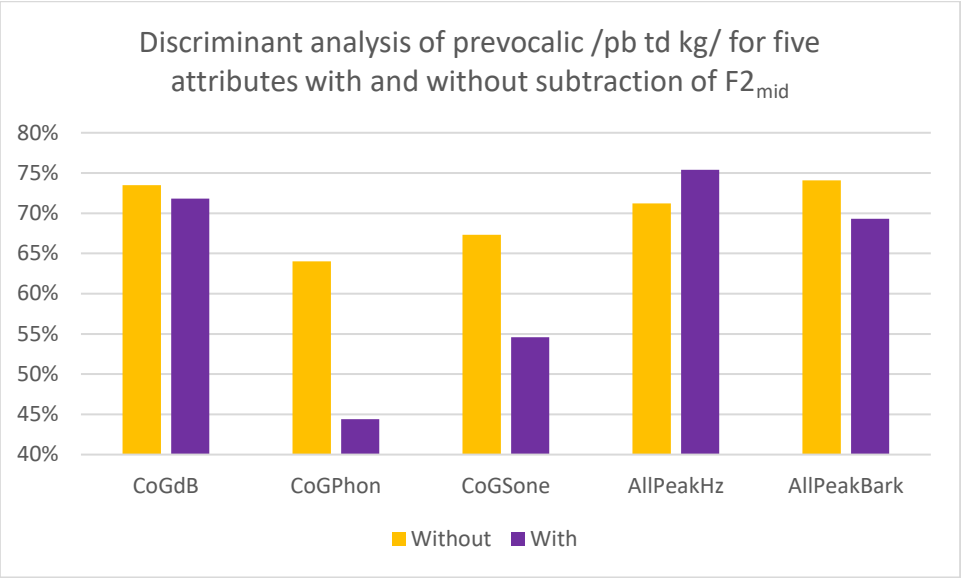


Figure 6.18: Discriminant analysis classification accuracy of the three all-frequency frequency-domain attributes, with versus without subtraction of the F2 frequency of the following vowel ($F2_{mid}$).
The data subset used is all plosives which both contain a release burst and are followed by a vowel (/p t k/ and /b d g/ were run separately, their results summed); N = 3,605.

Upon the subtraction of $F2_{mid}$, the classification accuracy of both AllPeakBark and the three kinds of CoG decreases. The decline for CoGdB and AllPeakBark is considerable (2.7 and 4.8 percentage points respectively) and for CoGPhon and CoGSone is disastrous (19.6 and 12.7 percentage points respectively). In contrast, for AllPeakHz the result is the reverse: the subtraction of $F2_{mid}$ *improves* the attribute, by 4.2 percentage points, and the attribute goes from being the third weakest to the strongest, with a classification accuracy of 75.4%.

The reason why the classification accuracy of AllPeakHz improves but not that of AllPeakBark nor Centre of Gravity is unclear. The subtraction of the vowel's F2 frequency was expected to improve the accuracy of all three kinds of attributes, based on the reasoning explored above and in Chapter 2: the burst of a velar consonant in particular is heavily affected by the F2 frequency of the following vowel; the higher the F2 frequency, the higher the frequency of the peak in the velar burst. To a lesser extent, the same effect is observed in alveolars, whose peak tends to be lower in frequency before back vowels. Thus it would appear that the AllPeakHz attribute somehow captures this coarticulatory effect better than the other two attributes, though this suggestion remains tentative.

In the above formula, the classification accuracy was the result of a simple subtraction of $F2_{mid}$ from the burst attribute. However, this assumes that the coarticulatory effect of the vowel's acoustics on the burst is of the same magnitude as the burst itself and so can simply be subtracted from the burst. But what if the influence of the vowel on the burst is *not* 1:1 as such

mathematics implies? If the amount of acoustic influence of the vowel on the burst attributes were moderated, would the classification accuracy improve?

To answer this question, the following procedure is developed: subtract the $F2_{mid}$ value from the burst attribute as before, but this time before subtracting it multiply it by a constant *c* (which varies the amount of acoustic influence of $F2_{mid}$ on the burst attribute and whose value ranges between 0 and 1). This is reminiscent of what was done in Chapter 5 to moderate the influence of $F2_{difference}$ on $F2_{onset}$ in our exploration of $F2_R$. We illustrate the formula for this idea using AllPeakBark but it will also be applied to AllPeakHz and CoG:

$$AllPeakBark_{new} = AllPeakBark - (F2_{mid} \times c)$$

where AllPeakBark denotes the original AllPeakBark value and $F2_{mid}$ denotes the F2 frequency in the middle of the vowel. As *c* is gradually increased from 0 to 1, the acoustic influence of $F2_{mid}$ on the burst attribute increases. The aim is to identify the value of *c* for which the classification accuracy of $AllPeakBark_{new}$ is highest. (If that value turns out to be 0, it indicates that factoring in the coarticulatory influence of the following vowel using $F2_{mid}$ *never* leads to an improvement in classification accuracy.)

Here are the results for CoG:

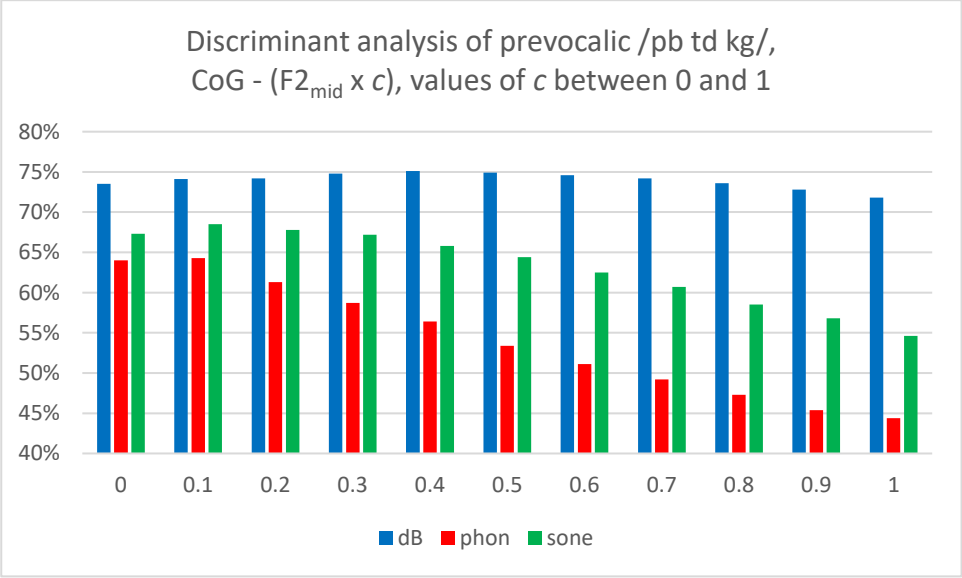


Figure 6.19: Discriminant analysis classification accuracy of Centre of Gravity, for a range of values of *c* between 0 and 1.

(The higher the value of *c*, the greater the influence of $F2_{mid}$ on the Centre of Gravity.) The data subset used is all plosives which both contain a release burst and are followed by a vowel; N = 3,605.

The classification accuracy of CoGHz when $c = 0$ (i.e. when the influence of $F2_{mid}$ is 0) is 73.5%. As the influence of $F2_{mid}$ is increased, the classification improves modestly, peaking at

75.1% for *c* = 0.4. For values of *c* greater than this, the classification begins to decline, and for *c* = 1 (the result for CoG – $F2_{mid}$ shown on the previous page) the classification accuracy has sunk so much that the attribute is 1.7 percentage points *worse* than if the influence of $F2_{mid}$ had not been utilized at all. In sum, $F2_{mid}$ does have the potential to improve CoGdB to a small degree but only if its influence is moderated by a carefully chosen value of *c*, around 0.4. Even with this value the improvement is arguably too small to warrant the use of the formula. The results for CoGPhon and CoGSone are still less encouraging: the classification accuracy does improve in these attributes if *c* is set to 0.1 or 0.2, but the improvement is so small (0.3 and 1.2 percentage points, respectively) that it is probably too small to be a true effect. Even if it were a true effect it is too small to justify the normalization.

Let us turn to the results for AllPeakBark:

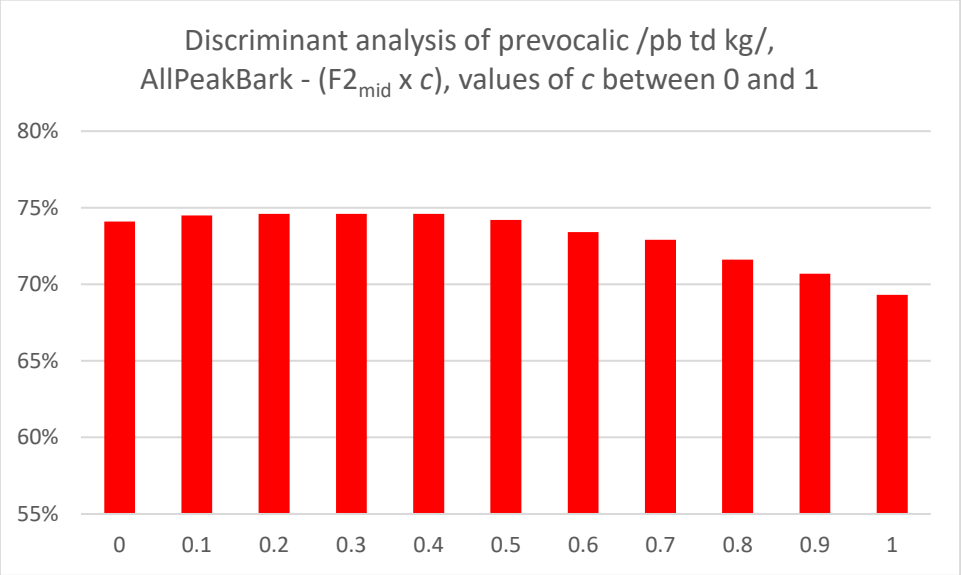


Figure 6.20: Discriminant analysis classification accuracy of AllPeakBark, for a range of values of *c* between 0 and 1. (The higher the value of *c*, the greater the influence of $F2_{mid}$ on AllPeakBark.) The data subset used is all plosives which both contain a release burst and are followed by a vowel; N = 3,605.

The classification accuracy when *c* is 0 (i.e. the influence of $F2_{mid}$ is 0) is 74.1%. As the influence of $F2_{mid}$ is increased, the classification accuracy improves marginally, peaking at 74.6% when *c* = 0.2, 0.3, or 0.4. When the influence of $F2_{mid}$ is increased any further, the classification accuracy begins to drop, reaching the 69.3% accuracy that was shown in Figure 6.18 above for *c* = 1. These results indicate that, no matter what value of *c* used, it is not possible to improve the classification accuracy of AllPeakBark using the F2 of the following vowel by more than 0.5 percentage points, which is too marginal an effect to justify the use of the normalization.

We finish with the results for AllPeakHz. For this attribute the effect of *c* has been examined over a larger region extending up to *c* = 1.5 (in order to be sure about the value of *c* for which its classification accuracy is strongest). Here are the results:

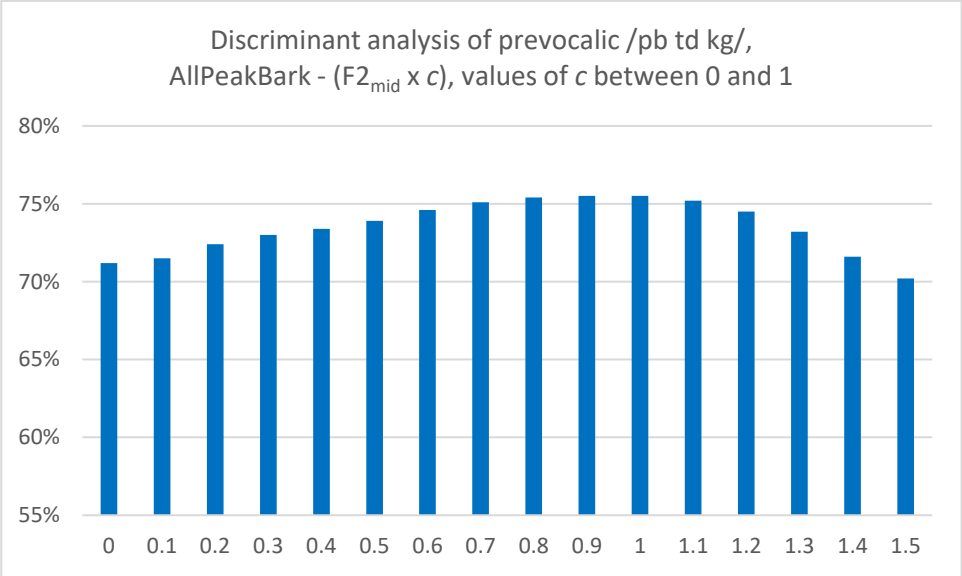


Figure 6.21: Discriminant analysis classification accuracy of AllPeakHz, for a range of values of *c* between 0 and 1. (The higher the value of *c*, the greater the influence of $F2_{mid}$ on AllPeakHz.) The data subset used is all plosives which both contain a release burst and are followed by a vowel; N = 3,605.

The results for AllPeakHz differ from those of AllPeakBark in that the factoring in of $F2_{mid}$ yields a noticeably larger improvement in classification accuracy. When *c* = 0, the classification accuracy is 71.2%, and rises to its peak of 75.5% for *c* = 0.9. This classification accuracy is almost identical to the accuracy when *c* = 1 (75.4%), the classification accuracy shown in Figure 6.18 earlier when *c* had not been entered. Thus the necessity of adding *c* for improving the classification accuracy is almost zero in the case of AllPeakHz, since the classification accuracy when *c* is not used turns out to be almost perfectly optimal anyway. Another important respect in which the results for AllPeakHz are different from those of the other two frequency-based attributes is that the improvement in the classification accuracy (5.3 percentage points) is much larger than for the other two attributes (1.6 percentage points for Centre of Gravity, 0.5 percentage points for AllPeakBark). Thus AllPeakHz seems to be the only attribute in any way capable of incorporating the coarticulatory effect of the following vowel to boost its own classification accuracy. Note also that the classification accuracy, 75.5%, is higher than the accuracy for any other attribute in this chapter.

That, then, is the influence of $F2_{mid}$ on three frequency-domain attributes. One limitation of $F2_{mid}$ is that it is only usable when the plosive happens to be followed by a vowel. We now turn to a formant-based normalization technique that is different in character: this involves the

speaker's *mean* F2 or F3 values. The theory underlying this approach is that while a listener is listening to a speaker, they are calculating ever more precise estimates of the speaker's mean formant values (see Section 5.4.2 for further discussion). The mean formant values tell the listener something about the speaker's vocal tract size: the larger the vocal tract, the lower the value.

But why gauge a speaker's release-burst properties relative to that speaker's mean formant values? The assumption is that speakers with mean formant values that are relatively low will also have burst centre of gravity (and burst peak) that are relatively low, assuming that the length of the vocal tract affects both acoustic events.

Here are the results of subtracting mean speaker F3 from the three frequency-domain attributes:

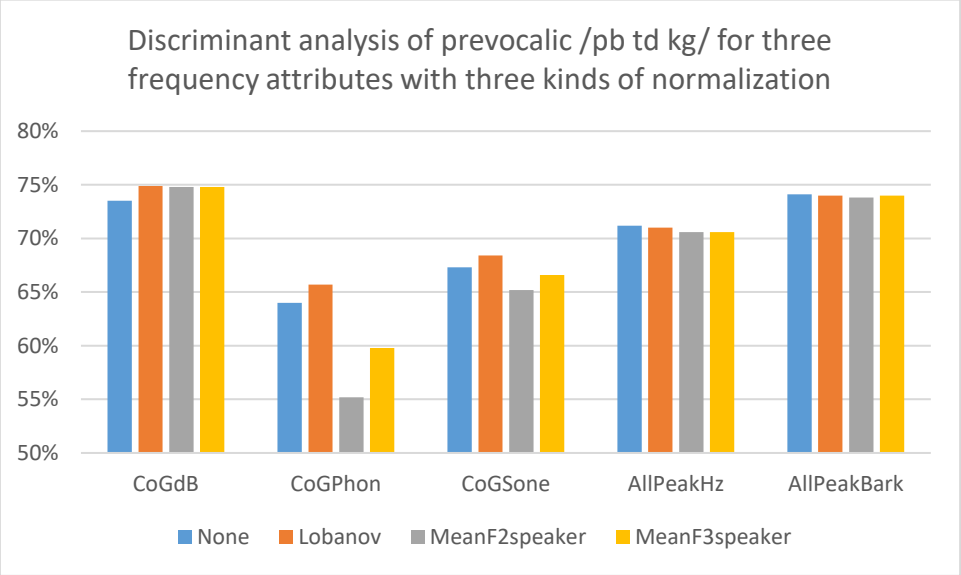


Figure 6.22: Discriminant analysis classification accuracy of the three frequency-domain attributes for three kinds of normalization, namely one kind of self-normalization (Lobanov) and two kinds of formant normalization (µF2speaker and µF3speaker).
The data subset used is all plosives which both contain a release burst and are followed by a vowel; N = 3,605.

The only one of the three attributes to show an improvement upon the incorporation of a mean formant frequency is Centre of Gravity, whose classification accuracy increases by 1.6 percentage points with Lobanov normalization, and by almost as much (1.3 percentage points) when either the speaker's mean F3 or F2 is subtracted from it. Thus for this attribute it makes little difference whether the normalization involves a speaker's mean vowel frequencies (formant normalization) or the mean of the burst attribute itself (self-normalization).

In the case of the other two attributes, AllPeakHz and AllPeakBark, the classification rate stays almost the same regardless of the normalization measure used. Indeed the classification *weakens* slightly, by between 0.1 and 0.6 percentage points. AllPeakHz, of course,

was the attribute out of the three that most improved when using $F2_{mid}$ as the normalizer (by 4.3 percentage points); it seems this improvement is not yielded by meanF2speaker and meanF3speaker, at least if the procedure used is one of simple subtraction rather than weighting the subtraction by the constant *c* that was used earlier.

With this in mind, we turn to the results of normalizing the above attributes when the influence of the speaker's mean F3 is moderated by the constant *c*. (This has not been performed on µF2speaker as the initial results in Figure 6.22 above show that µF2speaker and µF3speaker perform very similarly as normalizers.) Here are the results for the three kinds of CoG:

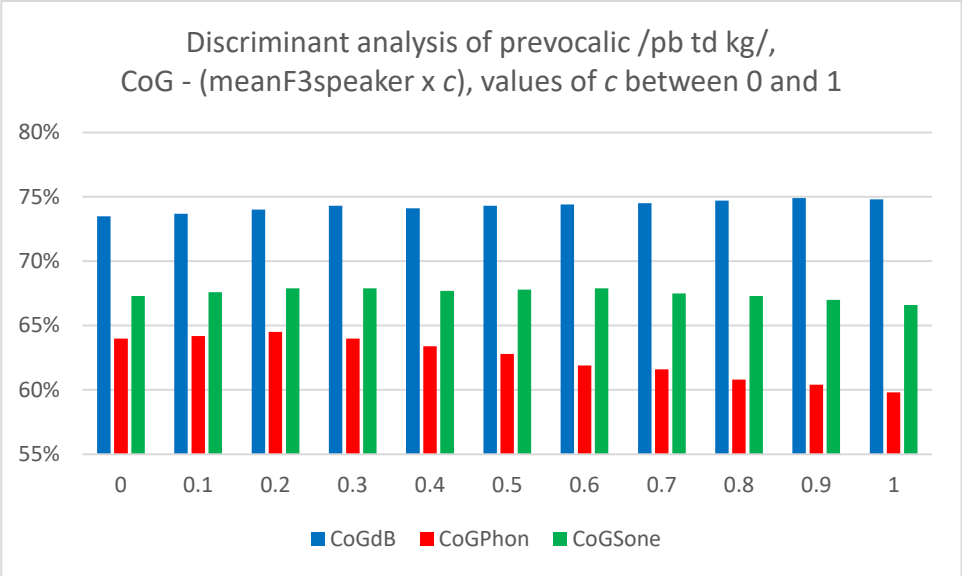


Figure 6.23: Discriminant analysis classification accuracy of Centre of Gravity (CoG) for a range of values of *c* between 0 and 1.
(The higher the value of *c*, the greater the influence of $F3_{mean}$ on Centre of Gravity.) The data subset used is all plosives which both contain a release burst and are followed by a vowel; N = 3,605.

For the Hz-dB version of CoG, the classification accuracy increases, albeit jaggedly, as the influence of µF3speaker is increased by *c*. Nevertheless the values of *c* greater than 0 and less than 1, with one exception (0.9), do *not* have a classification accuracy that exceeds that of *c* = 1 (which first appeared in Figure 6.22).

The results for the phon and sone versions of CoG are, in contrast, rather weak: although the classification accuracy does increase when the value of *c* is set to 0.1, the improvement is much too small to regard the normalizing as improving the attributes meaningfully, and when *c* is greater than 0.3 the classification accuracy for both attributes drops below the accuracy found when there is no normalization at all.

These are the results for AllPeakHz:

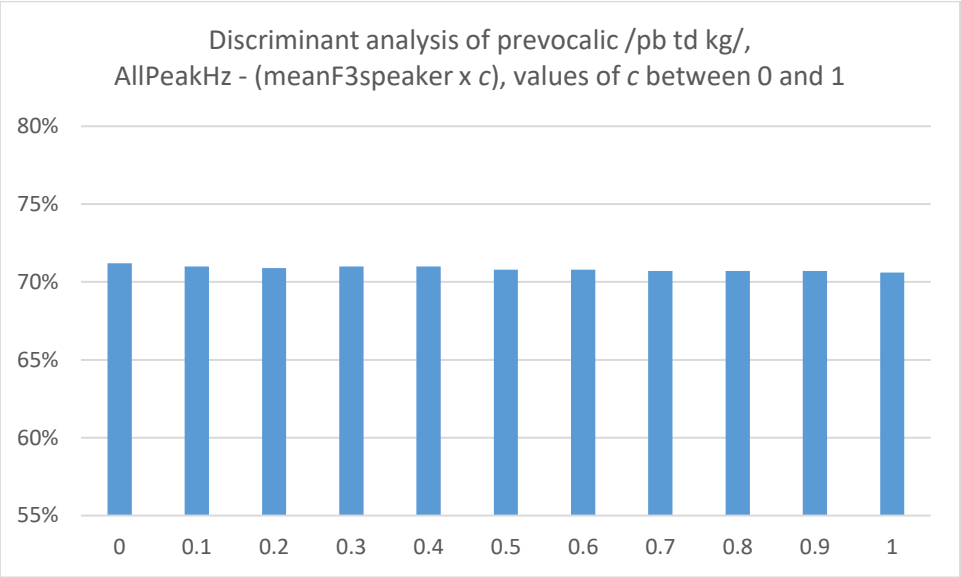


Figure 6.24: Discriminant analysis classification accuracy of AllPeakHz, for a range of values of *c* between 0 and 1. (The higher the value of *c*, the greater the influence of $F3_{mean}$ on the AllPeakHz.) The data subset used is all plosives which both contain a release burst and are followed by a vowel; N = 3,605.

The influence of speakerF3mean on this attribute is negative for all tested values of *c*. It seems that this attribute is incompatible with being normalized by a mean formant value.

These are the results for AllPeakBark:

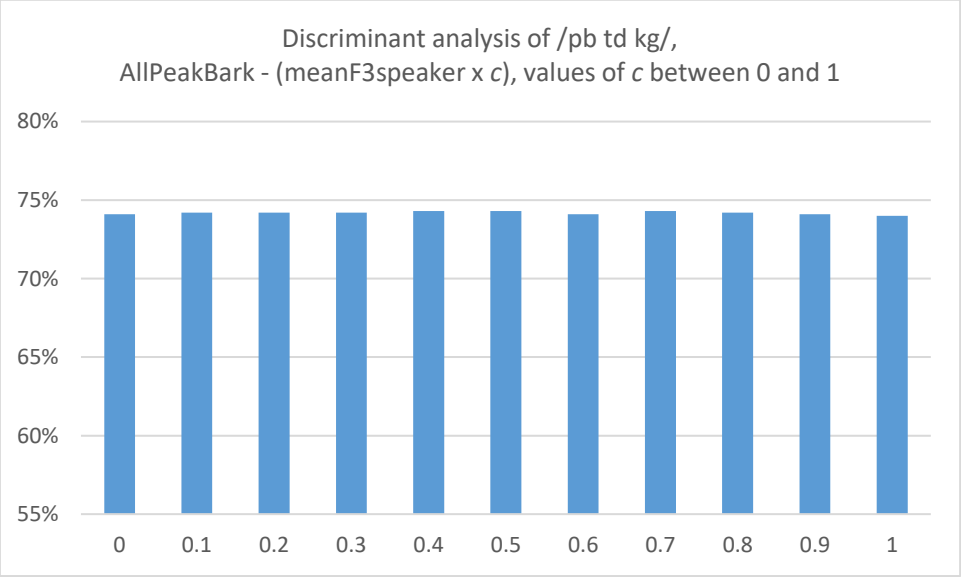


Figure 6.25: Discriminant analysis classification accuracy of AllPeakBark, for a range of values of *c* between 0 and 1. (The higher the value of *c*, the greater the influence of $F3_{mean}$ on the AllPeakBark.) The data subset used is all plosives which both contain a release burst and are followed by a vowel; N = 3,605.

Although some values of *c* do yield an improved classification accuracy over having no influence of µF3speaker at all, the gain in classification accuracy (from 74.1% for *c* = 0 to 74.3% for *c* = 0.4 or 0.5) is once again much too small to warrant using meanF3speaker as a

normalizer for this attribute. Thus the result for this attribute is essentially the same as that for AllPeakHz.

To summarize, the classification accuracy of AllPeakHz increased when it was $F2_{mid}$ that was subtracted from it; however, no normalization method involving mean values (whether Lobanov normalization, meanF2speaker or meanF3speaker) improved its classification accuracy. For CoGdB, $F2_{mid}$ did yield an improvement in classification (of 1.6 percentage points) but this modest improvement only appeared when the influence of $F2_{mid}$ was moderated by multiplying it by a constant *c* equal to 0.4 (when there was no constant, $F2_{mid}$ made the classification of Centre of Gravity worse). Lobanov-normalizing Centre of Gravity led to an identical improvement of CoGdB of 1.6 percentage points, while using meanF3speaker and meanF2speaker led to a smaller improvement in accuracy of 1.3 percentage points. Multiplying meanF3speaker by *c* did not reveal a value of *c* that resulted in anything beyond a minute change (0.1 percentage points) to the classification accuracy of CoGdB.

In brief, it seems that the most straightforward and beneficial means of improving the classification of CoGdB is to Lobanov-normalize it.

Finally, unlike the other two attributes the classification accuracy of AllPeakBark did not improve appreciably under any of the normalization conditions examined. Lobanov-normalizing and using either meanF2speaker or meanF3speaker all weakened its accuracy (albeit fractionally: 0.1 to 0.3 percentage points). Multiplying meanF3speaker by the constant *c* did reveal a value of *c* (0.4) that improved the classification by 0.2 percentage points, but this improvement is too trivial to justify using the normalization.

The strongest-performing attribute that we have seen is AllPeakHz-$F2_{mid}$ (c = 0.9), with a classification accuracy of 75.5%. Almost identical in accuracy is AllPeakHz-$F2_{mid}$ with no constant *c*, which yields 75.4% accuracy. A close second is the Lobanov-normalized variant of CoGdB with 75.1% accuracy. In third position is AllPeakBark with no normalization, which yields an accuracy of 74.1%. It should be borne in mind that, although these classification rates are relatively high, the testing ground is prevocalic tokens only (due to the requirement that $F2_{mid}$ be present).

An interesting difference between these three attributes lies in their performance on (prevocalic) /b d g/ relative to (prevocalic) /p t k/. On the /b d g/ tokens (N = 1,703), AllPeakBark achieves an accuracy of 78.3% whereas AllPeakHz-$F2_{mid}$ achieves 69.7% and CoG_Z 71.4%. On the /p t k/ tokens, however, it is AllPeakHz-$F2_{mid}$ and CoG_Z that dominate, with classification accuracies of 80.5% and 78.3% as compared to 70.2% for AllPeakBark (N = 1,902). It seems, then, that the performance of these attributes is broadly similar, with AllPeakBark strongest on voiced stops, AllPeakHz-$F2_{mid}$ and CoG_Z on voiceless. Furthermore,

given that the classification accuracies of the three attributes on the two contexts together (N = 3,605) are just 1.4 percentage points apart, it seems premature to see the present results as identifying one of the three attributes as being definitively better than the other two.

To summarize, the improvements in classification accuracy yielded by mean formant-frequency normalization (whether meanF2speaker, meanF3speaker, or standard scoring) are small in nature and appear only inconsistently. The only attribute to show substantial improvement using $F2_{mid}$ was AllPeakHz.

Overall, normalization – when it yields improvements at all – yields only modest improvements in classification accuracy. This has been a recurrent finding in this chapter.

6.4.7 Amplitude Normalization

The previous section considered normalization for frequency-domain attributes. We now turn to normalization for amplitude attributes. The normalization methods Norm and Lobanov which were presented in Section 6.4.4 can be thought of as a kind of 'self-normalization': the data points for each attribute, rather than being considered in terms of their absolute values, are instead expressed relative to the speaker's mean and standard deviation for that attribute. In common with the previous section, the present section investigates normalization methods in which the attribute is normalized using something *outside* of itself: the amplitude of the following vowel's second formant (termed $F2_{amp}$) or the amplitude of the following vowel's first formant (termed $F1_{amp}$).

The testing ground is the high-frequency peak attributes, namely HiPeakdB, HiPeakPhon, and HiPeakSone. These have been chosen because their classification accuracies were more accurate than those of the mid-frequency and all-frequency equivalents. That is, they are the amplitude-based attributes most worth attempting to improve. The names of the three F1-amplitude-normalized attributes are 'HiPeak-F1(dB)', 'HiPeak-F1(Phon)', and 'HiPeak-F1(Sone)'. Here are the results:

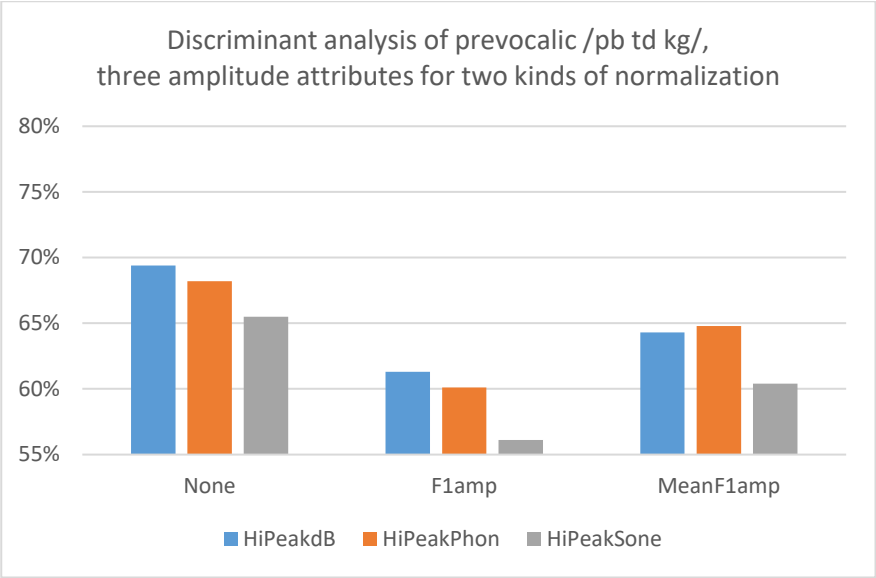


Figure 6.26: Discriminant analysis classification accuracy of the three high-frequency amplitude attributes, for two kinds of F1-amplitude normalization.

The data subset used is all plosives which both contain a release burst and are followed by a vowel. N = 3,518.

The classification accuracy of all three high-frequency attributes is reduced by the subtraction of the F1 amplitude of the following vowel ($F1_{amp}$), and this is true regardless of the amplitude scale used. The decrease is larger if the F1 amplitude of the particular vowel token is used ($F1_{amp}$) rather than the speaker's average F1 amplitude ($meanF1_{amp}$).

Clearly the mean F1 amplitude is a more promising (or rather, less damaging) means of normalization than the F1 of the particular vowel. Given this finding, an attempt to improve the performance of mean F1 amplitude will be made by regulating its degree of influence on the burst attribute by multiplying it by the constant *c*, whose value ranges from 0 to 1. This is analogous to what was done in the previous section with the frequency-based attributes. We begin with the results for the decibel variant, HiPeak-F1(dB):

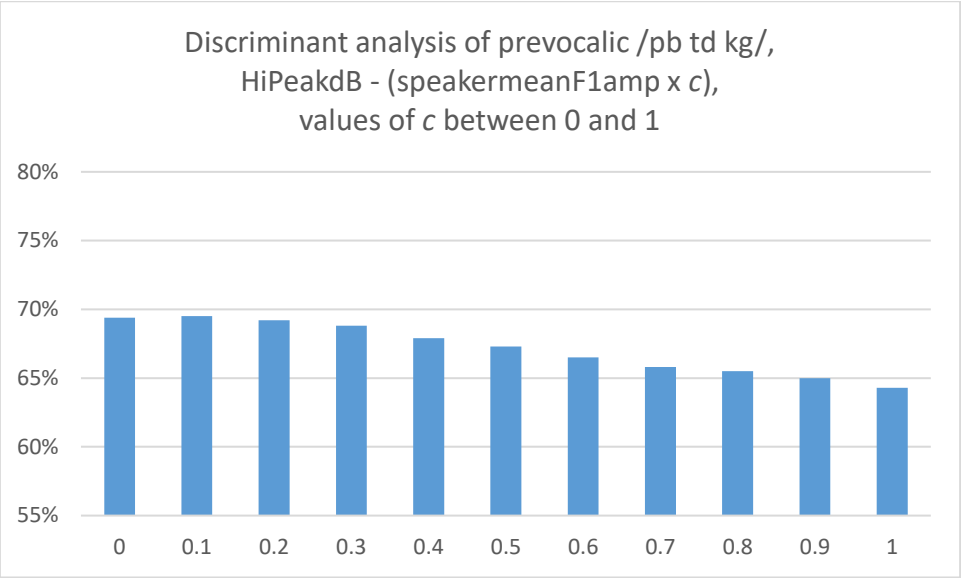


Figure 6.27: Discriminant analysis classification accuracy of HiPeak-F1(dB) for a range of values of *c* between 0 and 1.

(The higher the value, the greater the influence of $F1_{amp}$ on HiPeak-F1(dB).) The data subset used is all plosives which contain a release burst and are followed by a vowel with an extracted F1 amplitude. N = 3,518.

The classification accuracy improves when *c* is set to 0.1 but the improvement is minuscule (0.1 percentage points). All other non-zero values of *c* weaken the classification accuracy.

Here are the same results for the phon-scale equivalent:

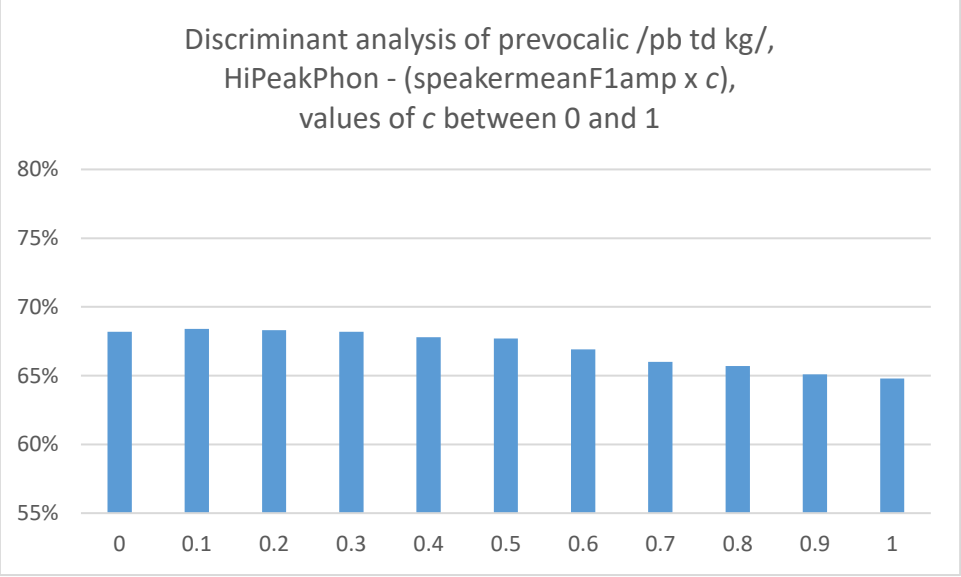


Figure 6.28: Discriminant analysis classification accuracy of HiPeak-F1(phon) for a range of values of *c* between 0 and 1.

(The higher the value, the greater the influence of $F1_{amp}$ on HiPeak-F1(phon)). The data subset used is all plosives which contain a release burst and are followed by a vowel with an extracted F1 amplitude. N = 3,518.

The classification accuracy again improves marginally when *c* is set to a low value (0.1 or 0.2) but for all other non-zero values of *c*, the result is a weakening of the classification accuracy.

And finally, the results for the sone-scale equivalent:

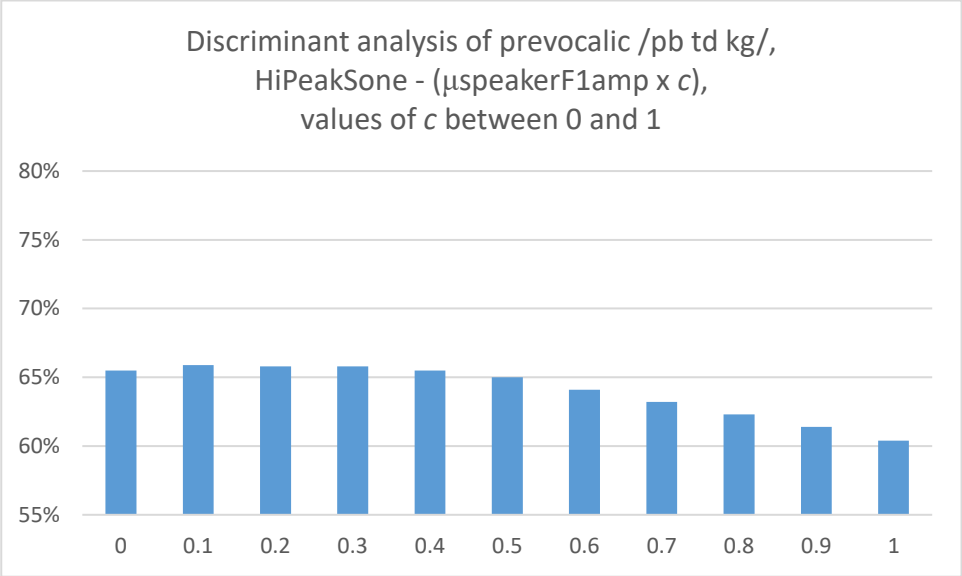


Figure 6.29: Discriminant analysis classification accuracy of HiPeak-F1(sone) for a range of values of *c* between 0 and 1.

(The higher the value, the greater the influence of $F1_{amp}$ on HiPeak-F1(sone).) The data subset used is all plosives which contain a release burst and are followed by a vowel with an extracted F1 amplitude; N = 3,518.

Once again the classification accuracy when *c* is small (in this case, between *c* = 0.1 and 0.4) improves marginally (by no more than 0.4 percentage points). For all values of *c* greater than 0.5, however, the classification accuracy is worse than when no F1 amplitude is included in the formula.

From the above three figures the following can be concluded: no matter what the amplitude scale, using a speaker's mean F1 vowel-onset amplitude as a normalizer of the burst's high-frequency amplitude has not led to an improvement in the attribute's ability to classify place of articulation correctly. The improvement to classification accuracy that does come when *c* is set to a low value is trivially small, far too small to justify the complexity involved in using such an acoustic attribute. Furthermore, when the above three high-frequency attributes are Lobanov-normalized, their classification accuracy improves by 2.3 percentage points for the decibel variant, 2.2 percentage points for the phon variant, and 2.7 percentage points for the sone variant. Thus Lobanov-normalization is superior to a normalization procedure involving the F1 amplitude.

Why might this be? Perhaps one factor is the fact that the vowel is voiced whereas the burst is voiceless. The structures used to generate voicing are in the larynx, whereas place of articulation is produced by articulators higher up, namely the tongue and lips. Because the two sounds are generated by largely independent structures, the mean amplitude of F1 in vowels could easily differ unpredictably from the mean burst amplitude of the same speaker. Also, the

laryngeal muscles can assume a variety of configurations that yield voice qualities with distinctive spectral properties (e.g. spectral tilt is shallower with relatively tense phonatory settings: Stevens, 1998: 99; Gobl and Ní Chasaide, 2010), with knock-on effects for mean F1 amplitude. The probability that all 19 speakers in the above data are assuming the same voice quality and the same loudness is minuscule. Furthermore, normalization by mean individual-speaker F1 amplitude is unlikely to succeed at resolving this complexity since voice quality varies considerably within an intonation phrase (part of a broader phenomenon known as declination), tending to be laxer towards the end of the phrase. In sum, mean F1 amplitude does not appear to be an appropriate method of amplitude-normalizing acoustic attributes. It seems that raw formant amplitudes are too different from the burst and too superficial a representation (confounding complicated source-filter interactions) to be an appropriate burst normalizer.

Although the present finding is negative, it is an important finding in that previous research such as Stevens et al. (1999) and Suchato (2004) have used $F1_{amp}$ in burst-based formulas, e.g. the attribute Av-Ahi (termed 'HiPeak-F1(dB)' in the present study) which subtracts the amplitude at the onset of the following vowel from the amplitude of the high-frequency peak in the burst. Suchato did not present the classification accuracy of Av-Ahi relative to Ahi on its own, which meant that there was no way of knowing whether and to what degree the inclusion of the F1 amplitude in the formula improved the attribute. The results of the present study suggests that employing formant amplitudes to enhance to classification accuracy of acoustic attributes is unprofitable.

6.5 Comparison of dB, Phon, and Sone Burst Attributes

In Section 6.4 the performance of 27 attributes was compared under a wide variety of conditions. With the exception of spectral tilt and the amplitude-based or frequency-based normalization techniques, the section was concerned with the performance of individual attributes. We now turn our attention to the performance of large combinations of attributes. For this purpose, the choice of statistic, as in Chapter 3, is random forests (see Section 3.1.7 for an overview).

In Section 6.4 the comparison of individual attributes from the decibel, phon, and sone spectra suggested that the attributes derived from the decibel spectrum performed stronger on average than those from the phon and sone spectra: in 6.4.1 the mean discriminant analysis classification of eight attributes from the dB spectrum was 57.6%, whereas the classification of the same attributes from the phon spectrum averaged 56.6%, and from the sone spectrum averaged 56.8%.

However, mean classification accuracy of attributes does not indicate how those attributes would perform when combined with each other.

The eight attributes taken from each of the three spectra are:

1. CoGdB/CoGPhon/CoGSone
2. SDdBFreq/SDPhonFreq/SDSoneFreq
3. AllPeakHz/AllPeakBark/AllPeakBark
4. AllPeakdB/AllPeakPhon/AllPeakSone
5. HiPeakdB/HiPeakPhon/HiPeakSone
6. HiTotaldB/HiTotalPhon/HiTotalSone
7. MidPeakdB/MidPeakPhon/MidPeakSone
8. MidTotaldB/MidTotalPhon/MidTotalSone

The results of Section 6.4 indicate that all attributes classified place of articulation above chance, 33.3%. Thus all of them will be used in the random forest. The random forest correlated reasonably closely with the data: $r^2 = 0.77$ for the 8 Hz-dB attributes, $r^2 = 0.78$ for the Bark-phon attributes, and $r^2 = 0.76$ for the Bark-sone attributes. The `predict` function (in the `party` package) was used to generate predictions on the out-of-bag (i.e. test) data based on the training data. That is, the algorithm first trains itself on two thirds of the data and sets aside the remaining third of the data for classification (for more information see James et al. (2013: 315-323) and Strobl et al. (2009)).

For the 8 attributes the overall percentage correct classification was 84.2% in the Hz-dB case, 85.5% in the Bark-phon case, and 84.0% for the Bark-sone case. Two things that are noteworthy about these figures: (1) they are similar to each other; (2) the spectrum with the lowest mean discriminant analysis individual-attribute accuracy – Bark-phon – turned out to yield the highest classification accuracy of the three attributes on the combined-attribute (random forest) test. Nevertheless the difference in accuracy is small, so too much importance should not be attributed to the difference. Instead, the fact that the ordering of the three spectra can flip on one statistic to another and that their accuracy remains close in all conditions suggests that no spectral representation is definitively superior. There appears to be no conclusive evidence that the choice of spectral representation affects the classification accuracy sufficiently to justify the choice of one spectral representation over another. As was shown in 6.4.1, the choice of frequency region in the burst (mid-frequency versus high-frequency) is a vastly more important factor in classification accuracy (13 percentage points).

We now present the results for the relative ordering of the eight attributes on each of the spectral representations, beginning with the Hz-dB representation:

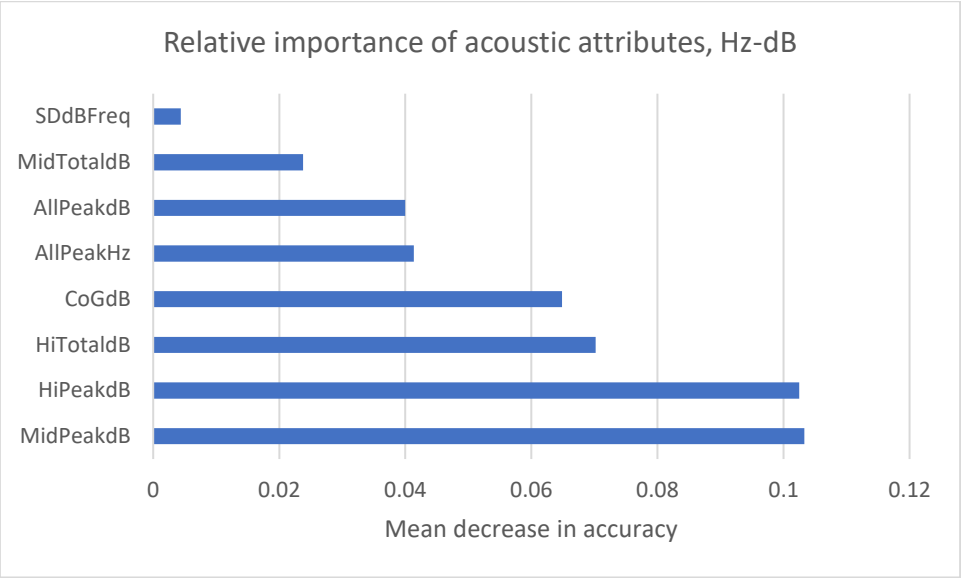


Figure 6.30: Relative importance of the acoustic attributes to the overall random forest classification accuracy for the eight Hz-dB attributes.
100 trees in forest.

MidPeakdB is the attribute that contributes the most to the classification, which is surprising given that all the mid-frequency attributes were found to be relatively weak on the discriminant analysis in Section 6.4. However, in the introductory discussion of random forests on 3.1.7, it was noted that the fact that random forests only allow the rounded square root of the total number of attributes to be considered in the construction of each tree, allows weaker predictors to have more of a voice in the decision of the forest rather than letting the best attribute dominate the forest by always being at the top of each decision tree, as would happen if each decision tree were allowed to choose from the full range of attributes. This may go some way to explaining why a weaker attribute like MidPeakdB contributes the most to the classification.

Nevertheless, three other attributes also contributed to nearly the same degree in the Hz-dB classification, two of which are from the high-frequency region. Recall that this region was found in the discriminant analysis results to yield many of the strongest single attributes, especially HiPeakdB.

Here are the equivalent results from the Bark-phon spectrum:

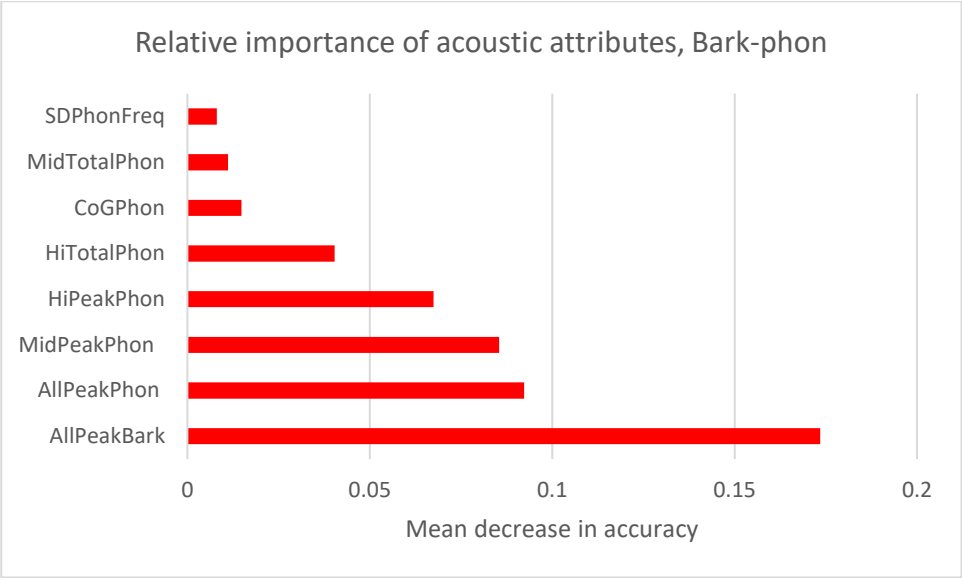


Figure 6.31: Relative importance of the acoustic attributes to the overall random forest classification accuracy for the eight Bark-phon attributes.
100 trees in forest.

Unlike the results for the Hz-dB attributes, there is one attribute that contributes to the Bark-phon classification far beyond any other, namely AllPeakBark. This attribute measures the frequency of the loudest component in the burst spectrum (the peak). The next most important attribute is AllPeakPhon, which also measures the peak but records its amplitude rather than its frequency. It was noted in the mixed-effects modelling of Section 6.3 that this attribute showed substantially lower values for bilabials than for velars and alveolars. AllPeakBark showed substantially higher values for alveolars than velars and bilabials. Together these two attributes could form the backbone of the three-way place contrast (further testing of this idea will be presented in Section 7.3). Nevertheless, MidPeakPhon shows almost as much a contribution as AllPeakPhon, so this suggestion remains tentative.

The results for the Bark-sone spectrum are shown in Figure 6.32. AllPeakBark is again by far the largest contributor to the classification. The ordering of AllPeakSone and MidPeakSone is the reverse of that of AllPeakPhon and MidPeakPhon, but in both classifications the importance of the two attributes is so close to each other that this is probably not a meaningful difference.

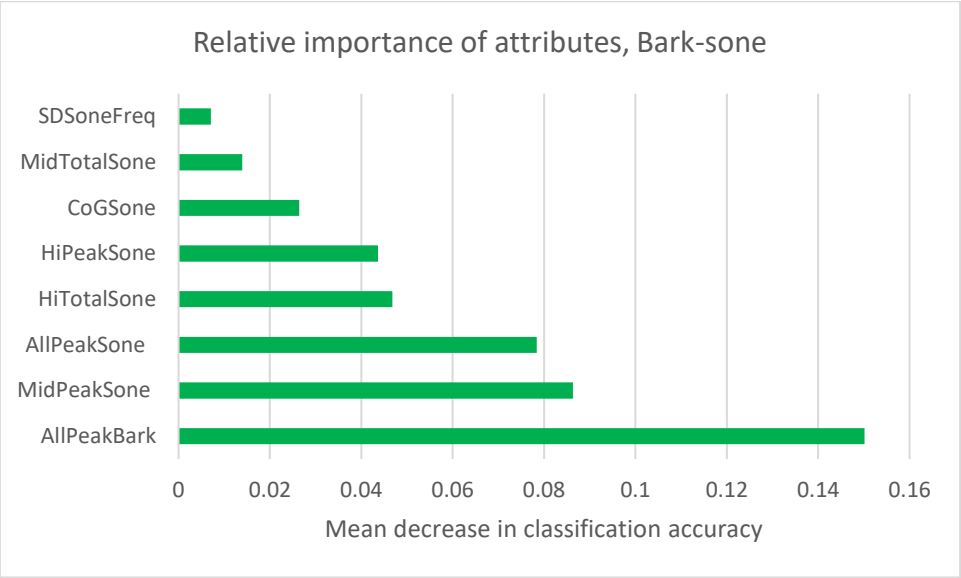


Figure 6.32: Relative importance of the acoustic attributes to the overall random forest classification accuracy, for the eight Bark-sone attributes.
100 trees in forest.

To summarize, on two out of three of the spectra it is AllPeakBark that turns out to be the strongest attribute by a considerable margin. The fact that this attribute turns out to be so important is not altogether surprising in light of the finding from the literature review that perceptual research (e.g. Li et al., 2010; Kapoor, 2010; Cvengros, 2011) has found the burst peak to be especially important for the perception of plosive place. The present findings, while being acoustic rather than perceptual in nature, point in a similar direction.

Nevertheless, the equivalent of AllPeakBark on the Hz-dB spectrum, AllPeakHz, did not turn out to be nearly as strong a predictor as AllPeakBark. If the frequency of the burst peak really is so important, why did AllPeakHz not also turn out to be important? The answer may have something to do with the fact that AllPeakHz does not warp the amplitudes of different frequency regions relative to each other in the manner done by the equal-loudness contours (unlike the Bark-phon and Bark-sone spectra). As was shown in the methodology chapter, this warping of the frequency regions represents the resonance properties of the outer and middle ear and has the effect of boosting frequency regions between around 2,000 to 5,000 Hz relative to higher and lower frequencies. Because the Hz-dB spectrum does not model this, it is possible that this could lead to a different spectral prominence becoming the 'peak' than in the Bark-phon and Bark-sone representations. This suggestion is, however, tentative (furthermore, recall that AllPeakHz and AllPeakBark showed extremely similar classification accuracies as single attributes in Section 6.4). Whatever turns out to be the true cause, it is sufficient for now to simply note that the frequency of the peak turned out to be the biggest contributor to classification accuracy on the two psychoacoustically-oriented spectral representations.

Random forests were also run (N = 100 trees) on the Lobanov-normalized variants of the eight attributes (presented in Section 6.4.4). The classification accuracy was 84.7% for the Hz-dB attributes, 85.2% for the Bark-phon attributes, and 84.2% for the Bark-sone attributes. These constitute a +0.5%, -0.3%, and +0.2% change on the non-normalized classifications, respectively. Thus the modest improvement that individual attributes showed in 6.4.4 when Lobanov-normalized (1.4 percentage points) effectively disappears when the attributes are combined in a random forest. Thus normalization by individual speaker has little or no effect on the combined classification of attributes derived from the burst spectrum.

This appears to suggest that differences between individual speakers in the realization of plosives bursts are either too small or too inconsistent to warrant normalization.

6.6 Summary

In this chapter, the performance of 27 burst-based acoustic attributes was examined over a variety of conditions. This examination included an investigation of the performance of a number of compound attributes, some based on amplitude (such as spectral tilt) and others based on frequency (such as the AllPeakBark-$F2_{mid}$).

The main findings of the present chapter are as follows:

- The choice of Hz-dB, Bark-phon, and Bark-sone spectra does not appear to be a particularly important factor in obtaining reliable classification of place of articulation. This has been seen in multiple respects: (1) attributes from the high-frequency region outperform those from the mid-frequency region on the discriminant analysis by an average of 13 percentage points, whereas the average classification accuracy of Hz-dB, Bark-phon, and Bark-sone attributes within these groups differs by a mere 1 percentage point; (2) the mean discriminant analysis accuracy of eight attributes from the three spectra were very similar: 57.6% for the Hz-dB spectrum, 56.6% for the Bark-phon, and 56.8% from the Bark-sone. Furthermore when these attributes were combined in a random forest, the resulting classification accuracies were 84.7%, 85.4%, and 84.2% respectively, again very similar.
- The only condition under which the Bark-sone spectrum appeared to yield (moderately) better attributes than the Hz-dB (and Bark-phon) spectra was in spectral tilt: TiltPeakSone classified at a rate of 73.6% as against 69.0% and 70.2% for the Hz-dB and Bark-phon equivalents.

- On the discriminant analysis, CoGdB and AllPeakBark were the strongest attributes. The former was strongest when voiced and voiceless put in a single classification, whereas the latter was strongest when they were classified separately.
- Normalization of attributes by individual speaker has surprisingly little effect on the classification accuracy of attributes. Lobanov-normalization yielded an average improvement in attribute classification accuracy of just 1.4 percentage points using discriminant analysis on individual attributes (though this difference did turn out to be statistically significant for most of the attributes). However, when the attributes were combined into a random forest, the Lobanov-normalized group classified almost the same rate as the non-normalized group: there was no more than a 0.5-percentage-point difference in the classification accuracy, and this was true for the Hz-dB, Bark-phon and Bark-sone spectra.
- Other methods of normalization (amplitude-normalization, frequency-normalization) showed even less potential at improving attribute accuracy. This appears to undermine the utility of certain compound attributes used in previous studies (e.g. Avi-Ahi: Stevens et al, 1999; Suchato, 2004). The results of the present study show that Ahi (= HiPeakdB) performs better *without* F1-based amplitude-normalization. Likewise CoG and AllPeakBark could not be improved much by subtracting $F2_{mid}$, and this was true even when the influence of $F2_{mid}$ on the burst was moderated using the constant *c*.
- A far more important factor affecting attribute accuracy than either normalization or the spectrum type is the choice of burst frequency region: attributes from the high-frequency region classify more accurately than analogous attributes derived from the mid-frequency region by a large margin: 13 percentage points.
- The mean classification accuracy of attributes on /p t k/ was 1.1 percentage points higher than on /b d g/. Splitting the bursts into separate classifications only led to a 0.5-percentage-point improvement in the mean classification accuracy of the attributes, though there were occasional exceptions (AllPeakBark +4.3 percentage points). It seems, then, that entering bursts from voiced and voiceless stops into the one classification has almost no detrimental effect on classification accuracy.
- In contrast, there was a considerable difference in the classification accuracy of prevocalic and non-prevocalic bursts: classification of non-prevocalic bursts was 8.4 percentage points less than prevocalic ones. This finding suggests that the acoustic changes brought on the burst by a following consonant affects its place-of-articulation information far more than voicing. This may also be one factor why the world's languages to varying degrees disfavour consonant clusters (i.e. there are human

languages where consonant clusters do not occur, e.g. Zulu (Harris, 1994: 162), whereas there are no human languages in which singleton consonants do not occur).

- When eight attributes were combined into a random forest, the attribute that was found to contribute most to the classification was AllPeakBark in the case of the Bark-phon and Bark-sone spectra. These results seem to suggest that the frequency of the burst peak is the most important piece of information in the burst. This finding dovetails with the perceptual findings about the importance of the burst peak presented in the literature review.

Chapter 7: Burst Features

The aims of the present chapter are threefold:

1. To examine how densely distributed place-of-articulation information is in the frequency domain of the burst.
2. To compare three types of feature for the release burst: (a) the traditional phonetic attributes presented in the previous chapter; (b) samples of the burst's amplitude at a selection of frequencies; (c) the first 12 coefficients of a discrete cosine transform of the burst.
3. To incorporate the information from the formants with the information from the burst so as to quantify its improvement to classification accuracy and model fit.

The layout of the chapter is as follows. Section 7.1 explains this chapter's aims in greater detail and provides a theoretical rationale for each aim. 7.2 examines the information density of the burst by gradually removing more and more of the 256 Bark-sone frequency channels and seeing what effect this has on classification accuracy. 7.3 compares the classification accuracy and model fit of the three attribute groups described in Aim 2 above. 7.4 furthers this comparison by adding information about the burst duration to the classification. 7.5 addresses Aim 3 above, namely the formant and other information from the preceding and following contexts is added to the classification to quantify how much it can improve the classification accuracy. 7.6 discusses the results of the chapter and concludes.

7.1 Overview and Theoretical Rationale

As stated above the aims of this chapter are threefold. Let us begin with the first aim of the chapter, which is to examine how dense the information in the frequency domain of the burst is for distinguishing place of articulation. This will be done by gradually removing more and more frequency channels and finding the point at which the classification accuracy of acoustic attributes begins to drop. The acoustic attributes used will be ones which rely on the entire spectrum, namely CoGSone + SDSoneAmp and AllPeakBark + AllPeakSone. The results of this procedure will be used to decide how many spectral samples to use in a comparison of three types of acoustic feature for the burst.

This leads us to Aim 2 of the present chapter, the primary aim, namely to compare the attributes used in the previous chapter with two alternative types of attribute, namely the spectral samples mentioned above and the first 12 coefficients of the discrete cosine transform (which was introduced in 2.3.1.9). The theoretical motivation for this comparison is as follows. In the literature review it was noted that phonetic science has been a long tradition of devising

attributes specially tailored to the release burst, including Halle et al.'s (1957) four-way split of spectral shape, Blumstein and Stevens' (1979) spectral templates, Lahiri et al.'s (1984) spectral tilt attribute, as well as more recent work by Stevens et al. (1999) and Suchato (2004). A selection of these tailor-made features has been tested in the previous chapter. An example of how such features are tailor-made is the fact that the boundaries of the high-frequency region (3.5-8 kHz) were defined to broadly coincide with the location of the alveolar burst peak and the boundaries for the mid-frequency region did the same for the velar peak (1.25-3 kHz). This style of research, in which the a priori knowledge of the researcher plays a strong role in the design of features, has been termed a knowledge-based approach (Abdelatty Ali et al., 2001; Suchato, 2004).

However, there seems to have been relatively little comparison in phonetics of how such knowledge-based features compare with certain alternatives. One alternative has been mentioned above: simply sample the amplitude of the burst at a variety of frequencies and use these as features (rather than predefining frequency regions such as the high- and mid-frequency regions). This approach serves as a benchmark against which to compare the tailor-made features, in the sense that taking spectral samples entails no predefining of frequency regions: a set of amplitudes is inputted to the classification, each corresponding to a different frequency; each amplitude can be thought of as a relatively raw acoustic attribute. This approach is a kind of 'null hypothesis' in the sense that no effort is made to process the spectrum beyond deciding the number of samples to take from it. Comparing this approach to tailored features allows one to see whether and to what extent having knowledge-based attributes improves classification accuracy. After all, it cannot be taken for granted that all these decades of designing handcrafted features in phonetics has led to features that are optimally performant.

The third group of features used in the present chapter are the first 12 coefficients of the discrete cosine transform (DCT). Recall from the literature review that this feature set underpins mel-frequency cepstral coefficients (MFCC), a front end widely used in automatic speech recognition. In that chapter the resemblance between the cycles (coefficients) of the DCT and the first four components of a principal component analysis was noted:

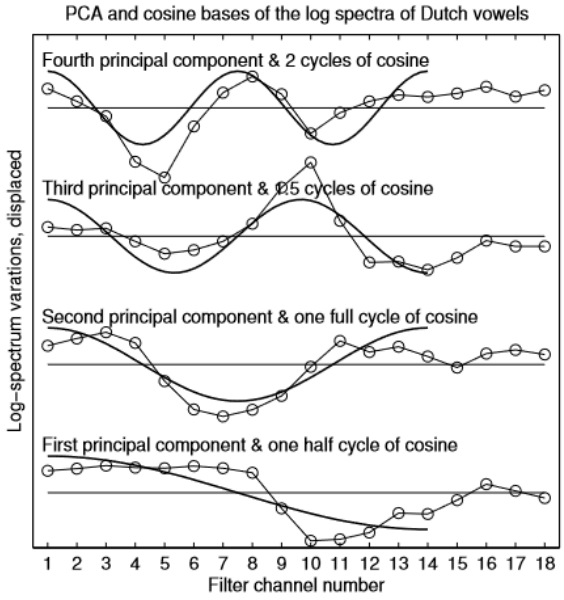


Figure 7.1: Comparison of the shape of the first four cycles of the DCT with the first four components of a principal component analysis (PCA).
Note the similarity of the two shapes. PCA from Plomp et al. (1967); diagram from Lyon (2017: 77).

In Figure 7.1 we see that the basis set used in the DCT is similar to that used in PCA. The transformation involved in the PCA, of course, is deliberately designed to make each variable account for as much of the variance in the data as possible while being orthogonal to the preceding variables. Thus given the similarity of the DCT basis set to PCA components, we expect the DCT coefficients to be strong at capturing the variance in the release burst. Whether and to what degree these features perform better than the spectral samples and the features used in the previous chapter is an empirical question.

A further aim of this chapter (Section 7.3) is to quantify the improvement to classification accuracy and model fit yielded when information about the burst's duration is included. Previous studies (e.g. Adelatty Ali et al. 2001, Suchato 2004) do not appear to have used burst duration as a feature. Inspection of spectrograms whilst annotating suggested that the duration should be shortest on average for bilabials, longer for velars and alveolars.

The final aim of this chapter is to combine the attributes from the burst with the attributes from the formants investigated in Chapter 5. The features from the preceding and following contexts will indicate (1) whether the segment is a vowel; (2) whether the vowel is front or back (or schwa); (3) F3 frequency at onset or offset; (4) $F2_R$.

7.2 The Information Density of the Burst

Each Bark-sone spectral slice consists of 256 channels, one every 0.1 Bark.[9] One might wonder to what extent having so many channels is necessary. To this end, we examine to what extent sparsifying the spectrum affects the classification accuracy of four attributes: CoGSone, SDSoneAmp, AllPeakBark, and AllPeakSone. The stages of sparsification are as follows: first, one out of every three channels will be deleted, which reduces the total number of channels from 256 to 171. For example, the channel at 0.2 Bark will be deleted, followed by the ones at 0.5, 0.8 Bark, and so on up the envelope.

Such sparsification is then successively increased, from 1/3 to 2/3 to 5/6 to 8/9 to 11/12 to 14/15 to 17/18 and so on. The point at which the sparsification begins to affect the classification accuracy appreciably is the point at which the pruning of the envelope genuinely causes information loss.

The first batch of results are for AllPeakBark:

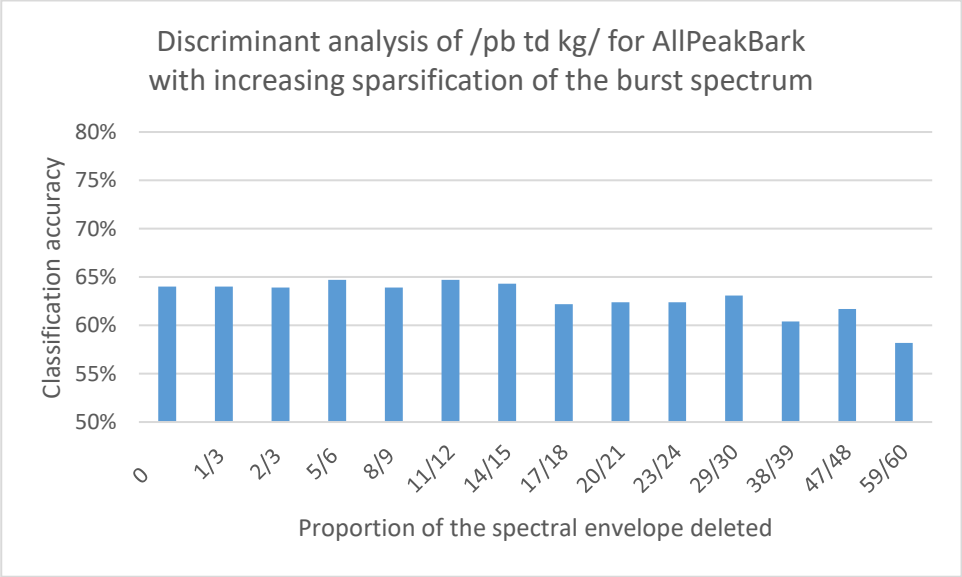


Figure 7.2: Discriminant analysis classification accuracy for the acoustic attribute AllPeakBark over a variety of spectral sparsification conditions.
Voiced and voiceless entered into the same classification, N = 5,471.

The effect of sparsification on AllPeakBark seems to be inconsistent. For small degrees of sparsification (e.g. 2/3 and 5/6) there is almost no change in classification accuracy, whereas with greater sparsification the attribute's performance seems to fluctuate unpredictably from one condition to the next. Nevertheless there does seem to be a marked reduction in classification accuracy under the three most extreme conditions tested, namely 38/39, 47/48,

[9] This is the default setting in Praat and was used in the present study. Praat allows the number of channels to be changed.

and 59/60. A smaller reduction (ca. 2 percentage points) can be seen for the sparsifications from 17/18 onwards. Under such conditions the difference in frequency between each remaining spectral component is over 3 Bark.

In terms of statistical significance, a McNemar test was run between the original (left-most) condition and each of the sparsified conditions. This revealed that the change in the classification was not statistically significant for all degrees of sparsification less than or equal to 8/9 (and for the 14/15 sparsification). For the 11/12 sparsification, the improvement in classification was statistically significant. However, it was significant at the $p < 0.05$ level, and as explained in Chapter 5 (5.4.1) the statistical significance should be set to $p < 0.01$ under the current conditions due to the large number of statistical tests increasing the risk of a Type 1 error.

For all sparsifications equal to or greater than 17/18, the reduction in classification accuracy is highly statistically significant ($p < 0.001$).

To summarize, the classification accuracy of AllPeakBark is robust so long as the sparsification is not greater than 14/15. This is impressive in the sense that the 14/15 sparsification means that over 93% of the original channels have been removed.

Here are the results for AllPeakSone:

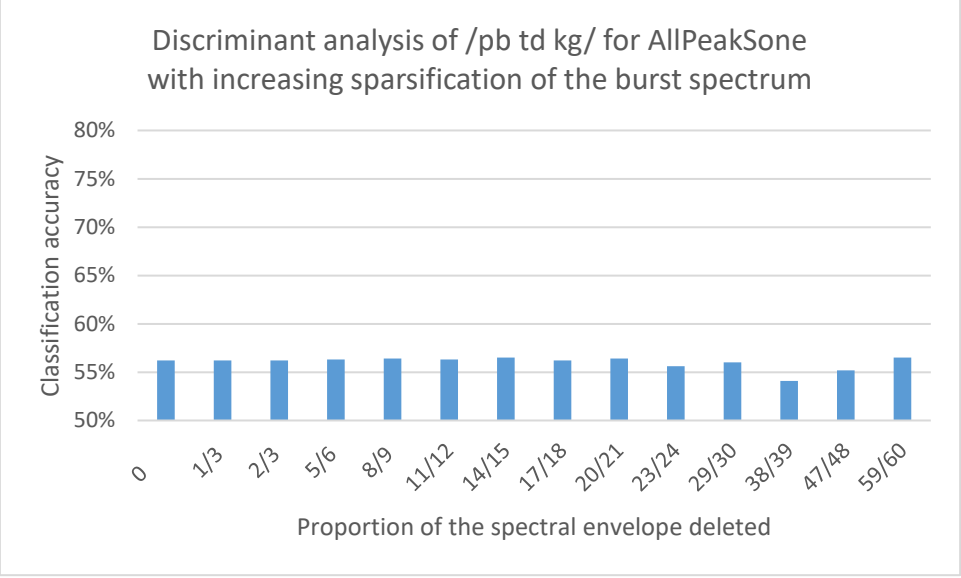


Figure 7.3: Discriminant analysis classification accuracy for the acoustic attribute AllPeakSone over a variety of spectral sparsification conditions.
Voiced and voiceless entered into the same classification, N = 5,471.

The classification accuracy of AllPeakSone, though lower on average than AllPeakBark, fluctuates less with increasing spectral sparsification: the accuracy remains largely intact up as far as 38/39 sparsification. This reduced fluctuation is reflected in the McNemar tests: none of the above sparsifications yields a classification that is statistically significantly different from

the unsparsified condition. Thus AllPeakSone seems to be even more robust to sparsification than AllPeakBark. This robustness is remarkable when one bears in mind how much of the original spectrum is deleted in the maximum sparsification condition: over 98%.

Here are the results when AllPeakBark and AllPeakSone are used together:

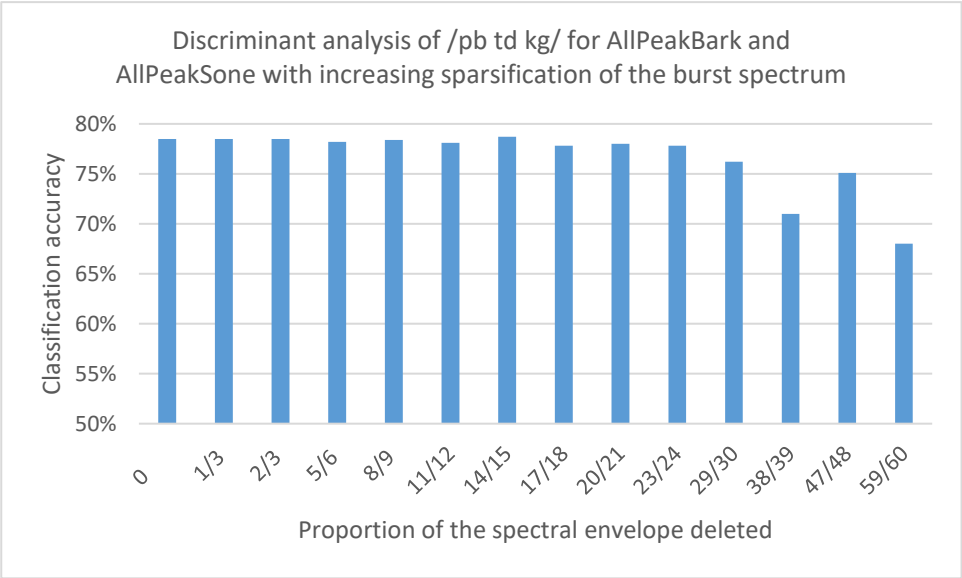


Figure 7.4: Discriminant analysis classification accuracy for the acoustic attributes AllPeakBark and AllPeakSone over a variety of spectral sparsification conditions.
Voiced and voiceless entered into the same classification, N = 5,471.

The sparsification has remarkably little effect on the classification accuracy of the acoustic attributes over most of the conditions examined. Indeed the classification accuracy remains almost the same for all sparsifications equal to or less than 23/24. This is quite surprising when it is remembered what a sparsification of 23/24 represents: almost 96% of the original spectrum's channels have been deleted, and yet the combined classification accuracy of the two attributes has dropped by just 0.7 percentage points. This 23/24 sparsification consists of a spectrum with just 11 out of the 256 channels remaining.

All the classifications that involve sparsification of 23/24 or less are not statistically significantly different from the non-sparsified classification (where statistical significance, recall is defined as $p < 0.01$). All sparsifications above this value have a lower classification than the non-sparsified sparsification that is highly statistically significant ($p < 0.001$ in all four cases).

We now turn to the results for centre of gravity (CoGSone) and standard deviation (SDSoneAmp). Recall that CoGSone is a frequency-based attribute and so has the same general function as AllPeakBark, whereas SDSoneAmp is an amplitude-based feature and is thus comparable to AllPeakSone. Here are the results for CoGSone:

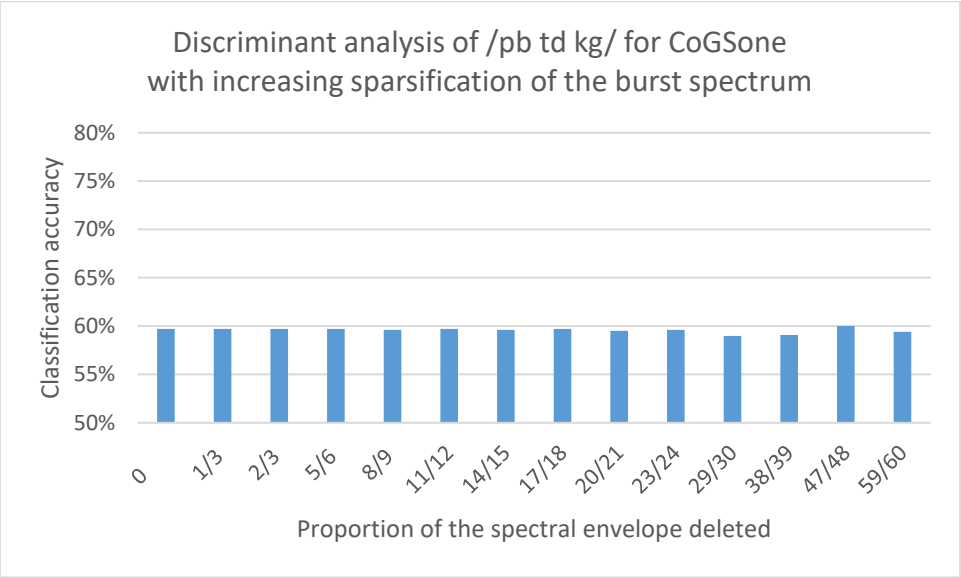


Figure 7.5: Discriminant analysis classification accuracy for the acoustic attribute centre of gravity (CoGSone) over a variety of spectral sparsification conditions.
Voiced and voiceless entered into the same classification, N = 5,471.

The attribute is remarkably robust to spectral sparsification. It is more robust than AllPeakBark, which tended to fluctuate unpredictably from one sparsification condition to the next (Figure 7.2). Even under the most extreme degree of sparsification (59/60, in which there is a spectral component only once every 6 Bark) the classification accuracy remains relatively strong (59.4%) and is stronger than AllPeakBark under the same condition (58.2%). (Nevertheless, final judgment on the robustness of whole-spectrum attributes should be reserved until CoGSone is combined with SDSoneAmp and compared to the combined accuracy of AllPeakBark and AllPeakSone, which will be done shortly but also in Section 7.3 using random forests.)

This robustness is also found in the tests of statistical significance: all sparsifications shown in Figure 7.5 above are not statistically significantly different to the non-sparsified condition. The 29/30 sparsification does, however, reach $p < 0.05$ but recall that statistical significance has been set at $p < 0.01$.

Here are the results for SDSoneAmp:

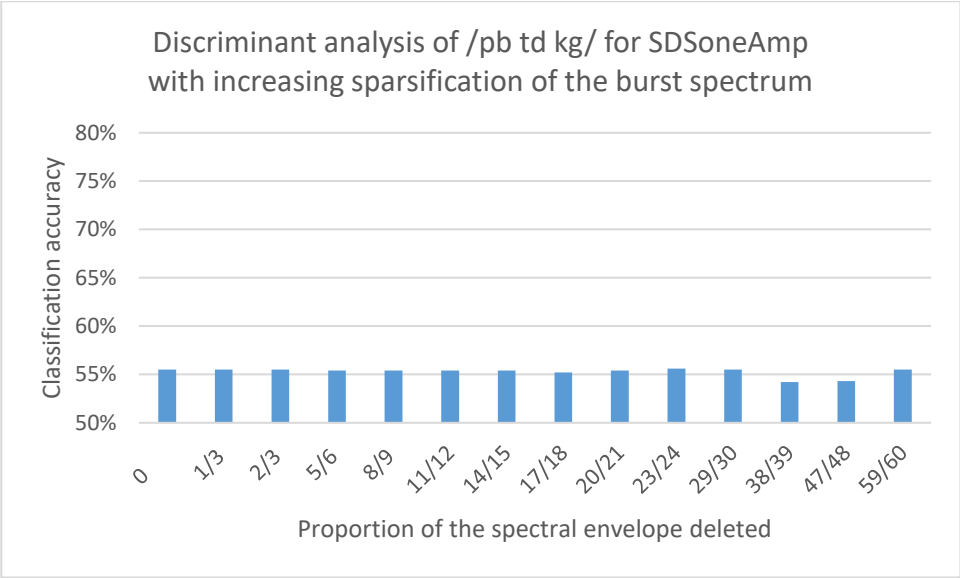


Figure 7.6: Discriminant analysis classification accuracy for the acoustic attribute standard deviation (SDSoneAmp) over a variety of spectral sparsification conditions.
Voiced and voiceless entered into the same classification, N = 5,471.

The consistency of SDSoneAmp over the sparsification conditions is less than that of CoGSone, which is apparent when Figure 7.6 is compared with 7.5. When 7.6 is compared with the results for AllPeakSone in Figure 7.3, its performance over each sparsification condition is remarkably similar to that of AllPeakSone: there is a similar dip in accuracy for the 38/39 and 47/48 conditions, followed by the same jump in accuracy for the 59/60 condition. This suggests that the two attributes tap into much the same information in the burst despite the fact that AllPeakSone picks a single spectral component whereas SDSoneAmp is calculated from all components. If the classification accuracies under all the sparsification conditions shown in Figures 7.3 and 7.6 are each averaged, the mean accuracy of AllPeakSone is 56.0% whereas SDSoneAmp averages 55.3%. This difference is unlike to be large enough to be meaningful: to a first approximation the two attributes do the same thing.

As with CoGSone and AllPeakSone, the classification remains remarkably statistically robust: nearly all the sparsifications above are not statistically significantly worse than the non-sparsified attribute with the exception of 38/39 and 47/48 ($p < 0.01$ in both cases). The 17/18 sparsification reaches $p < 0.05$ but not $p < 0.01$, the current threshold for statistical significance.

Turning now to the results when CoGSone and SDSoneAmp are both used:

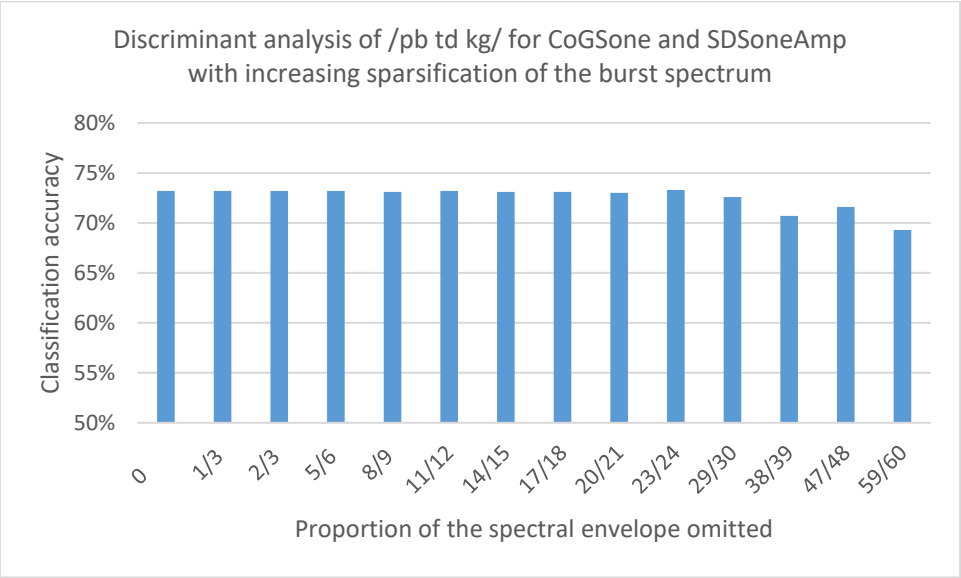


Figure 7.7: Discriminant analysis classification accuracy for the acoustic attributes centre of gravity (CoGSone) and standard deviation of amplitude (SDSoneAmp) over a variety of spectral sparsification conditions. Voiced and voiceless entered into the same classification, N = 5,471.

Similar to what was shown in Figure 7.4 regarding the two AllPeak attributes, the classification accuracy over most of the sparsification conditions remains remarkably consistent. It was noted in the case of AllPeak that the accuracy with a sparsification of 23/24 was still within 0.7 percentage points of the classification accuracy when there was no sparsification at all. For CoGSone and SDSoneAmp the accuracy for 23/24 sparsification is just 0.1 percentage points *higher* than the classification on the full spectrum, despite the fact a sparsification of 23/24 means that only around 4% of the original channels are present. When Figure 7.7 is compared with 7.4, there is a tendency for CoGSone + SDSoneAmp to be slightly less affected by the sparsification than the two AllPeak attributes, at least for all the conditions equal to or greater than 29/30 sparsification. On the other hand, the mean classification accuracy of CoGSone + SDSoneAmp over the entire range of 15 conditions is 72.6%, which is 4.0 percentage points less than AllPeakBark + AllPeakSone over the same conditions. Thus the whole-spectrum attributes CoGSone and SDSoneAmp are weaker on average than the peak attributes AllPeakBark and AllPeakSone, despite the jumpiness in performance of AllPeakBark relative to CoGSone. The fact that the two attributes based on the peak fare better on average than the two based on the entire spectrum is not altogether surprising, since the perceptual studies in the literature review (e.g. Li et al., 2010; Cvengros, 2011) generally showed the burst peak to be the most important piece of information for identifying place of articulation, whereas the spectral moments entail believing that all spectral components matter equally.

In terms of statistical significance, the sparsifications in Figure 7.7 above are mostly not statistically significantly different from the non-sparsified condition. It is only for

sparsifications equal to or greater than 38/39 that the classification is statistically significantly worse ($p < 0.001$ in all three cases). This is similar to the picture for the two AllPeak attributes (Figure 7.4), which showed the same pattern from 29/30 upwards.

In any case, perhaps the most important message of the sparsification results is how well all four attributes hold up under substantial thinning of spectral information. This suggests that the information in the burst envelope for detecting place of articulation is relatively sparsely distributed. This may be one of the reasons why the perception of speech is so robust to degradations in acoustic conditions: many frequency channels can be missing or corrupted without the 'gist' of the spectrum being lost. Perhaps another factor is that the burst is a relatively short acoustic event, which means that (following the time-frequency uncertainty principle; Oppenheim and Magnasco, 2013) one would not expect its spectrum to be as densely packed with information in the frequency domain as a longer acoustic events such as a vowel or the friction of a fricative.

The results suggest that a sample of the burst spectrum could be taken approximately every 2 Bark with little loss of burst information. Thus one of the approaches that will be adopted in the present chapter is to take 12 samples from the burst spectrum sampled 2 Bark apart, beginning at 2 Bark and terminating at 24 Bark. This will be termed the 'sparsified spectrum' approach or the 'spectral samples' approach. As noted in 7.1, its purpose is to serve as a baseline against which to measure the performance some acoustic attributes that were presented in the previous chapter (which will be termed the 'traditional-phonetic' attributes), as well as the 12 DCT coefficients. Comparing the classification accuracy of the traditional burst attributes to the spectral slices is one means of quantifying how much the use of tailor-made attributes improves classification accuracy above and beyond the accuracy yielded by simply using samples of the raw spectrum as attributes.

7.3 Comparison of the Three Attribute Groups

As in Chapters 3 and 6, the statistic employed for dealing with large groups of attributes is random forests. For further information on this statistic, see Section 3.1.7. The decision of how many forests to use was the same as that discussed in that chapter, i.e. 100 trees were used in all cases.

All 5,471 tokens that contain a release burst were inputted to the random forests. As in the previous chapter, the random forest was given no information about the identity of surrounding segments (which will not be added until Section 7.5).

Random forests were run separately for each of the below attribute groups. The attribute groups were as follows (all were derived from the Bark-sone spectrum):

1. The first attribute group consists of the first twelve coefficients of a discrete cosine transform (DCT) of the original 256-point Bark-sone spectrum. As described in Sections 2.3.1.9 and 7.1, this feature type undepins the MFCC front end used widely in ASR (MFCC do of course involve a somewhat different spectrum, namely a Mel frequency scale and logarithmic amplitude, but the DCT feature type is the same).
2. Twelve samples of the original 256-point Bark-sone spectrum sampled at the following points: 2.0 Bark, 4.0 Bark, 6.0 Bark, 8.0 Bark, 10.0 Bark, 12.0 Bark, 14.0 Bark, 16.0 Bark, 18.0 Bark, 20.0 Bark, 22.0 Bark, and 24.0 Bark. The rationale behind using these as features is to see to what extent tailor-made attributes improve the classification accuracy over this intentionally simplistic approach of using burst amplitudes without any feature design.
3. All 10 Bark-sone acoustic attributes tested in the previous chapter, namely: AllPeakBark, AllPeakSone, AllTotalSone, HiPeakSone, HiTotalSone, MidPeakSone, MidTotalSone, CoGSone, SDSoneFreq, and SDSoneAmp. These 10 attributes will collectively be referred to as the 'Traditional-Phonetic' attributes, as all but one of these attributes have been used in previous phonetic studies.
4. The following two Bark-sone acoustic attributes: AllPeakBark, AllPeakSone. These two attributes have been chosen because the results in the previous chapter of the random forests on the Bark-phon and Bark-sone representations established them as being in first and second or first and third position in terms of the degree to which they contributed to the classification accuracy. They thus have reason to be regarded as the strongest attributes from the previous chapter. The aim is to compare their performance to attribute groups 1-3 above, since this comparison tells one how well a relatively simple solution involving just *two* attributes performs relative to the first three attribute groups above in which there is a much larger number of attributes (between 10 and 12 in each).
5. To be sure that this two-way attribute combination really is the strongest two-attribute combination, the following attribute pairs are included for comparison:
    a. CoGSone + SDSoneAmp, or 'CoGSD' for short;
    b. HiPeakSone + MidPeakSone, or 'HiMidPeaks' for short;
    c. TiltPeakSone, or 'Tilt' for short. This attribute is included as it was found to be the strongest compound attribute in the previous chapter. Also, it serves as a benchmark against which to compare HiMidPeaks, since HiMidPeaks is composed of the same attributes as Tilt except that in HiMidPeaks the attributes are kept separate rather than one being subtracted from the other.

These are the results:

Attribute group:	Classification Accuracy (%)	r^2
12 DCT coefficients	85.1	0.78
12 spectral samples	84.2	0.76
All 10 Traditional-Phonetic attributes	82.1	0.72
2 AllPeak attributes	80.3	0.70
2 CoGSD attributes	72.9	0.54
2 HiMidPeaks attributes	79.6	0.70
1 Tilt attribute	74.6	0.70

Table 7.1: Comparison of three attribute groups in their representation of the release-burst information for identifying plosive place of articulation.
Separate random forests were run and each corresponds to a single row (100 trees were used in all cases). Attribute groups consisting of large numbers of attributes are in yellow, smaller attribute groups are in green. N = 5,471.

The three attribute groups give fairly similar classification accuracies, though the traditional-phonetic attributes lag behind the DCT and the spectral samples, which is especially apparent from the results from the r^2: 0.78 for the DCT, 0.76 for the spectral samples, but 0.72 for the 10 Traditional-Phonetic attributes. This indicates that the DCT features capture more of the variation in the burst than any other attribute type. As was noted in Sections 2.3.1.9 and 7.1, the shape of each cycle or coefficient in the DCT basis set has been shown to be similar to the basis set yielded by a principal component analysis (PCA) of the speech spectrum (Lyon, 2017: 77). The transformation involved in PCA, of course, is designed to yield variables that each account for as much of the variation in the data as possible while being orthogonal to the preceding variables. Thus, given the similarity of the DCT basis set to PCA components, the fact that this attribute group captures more of the variance in the data than the other attribute groups is not surprising.

Turning to the results for the one- and two-member attribute groups (in green in Table 7.1), three of them score the same on r^2, each having a score of 0.70, while one (CoGSD) scores markedly weaker, with an r^2 of just 0.54. However, on the random forest two of these attribute groups with an r^2 of 0.70 are clearly weaker, namely CoGdB with an accuracy of 72.9% and Tilt with a score of 74.6%. The two strongest attribute groups, AllPeak and HiMidPeaks, yield similar classification accuracies of 80.3% and 79.6% respectively. The 80.3% accuracy is just 1.8 percentage points less than that yielded when all 10 of the Traditional-Phonetic attributes were run in a random forest. Remember that the two AllPeak attributes (AllPeakBark and

AllPeakSone) are a subset of the 10 Traditional-Phonetic attributes. This means that adding those eight non-AllPeak attributes to the random forest increased the classification accuracy by just 1.8 percentage points and the r^2 by just 0.02.

Another interesting observation from the data in Table 7.1 is that the 12 DCT coefficients have an r^2 of 0.78 whereas the two AllPeak features have an r^2 of 0.70. This means that the DCT coefficients are capturing only ca. 11% more variance from the release burst than that captured by simply noting the frequency and amplitude of the burst peak. That is, two features capture almost 90% as much of the variance captured by 12 features. This could be interpreted as showing that the great majority of the information in the burst for distinguishing place of articulation is concentrated in the burst peak. This is not particularly surprising in light of the perceptual studies explored in the literature review (e.g. Li et al., 2010; Cvengros, 2011) that generally showed the burst peak to be the most important piece of information for listeners in identifying plosives' place of articulation.

When a random forest is run (the number of trees again being 100) in which only the first two coefficients of the DCT are used, the classification accuracy is 70.7% and the r^2 is 0.51. Thus the classification accuracy and variance captured by the two-attribute groups in Table 7.1 above is greater than that for the same number of DCT attributes. In sum, the DCT coefficients yield the highest classification accuracy and capture more of the data's variance but they achieve this by distributing the variance over a larger number of attributes, with a consequent trade-off of parsimony and interpretability. There thus seems to be a tension between simple, interpretable solutions such as AllPeak or HiMidPeaks and maximizing the classification accuracy using a larger number of attributes whose functioning is less straightforward to interpret.

To summarize, if attribute minimization and attribute interpretability are priorities for the researcher, then the frequency and amplitude of the burst peak can be used as features with only moderate loss of classification accuracy over using a larger attribute set such as the DCT coefficients that are purposefully designed to capture as much of the variance as possible. If, on the other hand, the researcher does not care about parsimony and interpretability and instead wants to prioritize maximizing the accuracy of recognition, then clearly the DCT is the strongest approach.

7.4 Adding Time-Domain Burst Information

All the burst attributes used in the previous chapter, as well as those used thus far in the present chapter, have been based on frequency and amplitude. This leaves open the question of to what extent burst information is also contained in the time domain. The segmentation policy followed

in the present study segmented the burst into transient and frication. The transient (as its name suggests) tends to be an ephemeral phenomenon, with a mean duration of 7.50 ms and standard deviation of 3.99 ms (N = 5,471). Thus its duration would not be expected to yield a feature with much ability to distinguish the three places of articulation (and, in any event, there are challenges with segmenting the boundary between the transient and frication that were noted in Section 4.3.5 and will be discussed again in 8.6.2). Instead, the duration of the transient and frication will be summed to yield the feature of burst duration.

Here are the results when the six attribute groups presented in the previous section have this time-domain information of burst duration added to their classification:

Attribute group:	**Classification Accuracy (%)**	**Improvement (% pts)**	**r^2**
12 DCT coefficients	87.6	+2.5	0.78
12 spectral samples	86.1	+1.9	0.76
All 10 Traditional-Phonetic attributes	83.4	+1.3	0.72
2 AllPeak attributes	82.1	+1.8	0.70
2 CoGSD attributes	76.5	+3.6	0.59
2 HiMidPeaks attributes	81.5	+1.9	0.70
1 Tilt attribute	78.4	+3.8	0.70

Table 7.2: The same random-forest methodology as in Table 7.1 except that the burst duration attribute is included in the classification of each attribute group.
(This means that there is one more attribute in the classification than that indicated in the table above, e.g. the DCT classification now has 13 attributes.) Comparison of three feature types in their representation of the release-burst information for identifying plosive place of articulation. Separate random forests correspond to each row (100 trees were used in all cases). Classifications involving large numbers of features in yellow, smaller numbers of features in green. N = 5,471.

Table 7.2 shows that the inclusion of burst duration improves the classification accuracy of the seven attribute groups by an average of 2.4 percentage points, though it has no effect on the r^2 values. This information on burst duration does not appear to have been used by previous studies. Suchato (2004), for example, decided to measure closure duration and VOT (both of which he found to be poor place-of-articulation predictors) but curiously did not measure burst duration.

Furthermore, most ASR front ends (see e.g. Huckvale (2013) for details) are not capable of measuring the burst duration because, as noted in the literature review, the typical acoustic model in ASR involves sampling the spectrum every 10 ms. Sampling the spectrum every 10 ms is probably too blunt an instrument to measure the burst duration with much precision. The burst duration is often less than 10 ms, which means that its acoustic effect may be found in only one or two of the spectral samples.

The present results suggest that the time-domain attribute of burst duration can boost the classification accuracy though only to a modest degree. When run as a single attribute in a random forest, the burst duration has a classification accuracy of 58% and an r^2 of 0.19 (N = 5,471).

7.5 Adding Contextual Information

The results of the previous section have compared the classification accuracy for three different feature types without providing information about the context in which the burst is situated. As established in Chapter 6 using mixed-effects modelling, the backness and stress of the following vowel can both affect the acoustics of the burst. In Chapter 5 we saw that the formant information can classify the place of /b d g/ with considerably accuracy.

In this section the following information is added to the classification in order to quantify its improvement of the classification accuracy:

(1) The voicing of the plosive (whether it belongs to /p t k/ or /b d g/, following the definition described in Chapter 4 and used throughout this study, i.e. /sp st sk/ are classified with /b d g/ due to their short voice onset times); two levels, 'v' for /b d g/, 'u' for /p t k/;

(2) Whether the plosive belongs to a stressed ('a') or unstressed ('b') syllable; note that the 'stressed' category includes all word-initial plosives;

From the segment following the plosive the following information is added:

(3) The identity of the segment, with three levels: vowel ('v'), liquid ('l'), non-liquid or pause ('c');

(4) The backness of the vowel, with four levels: front or central ('a'), back ('b'), schwa ('c'), and non-vowel ('d');

(5) $F2_R1.0$, normalized by $\mu F3_{speaker}$ (as tested in Section 5.4.2);

(6) $F3_{onset}$, normalized by $\mu F3_{speaker}$ (5.4.2);

From the preceding context the following information is added:

(7) The identity of the segment, with three levels: vowel ('v'), liquid ('l'), non-liquid/pause ('c');

(8) $F2_R2.0$ (tested in Section 5.4.5);

(9) $F3_{offset}$ (5.4.5).

The reasoning behind the choice of these pieces of information is as follows. In the previous chapter it was established that the classification accuracy is little affected by whether the bursts of /p t k/ and /b d g/ are classified separately or together. Nevertheless there were some noticeable differences between /t/ and /d/ in that /d/ had noticeably more mid-frequency and less high-frequency energy than /t/, and consequently it was found in the mixed-effects modelling of CoGSone that this feature had a noticeably lower modelled value for /d/ than /t/. This is one reason why including information about the plosive's voicing seems warranted. The other reason is that $F2_R$ was shown to work well on /b d g/ and poorly on /p t k/ (due to the aspiration of the latter series delaying the onset of voicing).

The information about the identity of the following segment seems warranted in that the extracted F2 frequencies associated with vowels and liquids (/l r w j/) are more likely to aid classification than those associated with non-liquid consonants. This is because the extracted F2 frequencies for non-liquid consonants are more likely to be erroneous or meaningless than those for vowels and liquids due to the presence of noise (in fricatives), silence (in plosives), and the low amplitude of F2 in the closure phase of nasals.

As for the use of $F2_R1.0$, this was shown in Chapter 5 to yield a decent classification accuracy similar to that yielded by $F2_{onset}$ + $F2_{mid}$. $F3_{onset}$ was also shown to boost the classification accuracy considerably over what F2 could yield on its own.

As for the information taken from the *preceding* context, the rationale is exactly parallel to that outlined for the following context.

Attribute group:	Classification Accuracy (%)	Improvement (% pts)	r^2	Improvement
12 DCT coefficients	89.5	+1.9	0.84	+0.06
12 spectral samples	87.2	+1.1	0.79	+0.03
10 Traditional-Phonetic attributes	85.3	+1.9	0.76	+0.04
2 AllPeak attributes	87.3	+5.2	0.79	+0.09
2 CoGSD attributes	85.0	+8.5	0.74	+0.15
2 HiMidPeaks attributes	86.6	+1.9	0.79	+0.09
1 Tilt attribute	85.4	+3.8	0.79	+0.09

Table 7.3: Comparison of the attribute groups in their representation of the release-burst information for identifying plosive place of articulation.

As detailed above, nine contextual variables have been added to the classification. Separate random forests run on each attribute group (100 trees were used in all cases). The two columns labelled 'Improvement' indicate how much the addition of these 9 contextual variables has improved the classification and fit to the data. N = 5,471.

For all attribute groups, adding the information about the surrounding phonetic context has boosted the classification accuracy. The strongest attribute group is still the 12 DCT coefficients, whose classification accuracy has risen close to 90%. However, the second strongest attribute group is now AllPeak, which has improved by 5.2 percentage points and outperforms both the 10 Traditional-Phonetic attributes and the 12 spectral-sample attributes. This is a somewhat surprising result since the classification accuracy of AllPeak was less than that of the 10 traditional-phonetic attributes before the addition of the attributes describing the surrounding context.

One might wonder to what extent the information from the preceding context and following context has aided the classification. When a random forest is run that is identical to that described for Table 7.3 but with the four preceding-context attributes removed, the classification accuracy drops by 0.3 percentage points for both the DCT coefficients and the 10 Traditional-Phonetic attributes (89.2% and 85.0% respectively), 1.2 percentage points for the

spectral samples (86.0%), and 1.3 percentage points for AllPeak (86.0%). What is noteworthy about all these figures is that the decline in classification accuracy is small. This appears to suggest that the preceding segment contributes relatively little to the classification accuracy of plosive place of articulation. This is not altogether surprising when the results of the mixed-effects model in Section 5.1.2 are borne in mind: the slope in the linear regression representing the influence of V1 on $F2_{onset}$ (in voiced stops) was just 0.15. Given the present results, it seems that the influence of V1 on the burst also seems to be relatively small, which is again unsurprising given that the burst is closer to $F2_{onset}$ (i.e. the following segment) than it is to the preceding segment.

The result when the equivalent four attributes for the *following* context are excluded (namely $F2_R1.0$, $F3_{onset}$, following segment type, and following segment backness) is that the classification accuracy drops by 0.7 percentage points for the DCT features (88.8%), 0.4 percentage points for the spectral samples (86.8%), 1.3 percentage points for the 10 traditional phonetic attributes (84.0%), and 2.9 percentage points for AllPeak (84.4%). All but one of these figures is larger than the decline when the preceding-context attributes were removed. This seems to dovetail with the finding of Section 5.4.5 that the information in the following vowel yields a higher classification accuracy than the information in the preceding vowel for distinguishing place of articulation. Nevertheless, the loss of classification accuracy when information about either the preceding or following context is omitted is far from drastic, averaging 0.78 percentage points for the former and 1.33 percentage points in the latter.

Finally, we turn to the omission of stress (which indicates whether the plosive is part of a stressed or unstressed syllable). When the stress attribute is omitted, the drop in classification accuracy is 0.6 percentage points for the DCT group of attributes and AllPeak (88.9% and 86.7% respectively), 0.7 percentage points for the spectral samples (86.5%), and 0% for the 10 Traditional-Phonetic attributes. It seems, then, that the information about syllable stress contributes to the random forest classification accuracy only slightly.

The results thus far have focused on those plosives that contain a release burst (N = 5,471). This constitutes the greater part of the overall dataset (87.0% of N = 6,284). We now turn to the results for the entire dataset. That is, we include cases where there is no release burst (N = 813). Given the absence of the release burst, the classification accuracy is expected to be lower than for the tokens with a burst that have been examined thus far.

Attribute group:	Classification Accuracy (%)	Change (% pts)	r^2	Change
12 DCT coefficients	85.7	-3.8	0.78	-0.06
12 spectral samples	85.2	-2.0	0.76	-0.03
10 'traditional phonetic' attributes	84.1	-1.2	0.74	-0.02
2 AllPeak attributes	85.4	-1.9	0.77	-0.02
2 CoGSD attributes	83.7	-1.3	0.72	-0.02
2 HiMidPeaks attributes	84.8	-1.8	0.75	-0.04
1 Tilt attribute	84.1	-1.3	0.77	-0.02

Table 7.4: Comparison of the attribute groups in their classification accuracy and fit to the data of identifying plosive place of articulation.

Unlike in the previous three tables, this table presents the classification accuracy for the entire dataset (N = 6,284), not just those plosives that happen to contain a release burst (N = 813). As detailed earlier, 9 extra attributes have been added to the classification that represent the preceding and following segments, and one further attribute (burst duration) has been added to represent the burst in the time domain. The columns labelled 'change' indicate how much the inclusion of the burstless tokens has disimproved the classification accuracy (from that indicated in Table 7.3). Separate random forest run on each (100 trees used in all cases). Note that when a token did not contain a burst, the burst variables were given the value '0' in the Excel file. N = 6,284.

In all four attribute groups the classification accuracy is, as expected, lower. However, the DCT features show the sharpest decline (-3.8 percentage points). Particularly noteworthy is that the AllPeak attribute group is within half a percentage point of the DCT classification, which seems to suggest that the use of a small number of acoustic attributes from the burst (3 as opposed to 13) can yield a similar classification accuracy to using a larger number of attributes, at least under these specific classification conditions.

7.6 Discussion

This chapter has two main findings:

(1) the information density of the burst in the frequency domain is relatively sparse, as indicated by the fact that removing over 90% of the Bark-sone channels had little effect on the combined

classification accuracy of CoGSone + SDSoneAmp and AllPeakBark + AllPeakSone respectively;

(2) 12 DCT coefficients represent the place of articulation information in the burst with greater accuracy than any other attribute group, as indicated by both the classification accuracy (85.1%) and r^2 (0.78). Nevertheless, AllPeakBark + AllPeakSone managed to come within 5 percentage points of this and had an r^2 (0.70) that was almost 90% as large as the DCT coefficients, making it the strongest two-way combination of burst attributes examined.

This latter result suggests that the largest part of the information in the burst for distinguishing place of articulation can be obtained from the burst peak, which is in accord with the perceptual findings that were presented in the literature review (e.g. Kapoor 2010, Li et al., 2010, Cvengros, 2011).

Another finding of the present chapter is that the two whole-spectrum attributes CoGSone + SDSoneAmp were notably weaker than the attributes that relied on spectral peaks, whether they be the absolute loudest peak in the entire spectrum (AllPeakBark + AllPeakSone) or the peak loudness in specific frequency regions (HiPeakSone + AllPeakSone and Tilt). This was reflected both in the classification accuracy and in the fit of the random-forest model to the data (indicated by r^2). Again, this was not particularly surprising in light of the perceptual studies presented in the literature review. It seems that the importance of the burst peak can be revealed not just by perceptual studies but also by classification studies such as the present one. In Chapter 6 (6.4.1) it was noted that the performance of the 'Peak' attributes relative to the equivalent 'Total' attributes was higher.

The results of the present chapter indicated that the information about the surrounding contexts is important, such as the information in the formants. For example, the accuracy of the two AllPeak attributes increased from 82.1% to 87.3% when this information was included. This may not sound like much, but another way of thinking about it is that the error rate was cut from 17.9% to 12.7%, i.e. an elimination of almost a third of the errors.

Another theme in the present chapter's findings is that there seems to be a tension between minimizing the number of features on the one hand and maximizing the classification accuracy on the other. The DCT coefficients achieved the highest accuracy but there were six times as many of them as the number of AllPeak attributes, which nevertheless managed to score within 5 percentage points of the DCT's accuracy. The greater the number of attributes used, of course, the more difficult it is for the researcher to interpret what it is that each of the attributes are capturing. This may be one reason why acoustic phonetics has a long history (as presented in Section 2.3.1 of the literature review) of developing acoustic attributes that are

specially tailored to the release burst (e.g. HiPeakdB, MidPeakdB, spectral tilt). These handcrafted attributes allow the researcher to know exactly what it is the attributes are measuring. For example, HiPeakSone measures the loudness of the loudest component in the burst's high-frequency region. In contrast, the DCT coefficients are less straightforward to interpret in terms of what spectral aspects they capture; Lyon (2017: 77) notes that the lowest DCT coefficient approximately represents the overall amplitude of the spectrum, with the second coefficient representing the spectrum's tilt (downward tilt being common in voiced speech, upward tilt in fricatives). For higher coefficients, however, because of their greater wiggliness (recall Figure 7.1) it is more difficult to be precise about what aspects of the spectral envelope they measure other than to remember the general point that the higher the coefficient, the greater its wiggliness, which means in effect that each coefficient is measuring progressively finer and finer spectral detail. Perhaps the best analogy that can be mustered is the brush strokes on a canvas: big broad brush strokes for gross aspects of the scene (sky, land, i.e. lower DCT coefficients) with progressively finer brush strokes for the details (tree branches, grass, i.e. higher DCT coefficients).

This lower interpretability of the DCT coefficients relative to tradition phonetic attributes may be one reason why they appear not to have been used by many previous studies of plosive's place of articulation in the field of phonetics. A similar point about the opacity of ASR features relative to formant frequencies and other simpler acoustic measures has been made by Harrison (2013: 56) in a forensic phonetic context and by Johnson et al. (2018: 106) in a clinical context.

7.7 Summary

- The information in the frequency domain for distinguishing plosive's place of articulation appears to be relatively sparsely distributed, as indicated by the fact that when over 90% of the 256 Bark-sone channels were omitted (leaving a channel only every 1.8 Bark), the classification accuracy of AllPeakBark + AllPeakSone and CoGSD + SDSoneAmp was within 2 percentage points of what it had been when the entire spectrum was at the disposal of these attributes.
- This led to the decision to include 12 samples of the burst at 2-Bark intervals as an attribute group against which to compare the attributes used in the previous chapter. A third attribute group, the DCT coefficients, was also included for comparison since these were theoretically expected to capture more of the variance in the burst than other attribute types.

- The results of this comparison showed that the DCT attributes did indeed yield the highest burst-only classification accuracy of the three attribute groups (85.1%, r^2 = 0.78), with the 12 burst samples a close second (84.2%, r^2 = 0.76). The third group of attributes, the traditional-phonetic attributes tested in the previous chapter, classified noticeably less well (82.1%, r^2 = 0.72). Nevertheless, AllPeakBark + AllPeakSone classified the strongest of the two-way attribute combinations examined, with a classification accuracy (80.3%) and r^2 (0.70) that was impressive considering there were only two features in the attribute group as opposed to the 12 features in the DCT and spectral-sample groups.
- The inclusion of burst duration as an acoustic attribute yielded a modest improvement in classification accuracy for all attribute groups examined, averaging 2.4 percentage points.
- The inclusion of information about the segments preceding and following the plosives (including vowel backness, $F2_R$ and F3) yielded a larger improvement in classification accuracy, averaging 3.5 percentage points in the seven attribute groups. On those tokens in the dataset containing a burst (N = 5,471), a 22-attribute combination of 12 DCT coefficients from the burst, the burst duration attribute, and the 9 attributes (including $F2_R$, F3) representing the two adjoining segments yielded the highest classification accuracy, 89.5%.
- When those plosives in the dataset not containing a burst were added to the classification (N = 813, yielding N = 6,284 in total), the classification accuracy of the 22-attribute group containing the DCT coefficients was again strongest (85.7%), with the 12-attribute group containing AllPeakBark and AllPeakSone being a close second (85.4%).
- Overall, the results of the chapter highlight the trade-off between choosing a large number of attributes that maximize classification accuracy (12 DCT coefficients or 12 burst samples) and a smaller number of attributes with somewhat lower classification accuracy but straightforward interpretability (AllPeakBark + AllPeakSone, HiPeakSone + MidPeakSone, Tilt). This tension between accuracy and interpretability is likely a major factor in the divergent choice of burst features found in the history of phonetic science vis-à-vis automatic speech recognition.

Chapter 8: Discussion

8.1 Introduction

This chapter brings together the results of the last seven chapters, puts them in a broader context, and highlights the limitations of the present study. Later, avenues for future research are suggested. Here are the four main aims of this study and the findings for each:

1. To test the performance of a technique for collapsing $F2_{onset}$ and $F2_{mid}$ into a single attribute, termed $F2_R$. The development of this technique was inspired by the observation that $F2_{onset}$ and $F2_{mid}$ tend to be closely correlated (the slope in a regression plot between the two typically being between 0.4 and 0.75). **Result**: $F2_R$ distinguished place with effectively the same accuracy as $F2_{onset}$+$F2_{mid}$, being within ±1 percentage point of $F2_{onset}$+$F2_{mid}$ at its strongest over most of the conditions examined.
2. To compare the performance of burst-based attributes with and without normalization by individual speaker. This was done because normalization by individual speaker of formant frequencies has been in widespread use whereas normalization of aperiodic events such as the burst seems to have been less widely used. **Result**: Lobanov normalization boosted the classification of the 27 attributes by an average of 1.4 percentage points, but even this modest improvement shrank or disappeared when the normalized attributes were put together in random-forest classifications (namely the Hz-dB, Bark-phon, and Bark-sone combinations examined in Section 6.5).
3. To examine the effect of different spectral representations (Hz-dB, Bark-phon, and Bark-sone) on the accuracy of the burst attributes. **Result**: the evidence as to the importance of spectral representation in the classificiation accuracy of the attributes was mixed. Although there was a tendency for the Hz-dB attributes to classify somewhat higher than the Bark-phon and Bark-sone attributes and this difference was in many cases statistically significant, the size of the effect was not particularly large on average: when six mid-frequency and six high-frequency attributes were compared the mean classification accuracy of the Hz-dB, Bark-phon, and Bark-sone attributes was within 1.2 percentage points of each other. In contrast, when the said six attributes were split by mid-frequency versus high-frequency, the high-frequency attributes outperformed the mid-frequency equivalents by an average 13.2 percentage points. This suggests that, in the broader scheme of things, the choice of spectral representation is minor relative to other factors such as the choice of frequency region in the burst.
4. To compare the performance of some traditional burst-based attributes with the first 12 coefficients of the discrete cosine transform (DCT). The motivation for this comparison

was that phonetic science has a long tradition of developing burst attributes that are tailored to the specific task of extracting place-of-articulation information from the burst, whereas automatic speech recognition (ASR) has long used attributes that function on all kinds of acoustic speech material and which are theoretically expected to capture more of the variance in the burst than the attributes that have been traditionally used in phonetic science. **Result**: The DCT coefficients classified higher than 10 'traditional-phonetic' attributes by 3 percentage points and captured more of the variance in the burst data (r^2 = 0.78 rather than 0.72). Nevertheless a combination of two of the traditional-phonetic attributes (AllPeakBark and AllPeakSone) performed impressively for such a small number of features, being within 5 percentage points of the DCT coefficients (r^2 = 0.70). Thus for the phonetic researcher who wants features that are straightforward to interpret, some of the traditional-phonetic attributes do well with only moderate sacrifice of classification accuracy.

In the first part of this chapter, each section discusses one of these four aims, beginning with Aim 1 (Section 8.2), Aim 2 (8.3), Aim 3 (8.4), and Aim 4 (8.5). Following this some limitations of the present study are discussed (8.6). The chapter ends by exploring how the $F2_R$ concept could be improved in future research (8.7).

8.2 $F2_R$

Aim 1 of the present study was to compare the performance of $F2_R$ with $F2_{onset}$ + $F2_{mid}$ under a variety of conditions. It was found that the classification of $F2_R$ at its strongest (i.e. for values of *c* ranging from 1 to 1.8) was similar to the classification of $F2_{onset}$ + $F2_{mid}$, e.g. 0.5 percentage points higher when the values were unnormalized (5.4.1), 0.8 percentage points lower when $F2_R$ normalized by $\mu F3_{individual}$ (5.4.2), 0.9 percentage points higher when tokens were separated by backness (5.4.3), and 1.1 percentage points higher when $F3_{onset}$ was added to the classification (5.4.3). This suggests that it is indeed possible to collapse the $F2_{onset}$ and $F2_{mid}$ attributes into a single attribute. The motivation for this merging of the attributes was explored in Section 5.2, namely to produce an attribute for place of articulation that is freer from the acoustic influence of the surrounding segmental context and, in the process, reduce the number of attributes in the classification system.

Despite the broad success of the $F2_R$ concept, it is important to recognize two potential limitations of the present study. The first limitation I do not regard to be a substantive limitation, but is worth mentioning nonetheless. The second limitation is more interesting, and will form the basis of the discussion in Section 8.7 as to how $F2_R$ might be improved in future research.

Although formant tracking errors were manually corrected in the pilot study, the results presented in the main study did not involve correcting formant errors. Let us consider the advantages and disadvantages of correcting tracking errors. The obvious advantage is that by compensating for the limitations of LPC, the researcher produces results that are closer to the 'true' formant frequencies. This attempt to strive after accuracy is reasonable. On the other hand, there are a number of decisions that a researcher has to make in practice that make the replicability of hand-corrected formant frequencies challenging. If a researcher sees that a formant's frequency appears to be 1,200 Hz on the spectrogram but the tracker measures it as 1,600 Hz, this is a 400-Hz discrepancy and probably all researchers would agree that this is sufficiently large to warrant correction. But if a discrepancy is smaller, say 100 Hz, it is less clear that all researchers would see this as worth correcting. This point is important because whilst manually correcting the pilot study's formant frequencies, numerous cases were noted where the formant tracker tracked F2 in back vowels such as [o] slightly lower in frequency than it appeared on the spectrogram. This 'error' appeared on so many occasions that it was unclear whether correcting it was really warranted, given that (1) it appeared relatively consistently, and (2) tended to be a relatively small discrepancy of approximately 100 or 150 Hz and was mostly confined to back vowels.

Trialling of different formant settings did not alleviate the discrepancy; instead, the discrepancy appeared to be particularly liable to occur whenever F1 and F2 were relatively close in frequency and F1 was much greater in amplitude than F2. This example illustrates the difficulty in deciding when to correct formant tracking errors. Such decisions strike me as being quasi-arbitrary in nature and introduce a difficult-to-replicate element into the data measurement. Another reason formant tracking errors were not corrected is that the present data were recorded under perfect acoustic conditions; if recording conditions had involved background noise, then perhaps a stronger case could be made for correcting formant tracking errors.

Despite this view, I nevertheless reran the classification with formant tracking errors in CV /b d g/ context corrected (N = 1,936). This involved changing 242 values of $F2_{onset}$, 242 values of $F2_{mid}$, and 440 of $F3_{onset}$. When a discriminant analysis was run consisting of $F2_{onset}$, $F2_{mid}$, and $F3_{onset}$, the result was 68.8% accuracy prior to correction, 69.5% afterwards, an improvement of 0.7 percentage points. As expected, correcting formant frequencies made little difference to the classification accuracy. Correction of formant errors is of course extremely time-consuming; these results suggest that when acoustic conditions are optimal, the return on investment for such correction is meagre.

One criticism that could be levelled at the $F2_R$ concept is that it does not approximate the 1950s $F2_{locus}$ concept closely enough. This became evident in 5.4.4 from the comparison of the mean $F2_R$ frequencies for /d/ before back vowels versus front vowels: for /d/'s $F2_R$ to have the same frequency in the two vowel contexts, *c* had to be set to 0.6. However, the results for classification *accuracy* showed that the accuracy was at its highest when *c* was set to 1.2. This meant that the best variant of $F2_R$ was one in which back-vowel $F2_R$ was higher in frequency than front-vowel $F2_R$. But of course the 1950s locus theory (which worked best of all on /d/, recall) posited that the $F2_{locus}$ for /d/ in all contexts should be *the same*, as was illustrated by Figure 2.11 on page 24. This discrepancy between $F2_{locus}$ and $F2_R$ will be discussed further at the end of the present chapter (Section 8.7), where a potential solution will be presented.

8.3 Normalization of Burst Attributes by Individual Speaker

Aim 2 of the present study was to compare the classification accuracy of burst attributes with and without normalization by individual speaker. Two kinds of normalization were tested: Lobanov-normalization, in which the mean value of a burst attribute for a particular speaker was subtracted from the attribute and the resulting difference divided by the speaker's standard deviation score for that attribute:

$$\text{Attribute}_{\text{normalized}} = \frac{\text{Attribute} - \mu}{\sigma}$$

The other kind of normalization, Norm, was very similar to the one above except that there was no division by standard deviation, i.e. the normalization was a somewhat simpler version of the formula above.

It was found that the Lobanov-normalization improved the classification of 25 burst attributes by an average of 1.46 percentage points, while Norm improved it by 1.25 percentage points (Section 6.4.4). Both figures constitute a relatively small improvement, and are arguably not large enough to justify the attribute normalization, even though the boost in accuracy did turn out to be statistically significant for most of the attributes ($p < 0.001$ for 14 of the 27, and $p < 0.01$ for a further 2). When the attributes were inputted to a random forest (one for Hz-dB attributes, one for Bark-phon attributes, and one for Bark-sone attributes) the 1.4-percentage-point boost shrunk and in one of the cases even reversed (i.e. the random forest accuracy was slightly higher when the forest consists of non-normalized attributes). These results, taken together, suggest that the normalization of burst-based attributes is not profitable.

The theoretical motivation for studying this phenomenon was the observation that normalization has been widely used on vowels. For example, Flynn (2012) compared the

classification accuracy of well over a dozen methods of vowel-formant normalization, and Adank et al. (2004) similarly examined a wide variety of vowel-formant normalizations (11 in total). The choice of Lobanov-normalization was based on the fact that it had performed strongest in Adank's study. Given this pervasive attention in the literature to the normalization of vowels as well as the paucity of previous burst studies using normalization, this topic seemed in need of further attention. The results of the present study suggest that normalization of burst attributes is unnecessary.

However, one might wonder whether there are hints in previous research that this result would be found. Forrest et al. (1988) recorded 5 males and 5 females uttering in isolation 14 words containing plosives. Although the scale and nature of their study are far from ideal, it is interesting to note that when the authors used an attribute model based on the male speakers' voiceless stops to classify the place of articulation of the female speakers' voiceless stops, it did so with about 94% accuracy. The authors note that, to the best of their knowledge, "this is the first demonstration of a high rate of cross-gender classification in the absence of any additional normalization procedure" (p. 123). Given the small size of their data, this finding, though interesting, should be taken with caution. Nevertheless, the results of the present study do also suggest that the role of individual-speaker differences in affecting the acoustics of place of articulation in stops is less than what one might have imagined, given the long literature on individual-speaker normalization in the context of vowels.

The present study examined just two kinds of (closely related) normalization technique. Given the vast literature on the normalization of vowel formants, it remains something of an open question whether some other normalization technique – whether a pre-existing method already utilized on vowels or a new method – would have greater success at improving the classification accuracy of burst attributes. The results of the present examination of burst normalization suggest not. Furthermore in 6.3.1 it was noted that there were no overlaps between the mean F2 frequencies of individual male and female speakers. In contrast, there were overlaps between the individual males and females in their mean CoG values, e.g. m09's mean alveolar CoG was 17.8 Bark whereas f03's was 16.1 Bark. This suggests that the individual-speaker variation in burst attributes in the dataset simply doesn't pattern in the same fashion as the variation in formant attributes. It is thus unlikely that a different choice of normalization technique would change the picture.

8.4 Spectral Representations

Aim 3 of the present study was to compare the performance of burst-based attributes on three kinds of spectral representation: Hz-dB, Bark-phon, and Bark-sone. When individual attributes

from each spectral representation were compared (6.4.1), the difference in classification accuracy did in many cases turn out to be statistically significant, with a trend for the Hz-dB variant to be the strongest of the three in most cases. However, mean classification accuracy of eight attributes on these three representations was within 1.1 percentage points of each other, with the Bark-phon attributes weakest of the three. When the eight attributes were combined in a random forest, the overall classification accuracy of the three was again close, within 1.5 percentage points, with Bark-phon attributes now being the strongest of the three. Taken together, these findings do not evidence one spectral representation being consistently more accurate than the others for representing acoustic attributes. Furthermore, the comparison of six mid-frequency and six high-frequency attributes showed the mean classification accuracy of the two groups to be 13.2 percentage points apart, whereas the mean classification of the Hz-dB, Bark-phon, and Bark-sone attributes across the two groups was within 1.2 percentage points of each other. This again suggests that in the broader scheme of things the choice of spectral representation is less important for burst attributes than other considerations such as the frequency region of the burst utilized.

Given these results, it is worth revisiting the motivation behind this comparison (4.5.2). The equal-loudness contours that underpin both the Bark-phon and Bark-sone spectra boost the amplitudes of the spectrum in a manner that is different from a conventional Hz-dB representation (whether this Hz-dB representation be pre-emphasized or not). Pre-emphasis boosts the amplitude of spectral components by 6 dB per octave, and carries this boosting through to all frequencies. Although the equal-loudness contours are somewhat similar to pre-emphasis in the general property of boosting higher frequencies relative to lower frequencies, there are a number of important differences, the most important of which being that the boosting of the equal-loudness contours below 1,000 Hz is greater than that above 1,000 Hz, and that the boosting stops above ca. 4,000 to 5,000 Hz (as was shown in Figures 4.8 and 4.9 in Section 4.5.2).

The present study did not employ pre-emphasis on the Hz-dB spectrum (see Section 4.5.1 for details) and yet even with all of these differences between the Bark-phon and Bark-sone spectra on the one hand and the Hz-dB spectrum on the other, there was little difference in the mean classification accuracy of attributes derived from these three spectral representations. In particular, it could have been expected that the attribute AllPeakBark would classify more accurately on the Bark-phon or Bark-sone spectra than its Hz-dB equivalent, AllPeakHz. This is because the lack of equal-loudness contours on the Hz-dB spectrum could result in a low-frequency peak becoming greater in amplitude than a high-frequency peak in an alveolar spectrum, for example. (When equal-loudness contours are not applied, then by

definition the lower-frequency components will be more prominent than they would be with the equal-loudness contours). Surprisingly, the accuracy of AllPeakHz was 0.6 percentage points higher than that of AllPeakBark, i.e. effectively the same accuracy (and the McNemar test revealed that the difference was indeed not statistically significant).

At first blush, one might think that this result suggests that the use of auditorily-oriented approaches in phonetics is unwarranted, at least if the researcher's aim is to boost attributes' classification accuracy. However, it is important to note that although the Bark-phon and Bark-sone spectral representations used in the present study have been termed 'auditorily-oriented', in reality they are quite far from being a comprehensive approximation to what the auditory periphery accomplishes.

Here are just a handful of the differences. The auditory periphery has a much richer representation of sound in the time domain than what was used in the present study (see Moore, 2012: 44-46 for an introductory discussion of phase locking). The Bark-phon and Bark-sone spectra used in the present study were obtained by warping the frequency and amplitude axes of the input from an FFT. FFTs are of course based on Fourier analysis, and Fourier analysis allows the neglect of time-domain information such as phase in favour of information about frequency (see Schnupp et al., 2011: Chapter 1 for a lucid exposition). In contrast, the auditory system has a far richer representation of variation of sound with time (termed 'fine-time structure' or 'temporal fine structure') than what can be achieved with FFTs. In this respect in particular, the Bark-phon and Bark-sone spectra used in the present study are remote from what occurs in the auditory system. Under real-life conditions, this fine-time structure is one critical factor that aids the separation of sound mixtures (termed 'grouping' in the literature; Bregman, 1990). See Section 9.2 for a concrete example of a burst in which the FFT-style neglect of time-domain information may be consequential.

Another difference between an FFT and what the auditory system utilizes is that an FFT entails the same time-frequency resolution at all frequencies, whereas the auditory system utilizes filters with higher frequency resolution (but lower time resolution) at low frequencies with lower frequency resolution (but higher time resolution) at high frequencies. (This factor will be revisited in the discussion of window length in 8.6.2.)

With the above considerations in mind, it should be clear that the 'auditorily-oriented' spectra used in the present study should not be thought of as being imitations of the auditory periphery. Therefore the results of the comparison should not be interpreted as showing that the use of auditorily-oriented spectral representations is not worthwhile; rather, the results should be interpreted as indicating that there is no consistently superior spectral representation out of the three specific spectral representations that were tested in the present work.

8.5 DCT Coefficients versus Traditional-Phonetic Burst Features

The final aim (Aim 4) of the present study was to compare the performance of the burst attributes that have traditionally been used in phonetics with an alternative popularly used in speech recognition whereby a discrete-cosine transform (DCT) is performed on the spectrum and the first 12 coefficients of the transform are used as attributes. In Section 2.3.1.9 a theoretical reason was outlined for expecting the DCT coefficients to be able to capture a greater amount of the variance in the burst. Sure enough, the DCT coefficients outperformed the traditional-phonetic attributes at distinguishing place of articulation based on the burst information, both in terms of classification accuracy and the r^2 fit of the random-forest model to the data. The difference in classification accuracy between the two types of attribute was between 3 and 5 percentage points depending on the number of traditional-phonetic attributes admitted to the classification. When the contextual information from the formants and the surrounding segment was included in the classification, the difference in classification accuracy between the DCT coefficients and the traditional-phonetic attributes did shrink, but the DCT coefficients nevertheless remained the strongest attribute type.

Given the similarities of the DCT coefficients to the components of a PCA, this result should not be altogether surprising. Indeed, it raises questions about why phonetic science has spent so much time developing ad hoc attributes of inferior accuracy to those that have been used in ASR for decades. However, as was noted when discussing the comparison between DCT coefficients and the traditional-phonetic burst attributes (Section 7.6), there is a trade-off to be made between classification accuracy on the one hand and attribute number or attribute interpretability on the other. It was found that AllPeakBark + AllPeakSone classified within 5 percentage points of the DCT coefficients, even though there were only two AllPeak attributes versus 12 DCT coefficients. It was also noted that when the number of DCT coefficients was pruned to the first two coefficients, their classification accuracy was far lower than that of AllPeakBark and AllPeakSone. Thus AllPeakBark and AllPeakSone, despite not being as accurate as the DCT coefficients, nevertheless manage to perform impressively well for just two attributes. This was also revealed in the r^2 value, which was 0.70 for the two AllPeak attributes as against 0.78 for the 12 DCT coefficients. This result could be interpreted as showing that almost 90% of the information for place of articulation in the burst is contained in the burst peak: simply knowing the frequency and amplitude of the highest-amplitude component in the burst was enough to yield 80.3% accuracy, which rose to 85.4% upon the addition of further features representing the burst duration, the formant frequencies and the type of adjoining segments.

It seems, then, that there is a law of diminishing returns: to extract as much place information from the burst as possible, it appears necessary to use an attribute type that is optimized to capturing variance, as is the case with the DCT coefficients (or perhaps also PCA components). On the other hand, the preoccupation of those researchers in the 'knowledge-based' approach to speech recognition (e.g. Adelatty Ali et al., 2001; Suchato, 2004) is to develop attributes that are interpretable, that are small in number, and that to some degree reflect our knowledge of the relation between speech production and speech acoustics. The results for AllPeakBark and AllPeakSone suggest that, because the information in the burst for place of articulation seems to be predominantly contained in the burst peak, using two attributes that measure this peak captures the majority of the information in the burst for place of articulation with only moderate loss of classification accuracy.

There is a tension between feature minimization / feature interpretability / feature robustness-to-noise on the one hand, and feature classification accuracy on the other. If the motive of the researcher is to maximize an acoustic model's classification accuracy, at least under the conditions of quiet examined in the present study, then the best burst attribute group is the DCT coefficients. If, on the other hand, the researcher is less interested in the absolute accuracy of burst attributes and more concerned with minimizing the number of burst features, then the two AllPeak attributes (AllPeakBark and AllPeakSone) are probably the best alternative, with HiPeak/MidPeak and Tilt a close second.

8.6 Limitations of this Study

This section identifies and discusses some limitations of the present study.

8.6.1 Inclusion of Fricative Realizations

One potential limitation of the present study was the inclusion of fricative realizations in the dataset (which constitute 1.1% of the tokens in the dataset, N = 71). As noted in the methodology chapter this inclusion was justified on the grounds that the present study is a study of place of articulation and the friction found in the (apical) alveolar fricative variants appeared to be sufficiently acoustically similar to the acoustics of affricated alveolar bursts to be kept in the dataset. On the other hand, the present study is not simply concerned with place of articulation but with *plosive* place of articulation; in this sense I now consider the inclusion of such tokens to have been unwarranted. Given that the fricative cases constitute just 1% of the total, it is unlikely that their inclusion affected the results by much. Nevertheless it is important to acknowledge this limitation, both to facilitate critical evaluation of the present study's results and to encourage future research to avoid such a decision.

8.6.2 Segmentation of the Burst

In terms of the segmentation criteria, I believe all aspects of them are satisfactory with one exception: the decision to segment the burst into the transient and frication.

This decision was prompted by the observation that a heavily affricated /t/ tends to have a transient whose acoustics are very different from that of the frication.

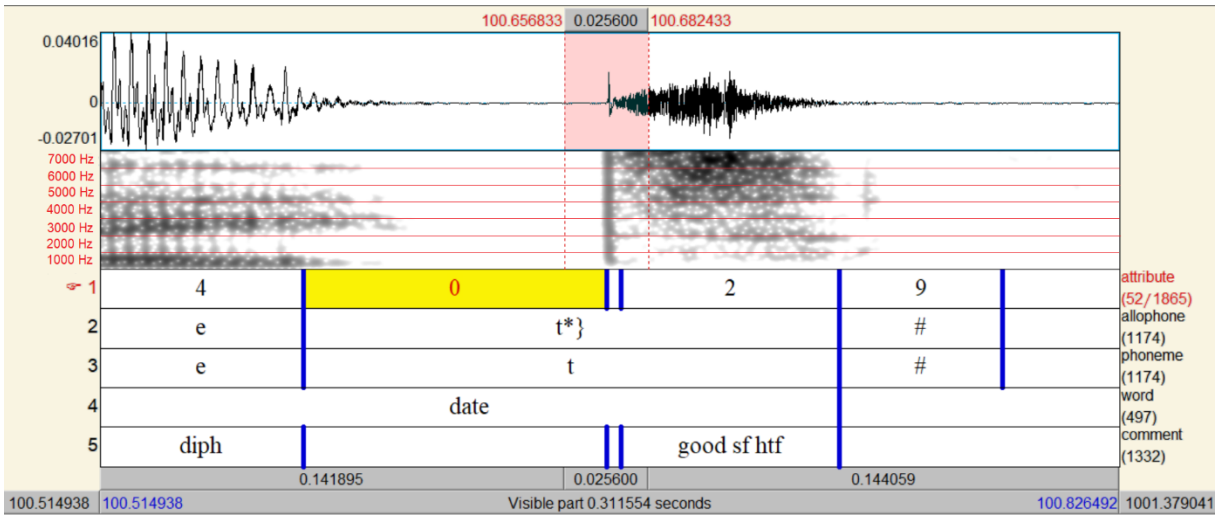


Figure 8.1: Heavily affricated /t/ burst, from f07 *date*.

The transient is 4 ms long, the frication 67 ms long. If the 25.6-ms window is centred at the beginning of the burst (shown in pink), it does not capture the alveolar burst shape accurately (see Figure 8.2).

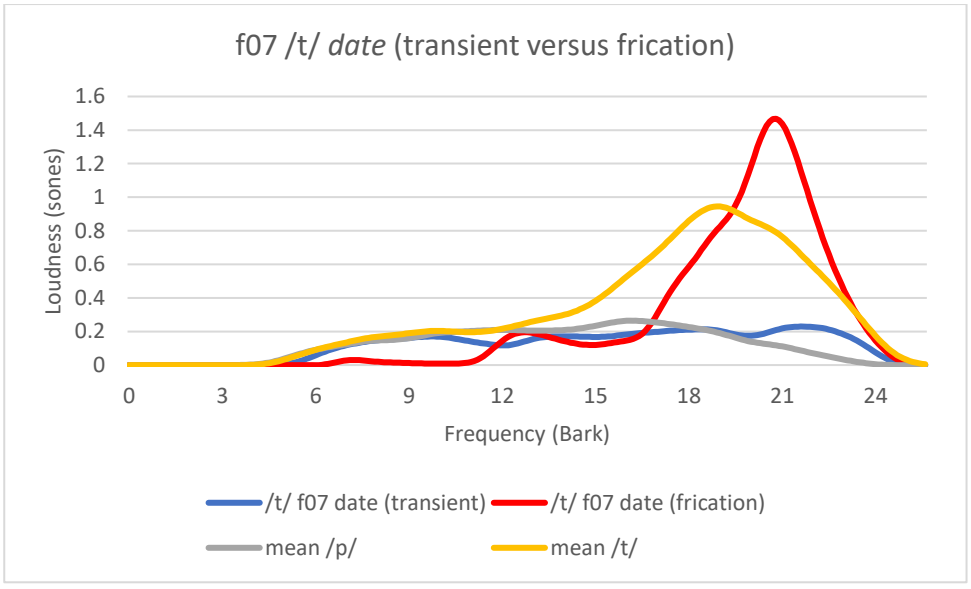


Figure 8.2: The /t/ spectral envelope from f07 *date* sampled at the transient (blue) and frication (red) compared to the mean /p/ (grey) and /t/ (orange) spectra.

This /t/'s transient spectrum resembles that of a /p/ whereas the frication is similar to /t/. The window that sampled the transient was in the position indicated in pink in Figure 8.2; the window that sampled the frication was centred in the 67-ms frication on the righthand part of Figure 8.2.

As can be seen in Figure 8.2, if the 25.6-ms window is positioned at the onset of the burst, i.e. at the transient, then it will yield a spectrum that would be misclassified as bilabial. If, however, the window is positioned in the middle of the burst, i.e. in the frication, the spectrum will be correctly identified as alveolar.

Given the existence of such cases, it seemed necessary to segment the burst into the transient and frication as it would allow the option of investigating the acoustics of such difficult cases further, whereas if the burst had been segmented into a single acoustic event it would have precluded such a possibility. The other reason that caused the need for the burst to be segmented into transient and frication was the fact that the window for measuring the burst was to be centred at the onset of the burst. The discovery of affricated /t/s in which the onset of the burst was /p/-like thus meant that putting the window at *that* part of the burst would have led to misclassifications of /t/ as /p/. In contrast, the segmenting of the burst into transient and frication allowed for the possibility of using the 'correct' information – from the frication – in such cases.

However, in practice, the decision to segment the burst into transient and frication created unforeseen difficulties. Although there were plenty of bursts in which it was straightforward to decide where the transient ended and the frication began (/p b/ tended to be straightforward due to their lack of frication), there were quite a few cases in which it was difficult to be sure where the transient ended and the frication began, and there were also cases where it was difficult to decide if the burst should be segmented into transient alone or transient-plus-frication.

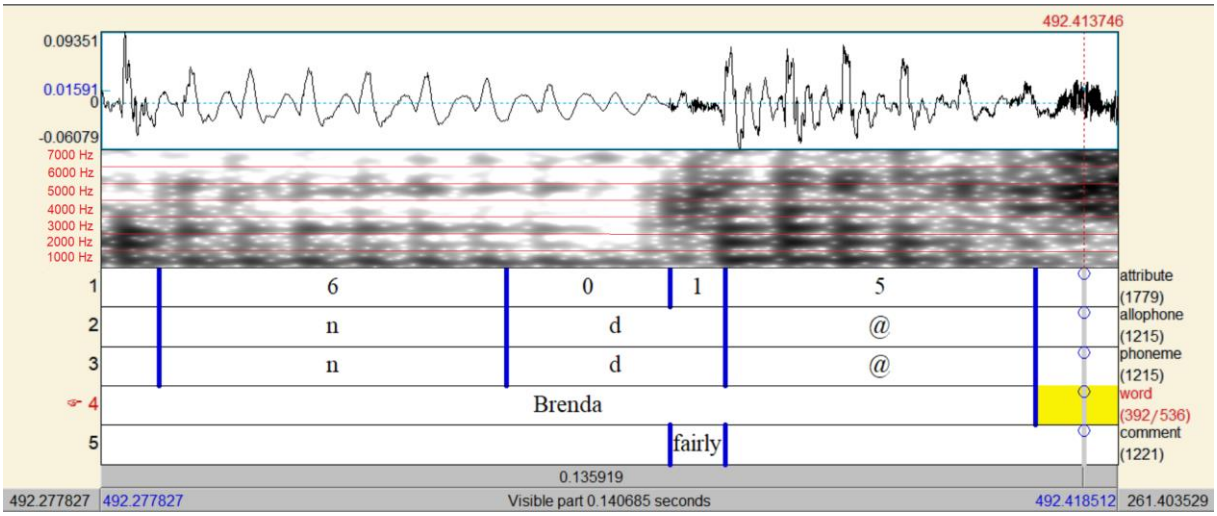
On the other hand, Figure 8.3 illustrates that the acoustics of the burst can be erroneously /p/-like at the start of the burst of at least some affricated /t/ tokens. This observation, recall, was what prompted the decision to segment into transient and frication in the first place. In hindsight the solution would have been to centre the window at the middle of the burst rather than at the onset. So why was this choice not taken? Placing the window at the middle of the burst would have increased the risk of the window being contaminated by the vowel onset. Recall from Chapter 4 that the window length used on the burst in the present study was 25.6 ms. The preoccupation with avoiding the vowel onset would have been alleviated if a shorter window, say 12.8 ms, had been chosen. The choice of a 25.6-ms window was motivated by three considerations: (1) it had been used on the burst in a previous study of plosives (Blumstein and Stevens, 1979); (2) the longer the window, the better the frequency resolution; and (3) a longer window averages together a longer chunk of the burst, which means the resulting spectral envelope would be less sensitive to the precise position of the window. Let us scrutinize each of these considerations.

The previous study of plosives that used a 25.6-ms window (Blumstein and Stevens, 1979) nevertheless discusses the possibility that a shorter time window may, in certain cases, have yielded burst spectra with a shape that is more representative of velar and alveolar place of articulation (p. 1013). They also discuss the possibility of using a window of a different length for different frequency bands, i.e. a relatively long window for low-frequency components with shorter windows for progressively higher frequencies. Such an analysis is what occurs in the auditory system (ibid.) and also underpins the wavelet transform such as the Haar, Le Marie-Meyer, and Malvar wavelets (Pintér, 1996; Addison, 2002: 312-313). Thus even though Blumstein and Stevens did use a 25.6-ms window, closer examination of their discussion reveals that they did not endorse such a long window length.

Regarding consideration (2), namely the desire to maximize spectral frequency resolution, this appears to be a relatively unimportant consideration for an acoustic event such as the burst, since the burst tends to be fairly brief, and brief acoustic events tend to involve less frequency resolution, following the time-frequency uncertainty principle (Oppenheim and Magnasco, 2013). (The mean burst duration in the present dataset is 19.5 ms, with a standard deviation of 18.2 ms; N = 5,471.) Indeed the results of the spectral sparsification exercise in Section 7.2 support the idea that the acoustics of the burst can be adequately captured with relatively poor frequency resolution (recall from Figure 7.3 that the combined classification accuracy of AllPeakBark and AllPeakSone dipped by just 0.5 percentage points when 20 out of every 21 of the burst's 256 channels were removed).

With regard to consideration (3), it may well be true that a longer window – because it averages together a longer portion of the burst – would have a spectrum that is less likely to change shape depending on the precise position of the window. This is what was found in pre-study trialling of a 7-ms window. When the window was moved over the burst in 1-ms increments, the envelope appeared to change shape from one millisecond to the next more than when a longer window was used. But in any event had the window been centred in the *middle* of the burst, there is no reason to expect that the spectrum would not have been representative of the place of articulation, regardless of the window length.

Finally, it should be noted that the 25.6-ms window, even though it was centred at the beginning of the burst (i.e. as far away as possible from the vowel onset), seems nevertheless to have been affected by the onset of the vowel in a few cases. In these cases the burst happened to be abnormally brief and low in amplitudde (which meant that the VOT was unusually short):



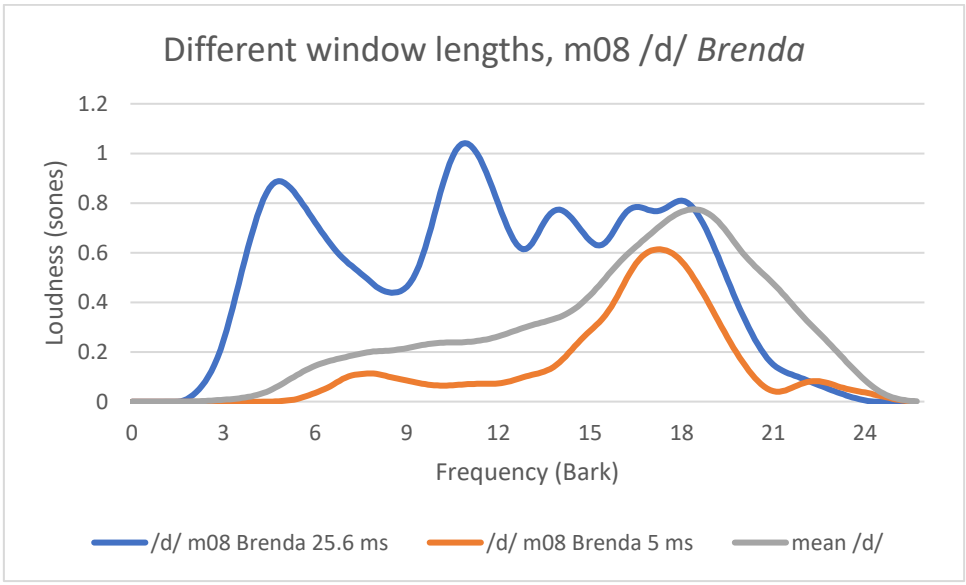


Figure 8.3: Top image: Waveform, spectrogram, and annotation of the /d/ in *Brenda*, uttered by m08. Bottom image: Spectrum of the /d/ in m08 *Brenda* compared to the mean /d/ spectrum for the entire dataset as well as the spectrum that would have been yielded had a 5-ms window been used.
In the top image, the pink selection shows the 25.6-ms window; note how the window extends into the onset of the following vowel. In the bottom image, note how atypical the spectrum of 25.6-ms window is due to the presence of formants from the following vowel.

It can be seen that the 'burst' of the /d/ in m08's utterance of *Brenda* is in fact dominated by the F1 and F2 peaks of the following vowel onset when the window is 25.6 ms long. The 5-ms window manages to avoid the following vowel onset and hence shows a more typically alveolar burst spectrum. Nevertheless, the burst in this particular /d/ is unusually short in duration (7.6 ms; mean /d/ burst duration in the entire dataset is 17.2 ms, N = 1,251) such that when the word *Brenda* is played with and without the /d/ burst, no audible difference could be detected. Thus the unusual nature of this particular case has as much to do with the heavy voicing continuing from the nasal and the unusually short burst (and consequently short VOT) as it does with the choice of window length.

In the context of fricatives, an alternative to the discrete Fourier transform known as the multitaper method (Thomson, 1982) has been widely used, and has recently been used by Johnson et al. (2018) on the /t/-/k/ contrast. However, as Reidy (2016: 2522) notes, the decision between the mutlitaper and DFT does not have much of an effect on acoustic attributes (such as peak frequency), as he demonstrated in another study (Reidy, 2015).

To summarize, segmenting the burst into transient and frication is difficult. We have seen how the decision to do so was connected to the decision to place the window at the onset of the burst, which in turn was connected to the decision to use a relatively long window.

There are three prescriptions that can be made for future studies of plosives:

1. Do not attempt to segment the burst into the transient and frication;
2. Avoid long windows for the burst, e.g. use 6.4 ms or 12.8 ms *but not* 25.6 ms;
3. If only one sample of the burst is desired (rather than, say, a multitaper), centre the window in the middle of the burst rather than at the beginning.

8.7 Improving $F2_R$

As discussed in 8.2, the collapsing of $F2_{onset}$ and $F2_{mid}$ into a single attribute, $F2_R$, was broadly successful. Nevertheless it was noted in 8.2 that the comparison of mean /d/ $F2_R$ frequencies before front vowels and back vowels indicates that $F2_R$ is different from $F2_{locus}$ in that $F2_R$ values tend to be higher before back vowels than front vowels, whereas the 1950s locus theory indicates that $F2_{locus}$ should be the same (for /d/) regardless of the backness of the following vowel. This discrepancy between the two suggests there is room for improvement in the $F2_R$ concept.

8.7.1 An Analogy from Vision

To see how $F2_R$ might be improved, we compare $F2_R$ to a perceptual phenomenon in vision known as colour constancy. A ball of snow appears white indoors whereas a lump of coal appears black outdoors, even though the intensity of the light hitting the retina from the outdoor coal is higher than that of the indoor snowball (Pinker, 1997: 7-8). This ability of the visual system to filter out the brightness of the lighting conditions when gauging the colour of an object is known as colour constancy (Palmer, 1999: 133-137). There are numerous visual illusions that illustrate the phenomenon, such as the following:

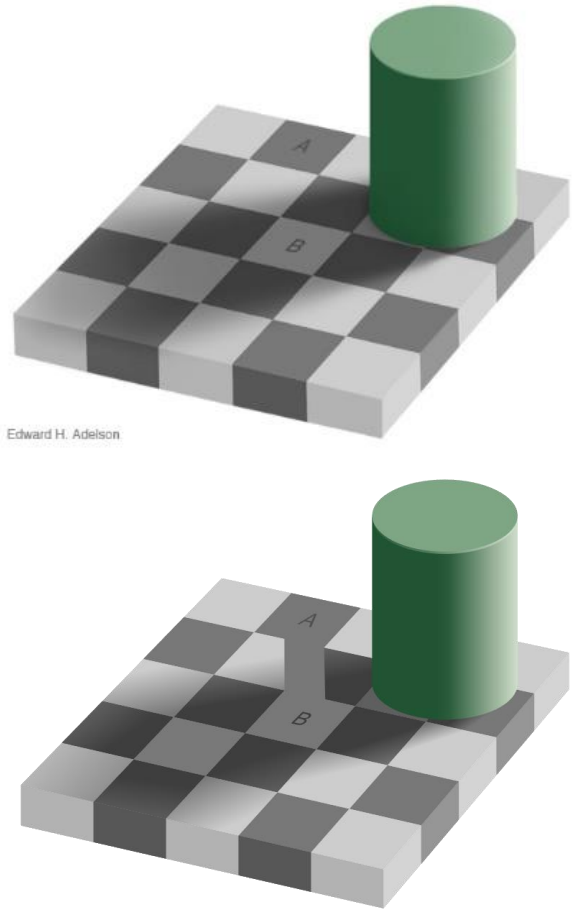


Figure 8.4: Image illustrating colour constancy.
In both images the tiles labelled A and B have the same light-intensity. Nevertheless the tile in the shade is perceived as white whereas the tile in the light is perceived as grey. The fact that the two tiles really have the same light-intensity is apparent from the lower image in which a grey streak has been used to join the two tiles. Source: Adelson (1995).

Colour constancy is an example of a broader phenomenon in our sensory modalities, that of perceptual constancy (for a selection of scholarly contributions on the topic, see Walsh and Kulikowski, 1998). Perceptual constancy refers to the tendency of perceptual systems to remove contextual effects in order to categorize an object or event as the same under different conditions. This is necessary because the properties of objects are independent of the brightness of the scene in which they are situated: a coal is still hard and sooty whether it is found indoors or outdoors. Thus a perceptual mechanism such as colour constancy that filters out the effects of lighting when processing a scene is one that aids the perceiver in seeing an object as the same object under different lighting conditions.

In phonetics there has been much discussion of the nature and degree of invariance found in speech perception (e.g. Liberman and Mattingly (1985); Lindblom, 1996; Fowler, 1990, 1996). The present discussion is not intended to rehash these longstanding debates.

Instead, the aim is to place $F2_R$ in a broader context by highlighting its commonalities with colour constancy.

Turning back to Figure 8.5 above: what kind of algorithm might allow one to perceive tile A as grey and tile B as white? Colour constancy is a vast topic in vision research (Dannemiller, 1989; 1998), so the following sketch is intended merely to provide an analogy for the reader to understand $F2_R$ better. An algorithm for perceiving tiles A and B as different colours would need to begin by obtaining the mean intensity of the pixels in the image region in which each tile is situated. In the case of tile A this mean pixel intensity will be higher than for tile B, since tile A is surrounded by light tiles on three sides whereas tile B is surrounded by dark tiles (= low-intensity pixels) on all sides. Once the mean intensities of these two regions of the image have been obtained, the intensity of each pixel within each region can be subtracted from the mean intensity of its region (Pinker, 1997: 28-29). This notion is formulated in the following formula:

colour of a tile = intensity of its pixels – mean intensity of the pixels in surrounding context

The result of applying such a formula is that a medium-grey pixel located in the shadow (i.e. in the region of the image of relatively low mean pixel intensity) will be perceived as brighter than a pixel of identical intensity located in the light (i.e. in the region of the image of relatively high pixel intensity). This is because the intensity of the medium-grey pixel in the shadow (tile B) is much *higher* than the mean intensity of the tiles in the shadows, whereas the medium-grey pixel in the light (tile A) is much *lower* in intensity than the mean intensity of the tiles in the light.

What does $F2_R$ have in common with colour constancy? Examine the following spectrograms:

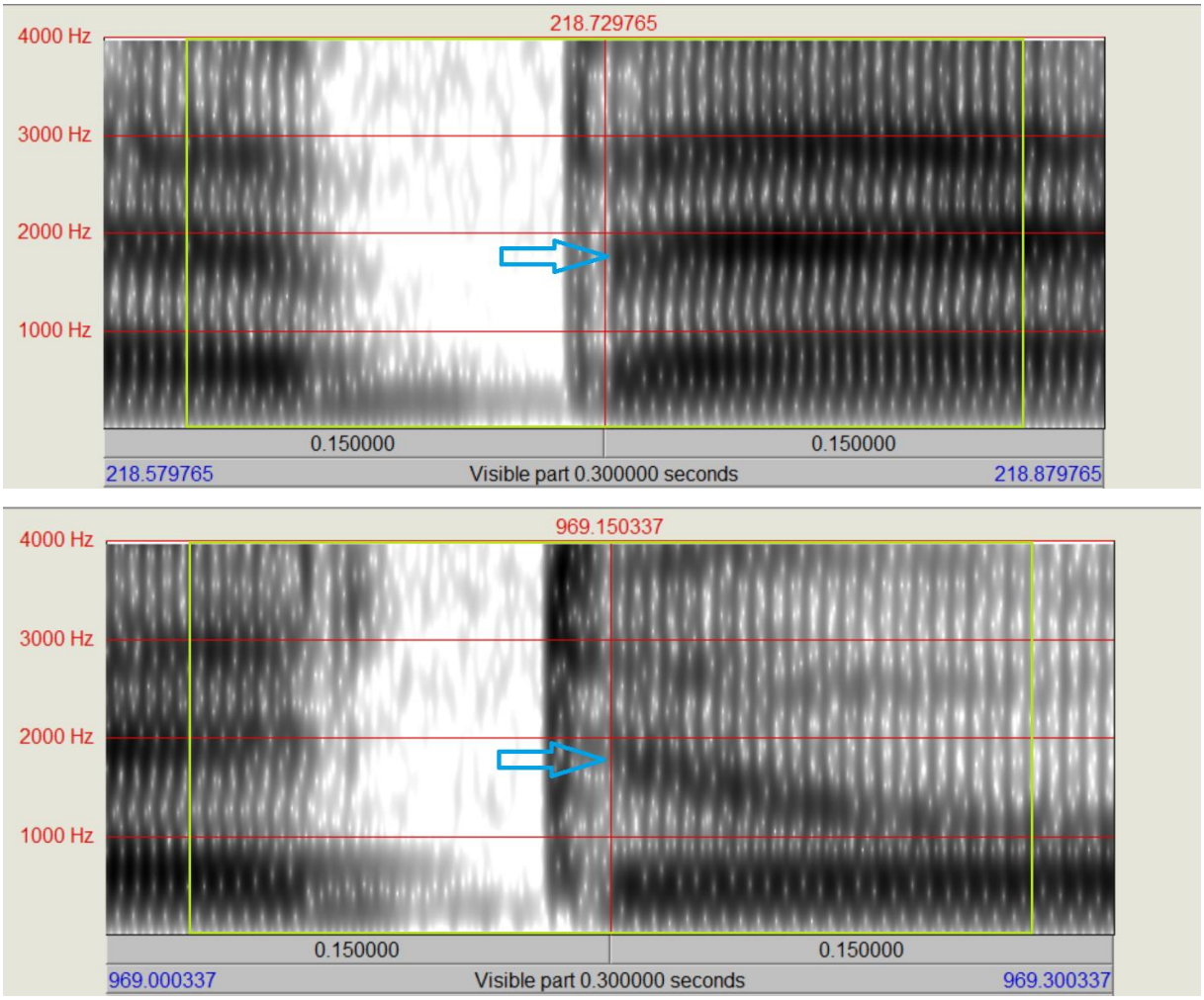


Figure 8.5: Comparison of $F2_{onset}$ in [ˈɛːˈbɛː] and [ɛːˈdoː], uttered by the same (female) speaker.
The lime-green boxes delineate the region within 125 ms of $F2_{onset}$ (i.e. a 250-ms zone). The blue arrow indicates 1,750 Hz, the approximate frequency of $F2_{onset}$ in both images.

In both cases, $F2_{onset}$ is approximately 1,750 Hz. However, if we examine the F2 frequencies within 250 ms of $F2_{onset}$ (i.e. within the lime-green box in the two images) the average frequency of F2 is 1,859 Hz in [ɛːˈbɛː] but 1,477 Hz in [ɛːˈdoː].

This means that the $F2_{onset}$ of [b] in [ɛːˈbɛː] is *lower* in frequency than the surrounding context whereas in [ɛːˈdoː] it is *higher* in frequency. Notice how this is parallel to the case of tiles A and B in Figure 8.5. The only difference is that tiles A and B involved a difference in mean *light-intensity*, whereas [ɛːˈbɛː] and [ɛːˈdoː] involve a difference in mean *F2 frequency*.

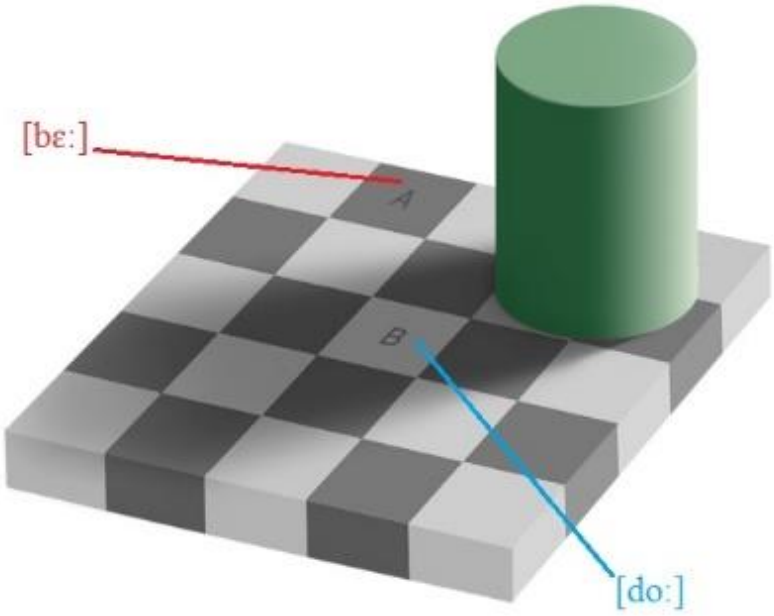


Figure 8.6: Analogy of colour constancy and $F2_{onset}$ variation.

In the analogy illustrated by Figure 8.6, sound frequency is represented by the light-intensity of a given pixel: the higher the $F2_{onset}$ frequency, the higher the light-intensity. The place of articulation of the plosive is represented by the colour of the tile: white tiles are alveolar, grey tiles are bilabial. The shadow cast by the green cylinder represents the acoustic effect of [o:] whereas the part of the board that is free of shadow represents the effect of [ɛ:].

8.7.2 A New Kind of $F2_R$

Having presented an analogy between $F2_R$ and colour constancy, the question arises: how might $F2_R$ be improved by this insight? In the present study the $F2_R$ concept was implemented by subtracting $F2_{mid}$ from $F2_{onset}$ and then subtracting this frequency difference from $F2_{onset}$. This relatively simple implementation has been sufficient as a proof of concept, i.e. that $F2_R$ has a similar classification accuracy on /b d g/ to using $F2_{onset}$ + $F2_{mid}$. However, a more elaborate way of implementing $F2_R$ was hinted at by the lime-green box in Figure 8.5: a window (say, 250 ms in length) would be centred at $F2_{onset}$ and the mean F2 frequency falling within that window would be used in place of $F2_{mid}$. This implementation of $F2_R$ would be similar to colour constancy on the checker board in the sense that the mean F2 frequency in the lime-green window is analogous to the mean pixel intensity of the pixels surrounding tiles A and B.

Thus the implementation of $F2_R$ would change from this:

(1) $F2_R = F2_{onset} - (F2_{mid} - F2_{onset})$

to this:

(2) $F2_R = F2_{onset} - (\mu F2_{onset250ms} - F2_{onset})$

where $\mu F2_{250ms}$ represents the average F2 frequency in the 250-ms environment surrounding $F2_{onset}$ found in the lime-green box in Figure 8.5. (Note that the choice of 250 ms to represent the context is preliminary. Further discussion of this duration later.)

One might wonder what the benefit of this more elaborate $F2_R$ would be. There are two possibilities, both of which were touched on in Chapter 5. Recall from Section 5.1 that when mixed-effects modelling was applied to $F2_{onset}$, it was found that the influence of $F2_{mid}V_2$ on the frequency of $F2_{onset}V_2$ was 4.33 times greater than the influence of $F2_{mid}V_1$ on $F2_{onset}V_2$ (slopes = 0.65 and 0.15 respectively). From Figure 8.5 it can be seen that the number of F2 measurements (i.e. red dots within the lime-green box) coming from V_2 relative to V_1 is 20 as against 6. This means V_2 is contributing 3.33 times more to $\mu F2_{onset250ms}$ than V_1. This value is somewhat smaller than the 4.33 value found in the mixed-effects modelling, but it does illustrate the notion that a window that averages the F2 frequencies occurring within a certain distance of $F2_{onset}$ would contain more information from V_2 than V_1. Thus a window centred on $F2_{onset}$ is a possible mechanism for implementing the idea that the frequency of $F2_{onset}$ is slightly correlated with V_1 but much more strongly correlated with V_2.

A second reason for putting a ca. 250-ms F2 average in the $F2_R$ formula in place of $F2_{mid}$ is as follows. In Section 5.4.4 the $F2_R$ values for front-vowel and back-vowel /d/ were compared. It was found that when the constant *c* was given the value 0.6, the mean $F2_R$ frequencies for back-vowel and front-vowel /d/ were approximately equal. However, if *c* was set to a value greater than 0.6, then the mean $F2_R$ frequency of back-vowel /d/ became higher and higher, whereas the mean $F2_R$ frequency of front-vowel /d/ remained approximately the same. In the [ɛːˈdoː] example given in Figure 8.5 above, the value of $\mu F2_{onset250ms}$ is 1,477 Hz whereas the value of $F2_{mid}$ is 965 Hz. What this means is that if $F2_{mid}$ were replaced by $\mu F2_{onset250ms}$ in the $F2_R$ formula, the value of $F2_R$ would be less extreme. This might solve the problem with $F2_R$ that was discussed in Sections 5.4.4 and 8.2 whereby $F2_R$ for back-vowel /d/ was higher than than for front-vowel /d/, which rendered $F2_R$ out of step with the 1950s $F2_{locus}$.

To see how this idea would work, a concrete example is in order. Up until now the 250-ms window has been described as being centred on $F2_{onset}$, i.e. as though it only applied to $F2_{onset}$. In reality, the window could be used on other parts of the syllable as well. In fact, the window could be a sliding window that replaces the every observed F2 value with a new F2 as follows:

(3) $F2_{new} = F2_{observed} - (\mu F2_{250ms} - F2_{observed})$

In this approach, every F2 frequency is not perceived in its own right; instead it is perceived *relative* to the mean F2 frequency in its environment. (Note how this is analogous to the checker-board pixels of Figure 8.6: the colour of each pixel is perceived *relative* to the mean pixel intensity in its environment.)

The procedure consists of the following steps:

1. F2 is extracted every 6.25 ms (following the Praat default). The F2 at each moment in time is averaged with the three preceding and three following F2 frequencies. This is done for all F2 frequencies in the F2 trajectory. The result is a smoother trajectory. Here is an illustration of the idea:

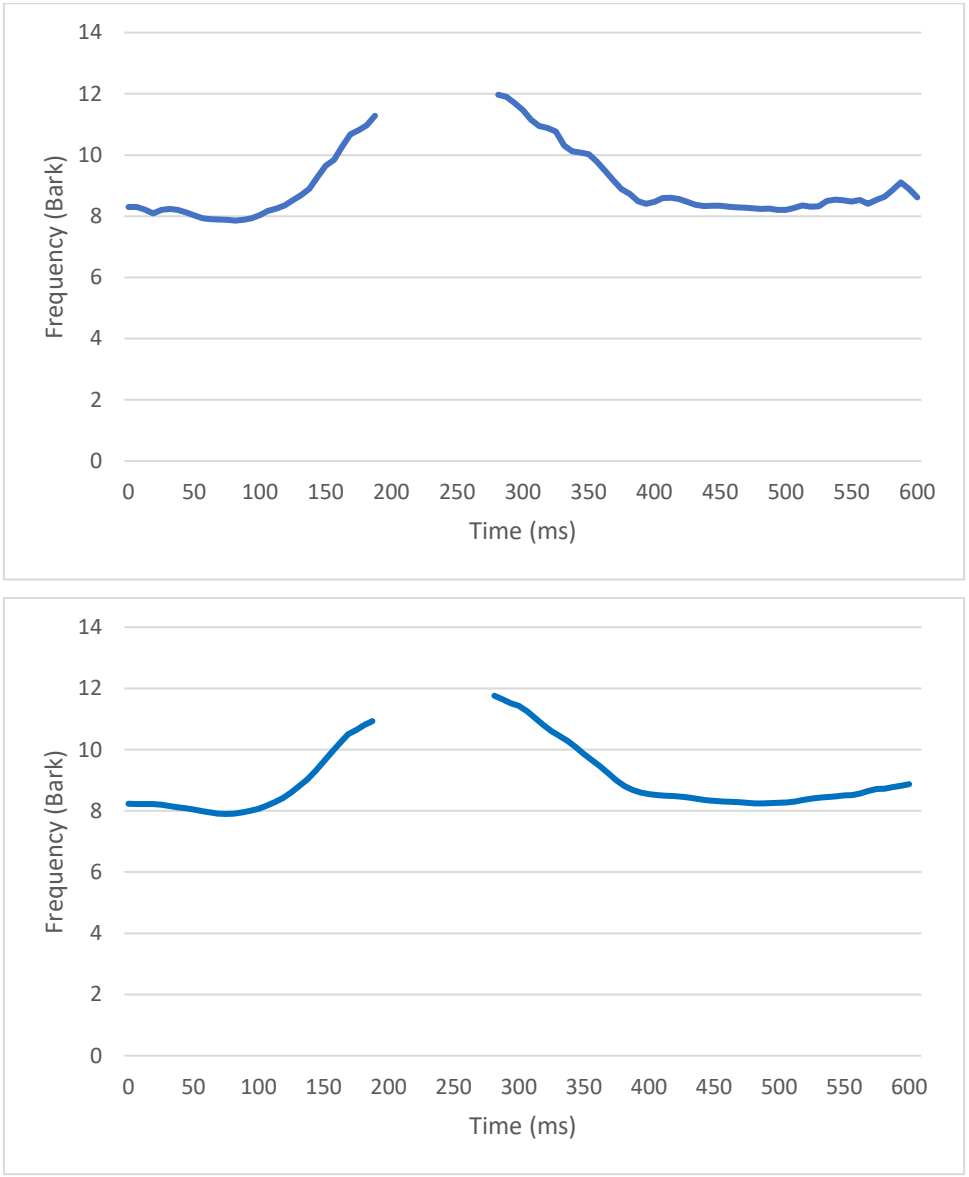


Figure 8.7: Schematic spectrogram of the F2 trajectory in the utterance [oːˈdoː], as spoken by f01 in the pilot study's material.

The upper figure shows the F2 frequencies as extracted every 6.25 ms from Praat (using the default settings, i.e. maximum formant = 5,500 Hz). The lower figure shows the same contour except that the smoothing described in Step 1 above has been applied, i.e. each datapoint has been replaced by an average of itself plus the three preceding and three following datapoints in the F2 trajectory. Note that the gap in the middle of the contour represents the closure, burst, and aspiration phases of the plosive, from which no formant frequencies are extracted.

2. In the next step, each of these smoothed F2 values forms the input to the filter described in Formula (3) above. Thus:

(4) $F2_{new} = F2_{smoothed} - (\mu F2_{smoothed250ms} - F2_{smoothed})$

Where $F2_{smoothed}$ is the output of Step 1 (Figure 8.7) and $\mu F2_{smoothed250ms}$ is the mean frequency of all the F2 values that occur within 125 ms of a given F2 datapoint. The formula is applied to every single datapoint in the contour. Here is a comparison of the contours before and after the application of this step:

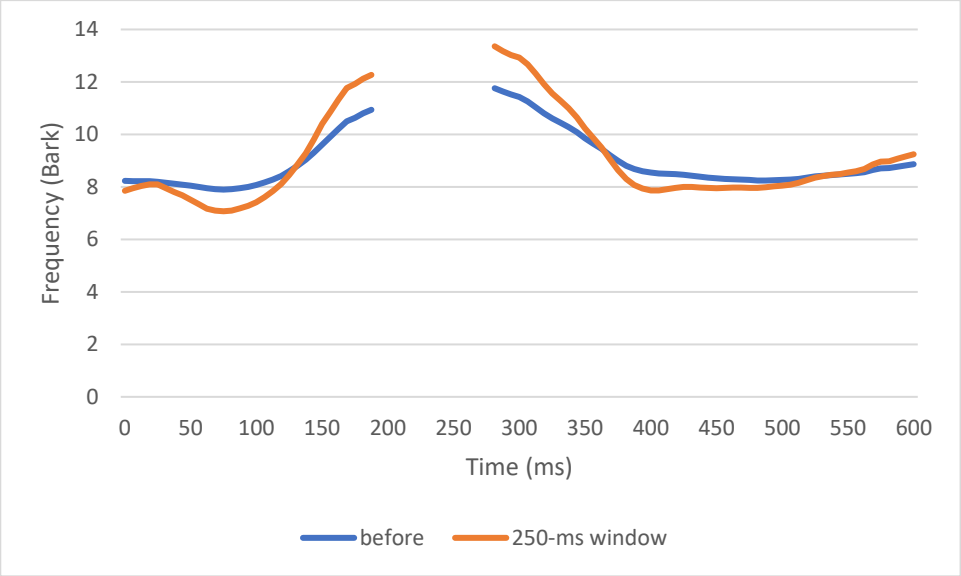


Figure 8.8: Schematic spectrogram of the F2 trajectory shown in Figure 8.8 before and after the application of Formula (4).
Formant settings in Praat are the same as those specified in Figure 8.8.

The most important result of applying Formula (4) is that the $F2_{onset}$ and $F2_{offset}$ frequencies have risen. This rise reflects the filter's purpose of removing the coarticulatory influence of surrounding vowels on $F2_{onset}$ and $F2_{offset}$. The $F2_{onset}$ frequency has risen from 11.76 to 13.36 Bark. Note that this 13.36 Bark very similar to f01's mean /d/ $F2_{onset}$ in the dataset (12.9 Bark, N = 108). Given that f01's mean /d/ $F2_{mid}$ value in the dataset is 12.1 Bark, then her /d/ $F2_{locus}$ frequency is probably slightly higher than her mean $F2_{onset}$ value of 12.9 Bark. In other words, the 13.36-Bark $F2_{onset}$ that occurs after the application of the filter seems to be a close approximation of the 'correct' $F2_{locus}$.

Why is this important? If the standard $F2_R$ formula (in which c = 1) is applied to the above case, then the $F2_R$ value would be 15.65 Bark, or 3,005 Hz. It need hardly be said that this is an excessively high value for $F2_{locus}$ in an alveolar. Instead, as outlined above /d/'s $F2_{locus}$ in a female speaker would be expected to be in the vicinity of 12.9 Bark, which translates to between ca. 2,000 and ca. 2,250 Hz.

Admittedly the $F2_R$ formula tested in this study was varied using the constant c, and it was noted that when c = 0.6, the $F2_R$ values for back-vowel /d/ matched those of front-vowel /d/. However, it was also also found (Section 5.4.2) that the accuracy of $F2_R$ was not highest when c was 0.6, but rather when c was slightly greater than 1 (1.2).

In summary, there are indications that the use of $\mu F2_{onset250ms}$ instead of $F2_{mid}$ in the $F2_R$ formula could yield better $F2_R$ values than the $F2_R$ formula employed in the present study.

The overall purpose of this section has been to identify some of the shortcomings of the $F2_R$ concept as it was implemented in the present thesis, and to suggest how the concept could be developed further in future research. Analogies from vision have been drawn on to aid the reader's understanding of the abstract logic of the process. This topic deserves further exploration but unfortunately constraints of space mean that it will not be developed further in the present work. However, it is hoped that the exposition has shown that the new kind of $F2_R$ demonstrated above has potential, and that the removal of coarticulatory effects using a filter of this general kind is worth exploring.

Further investigation of the concept should investigate:

(1) the effect of different window lengths (e.g. 150 ms versus 250 ms versus 350 ms) on the output $F2_{onset}$ frequencies;

(2) whether the length of the window could be set in a non-arbitrary fashion, e.g. if the speaker were uttering at a rate of 4.8 syllables per second this could be used to set the window length, viz. 1/4.8 yields a window length of 208.33 ms;

(3) whether the window should weight all frequencies occurring within it equally or whether the frequencies occurring towards the middle of the window should be weighted more heavily than the ones occurring towards the edges (this latter option might reduce the wiggle observable in the lefthand orange F2 transition in Figure 8.8 above).

Chapter 9: Conclusions

Here are the main results of this study:

1. $F2_{onset}$ and $F2_{mid}$ can be collapsed into a single attribute, $F2_R$, without compromising classification accuracy by much if at all.
2. Normalizing burst attributes by individual speaker does not improve the classification accuracy of the attributes sufficiently to justify its use.
3. Overall, the classification accuracy of burst attributes does not appear to be appreciably affected by the choice between a Hz-dB, Bark-phon, and Bark-sone spectrum. Far more important is whether the attributes come from the mid-frequency or the high-frequency of the burst.
4. Traditional phonetic burst attributes, such as the AllPeak attributes and HiMidPeaks attributes, classify place of articulation less well than 12 DCT coefficients but are fewer in number and easier to interpret.

9.1 Final Discussion of Findings

It is interesting to note that two of these findings (the results for Aims 2 and 3) are negative. Thus the widespread use in phonetic science of a Hz-dB spectrum and the practice of not normalizing burst attributes by individual speaker appear to be relatively inconsequential. The present study's findings on these two matters seem to provide at least some support for existing research practice on these matters.

The comparison between the DCT coefficients and the traditional-phonetic attributes (Aim 4) was intended to raise the issue of whether phonetics' longstanding practice of developing tailor-made burst features has been such a good idea, given that there is an alternative attribute set that captures more of the variance in the burst and which can be used as attributes for any spectral type, not just bursts. I argued that the traditional-phonetic burst features, while being less accurate at distinguishing place, are nevertheless fewer in number and easier to interpret than the DCT coefficients. Thus the fact that the DCT coefficients have been shown by the present study to yield greater classification accuracy of place is unlikely to end the use of traditional-phonetic burst attributes, given the conflicting aim of having easily interpretable attributes. This importance of phonetic acoustic attributes being interpretable has been made before (Harrison, 2013: 56).

In terms of the formant-based attributes (Aim 1), it appears that the collapsing of $F2_{onset}$ and $F2_{mid}$ into a single attribute is viable. However, one shortcoming of $F2_R$, which was highlighted and explored in Sections 5.4.4, 8.2, and 8.7, is that it produces F2 values for /d/ that

are higher before back vowels than before front vowels (at least when the constant *c* is set to any value greater than ca. 0.6). In this respect it is out of sync with the $F2_{locus}$ concept of the 1950s. One potential solution was presented, namely to replace the $F2_{mid}$ component of the $F2_R$ formula with a kind of local (syllable-length) average of F2. We saw in 8.7 how this did indeed yield less extreme values of $F2_R$ in [oːˈdoː] than the variant of $F2_R$ employed in the present thesis that used $F2_{mid}$. In sum, $F2_R$ is an encouraging idea but may need refinement. It is hoped that the presentation in Chapter 8 will serve as a roadmap for such improvement.

The present study attempted to put the phonetics of plosives' place of articulation in a more cross-disciplinary context than appears to have been done by most previous phonetic studies of the topic, such as the comparison of the traditional-phonetic attributes with the DCT coefficients, the comparison of the Hz-dB, Bark-phon and Bark-sone spectra, and the discussion in the literature review of forward masking from the burst and its effect on $F2_{onset}$. The aim of this cross-disciplinary outlook has been to stimulate further awareness and discussion of issues that sometimes appear to have been neglected in phonetics. For example, the aim of presenting the simulation study of forward masking (Xie, 2013) and listeners' responses to F2-less stimuli (Cvengros, 2011) was to increase awareness of such perceptual factors in phonetics.

9.2 Avenues for Future Research

It is worth reflecting on what the present study did not do. There are two topics in particular that deserve further attention. One of them was touched upon in the previous chapter, that of a more sophisticated version of the $F2_R$ formula in which $F2_{mid}$ is replaced by some kind of local average of F2. The exploration of the potential in this idea was cursory so it is clear that there is more that could be undertaken to refine and develop this idea. In particular, there is a need for a more principled means of setting the duration of the selection that calculates the local F2 average. The presented used a 250-ms selection. Further exploration of selection lengths as well as the underlying principle for setting the selection length (e.g. syllable rate per second, vowel duration) will be necessary before this elaborated version of $F2_R$ can be brought fully to fruition.

The second major thing that the present study did not do is investigate place of articulation using a full-blown auditory model. In this respect the study is little different than nearly all the previous acoustic phonetic studies that were examined in the literature review. There are a number of challenges with the use of auditory models, just as there is in any kind of cross-disciplinary research. Nevertheless, the richness of information in the time domain, as well as the incorporation of automatic gain control (one of the factors that generates the forward masking discussed in the literature review) are just two out of many factors that make full-

blown auditory models considerably different from the Bark-phon and Bark-sone spectra used in the present study.

Here is a concrete example of a release burst in which an auditory model might prove superior to what has been used by the present study and other phonetic studies:

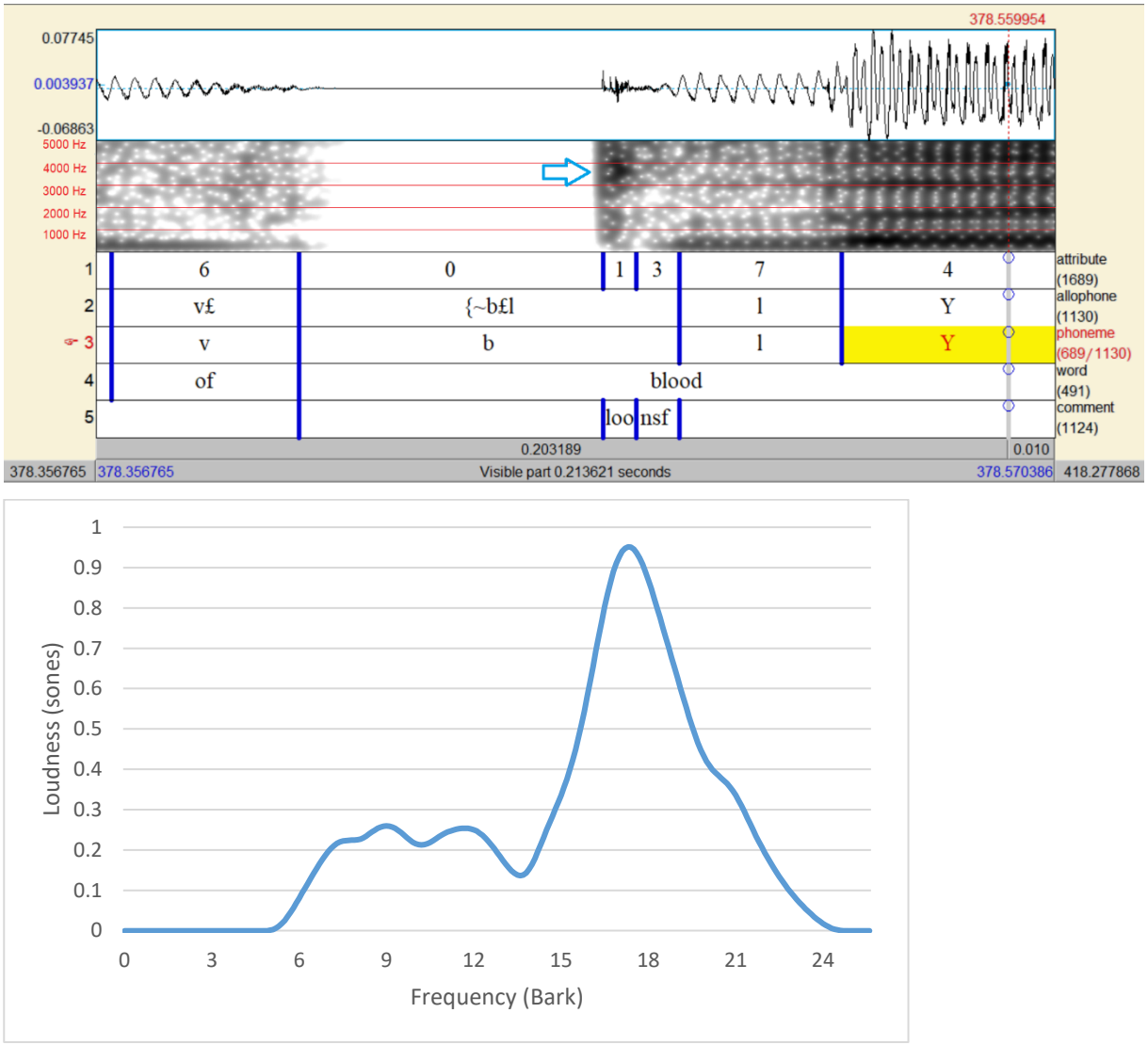


Figure 9.1: Spectrogram (top) and Bark-sone spectral slice (bottom) of the release burst in f09's utterance of the /b/ in *blood*.

The blue arrow points to the burst peak at 3,600 Hz, and the spectrogram displays from 0 to 5,000 Hz.

The release burst contains a sharp peak in the high-frequency region. This is, of course, the quintessentially alveolar pattern. However, when the burst is listened to – either on its own or as part of the overall word – it does not sound alveolar. This might seem strange given that in the slice the burst looks extremely alveolar. However, close inspection of the spectrogram reveals that this burst's duration is just 7.5 ms, whereas the mean /d/ burst duration in the present dataset is 17.2 ms (N = 1,251) and the mean /b/ burst duration is 6.1 ms (N = 825). Furthermore, even though the envelope has a high-frequency peak, this peak appears to last even less than

the rest of the burst, ca. 4-5 ms. This kind of detailed temporal information about how long energy is present in specific frequency bands is, however, lost in the spectral slice. It seems that if the high-frequency peak is very brief, as it is in Figure 9.1, then the burst sounds bilabial rather than alveolar.

Another respect in which a full-blown auditory model could shed insight is on the matter of how important the formant transitions are relative to the burst. Cvengross (2011) showed that listeners are surprisingly good at identifying plosives that lack an F2 transition: just 3.4 percentage points lower. In contrast plosives that lack a release burst were far more difficult for listeners to identify: 39.4 percentage points lower. This massive difference between the perceptual utility of the burst and the F2 transition was suggested to be a byproduct of forward masking. A full-blown auditory model (i.e. one with an automatic gain control loop, unlike the 'Bark-phon' and 'Bark-sone' spectra used in the present study) would represent this enhancement of the burst relative to the F2 transition, and would also allow one to quantify exactly how much masking of the F2 transition there is depending on such factors as the plosive's place of articulation articulation and whether the following syllable is stressed.

This lack of consideration of auditory factors does have consequences in phonetics. For example, Sussman et al. (1998: 246, 287) argue that the F2 transition is the "single most important cue in speech perception". It is difficult to see how such a viewpoint can be reconciled with the effects of forward masking on the prominence of the F2 transition relative to the burst (Xie, 2013; Section 2..3.7), nor with Cvengros's (2011) empirical findings on listeners' accuracy at identifying plosives lacking an F2 transition relative to plosives lacking a burst.

There seems to be a need for a cross-disciplinary research programme that combines the in-depth knowledge of speech sounds that has typified phonetic research, with the auditorily-oriented acoustic knowledge of hearing researchers. The result of such a collaboration would be a fuller, more integrated understanding of the acoustics of speech. The present study attempted to kindle such cross-fertilization, but due to the shortcomings of the present study's FFT-based Bark-phon and Bark-sone spectra as representations of the auditory periphery, it is clear that more work needs to be done before the present first step leads to a more solid investigation of the phonetics of plosive place of articulation from an auditory point of view. In particular, a more precise picture of the temporal structure of the burst would likely aid the correct classification of challenging cases such as the one in Figure 9.1. There is a richness of data in the small-scale temporal structure of acoustic events such as the burst that does not appear to have been adequately explored by most phonetic studies, and this neglect of temporal information in phonetics (and in ASR; Hermansky, 2011) can arguably be linked to the over-

reliance in the field on the Fourier transform and insufficient use of temporally richer, auditorily-oriented models (on which more shortly).

An objection I have sometimes encountered towards using auditory modelling in phonetics is that our knowledge of the brain's perceptual systems is incomplete whereas our conventional methods of acoustic measurement, though they may not incorporate our knowledge of auditory processes, nevertheless facilitate comparison with previous studies, given the constancy of acoustic methods utilized. I sympathize with this argument and it seems to me to have some merit. What I think is worth emphasizing, however, is that a plurality of approaches can complement each other. To give just one example out of many: it was shown in the literature review that many studies (e.g. Cole and Scott, 1974 a, b; Dorman et al., 1977; Stevens and Blumstein, 1978) sought to understand the perceptual role of the burst relative to the transitions. Their chosen method of investigation was the use of stimuli in which the burst and/or the transitions were removed. What was missing from all this discussion in phonetics (and what remains missing from a large part of the discussion to this day) is an awareness of the fact that forward masking is likely to be reducing the prominence of $F2_{onset}$ relative to the release burst.

Wang and Brown (2006: 5) note that broadband noises produce the strongest forward masking, and the release burst is perhaps the most broadband sound in speech there is. Yet discussion of how such auditory factors affect the relative importance of speech cues seems to be fairly rare in phonetic accounts. Two exceptions to this, as we have seen, are Cvengros (2011) and Xie (2013). Greater awareness of such auditory phenomena would thus advance phoneticians in their quest to determine the relative importance of phonetic cues in a given phonemic contrast.

To be specific, there is a need to quantify the amount of forward masking of $F2_{onset}$ caused by the burst in /b d g/. In the literature review, we saw that Xie (2013) has done this; however, due to the small scale of his study he was only able to do this for utterance-initial plosives in nonce monosyllables. The mixed-effects model in Section 6.3 showed that the amplitude of the burst in /b d g/ varies depending on the stress of the following vowel. That is, the burst is on average lower in amplitude in unstressed syllables. This would lead one to expect $F2_{onset}$ to play more of a role in such /b d g/ tokens because of masking being less. But because the focus has predominantly been on plosives in stressed monosyllables, the amount of forward masking of $F2_{onset}$ by the burst in *unstressed* syllables is unclear. What is also unclear is whether and to what extent $F2_{onset}$ would be masked less in prevoiced /b d g/ than devoiced /b d g/.

It was noted in the literature review (2.3.7) that Xie (2013) found the amount of masking caused by the burst varied depending on the burst's place of articulation. In particular, there is

reason to expect /b/ to mask $F2_{onset}$ less than other plosives, both because /b/ frequently lacks an audible release burst (Li et al., 2010) and because the release burst, when it is audible, tends to be lower in amplitude than that of the other places of articulation (Zue, 1976). All of these expectations need to be investigated on a much larger, quantitative scale than appears to have been done by previous studies. One hint that these expectations about masking in /b d g/ are not implausible is the dominance hypothesis that we saw in Figure 2.16 (Sussman, 1998: 257). Recall that this hypothesis posits that when $F2_{onset}$ is ambiguous in its place of articulation information, listeners will be biased to perceive /b/ in favour of /d/, and /d/ in favour of /g/. Given that the /b/ burst tends to be lower in amplitude than /d g/, it would be expected to mask $F2_{onset}$ less than /d g/, which would leave $F2_{onset}$ more salient than in /d g/.[10] Again, these ideas need further quantitative exploration with a proper auditory model.

One objection I have sometimes encountered in phonetic science to the use of auditory models is that no model is the same as human perception itself and since we do not fully understand speech and sound perception anyway, it might be unwise to abandon our existing acoustic models for some auditory model that could be replaced by a new model in a few years. Again, my suggestion is not to abandon anything, but rather to expand the range of possible approaches used in phonetics. It is true that we do not understand speech perception and auditory perception in as much detail as we would like. And it is also true that any model is *just* a model in the sense that it is not the same thing as the phenomenon it seeks to emulate: it also reflects the assumptions and decisions of the designer of the model.

Nevertheless, it is worth noting that Saremi et al. (2016) have examined the performance of seven auditory models on a variety of tasks (e.g. excitation patterns, frequency selectivity, signal-in-noise processing) and compared them to existing physiological data. Although all the models had strengths depending on the task, they found the CARFAC model to have the best overall performance, due to its outstanding agreement with experimental data and reasonably low computational cost (p. 1630). The design of this model has been detailed extensively in Lyon (2017) and has recently been implemented by Xu et al. (2018) on a field-programmable gate array for use in sound and speech recognition. Saremi et al.'s test of the model did, however, reveal two anomalies, which were traced to the model's inclusion of quadratic distortion (a psychoacoustic phenomenon), but Saremi and Lyon (2018) have recently shown that these anomalies can be removed by zeroing the model's quadratic-distortion parameter. The CARFAC model is freely available on GitHub for C++, MATLAB, and Python (Google,

[10] Rachel Smith (personal communication) points out that forward masking could also be used to explain why /g/ trumps /d/ in this hierarchy, namely that because the burst peak of /g/ is in the mid-frequency region, it is likely to generate more forward masking of $F2_{onset}$ than the burst peak of /d/ whose burst peak lies further away in the cascade of auditory filters from the F2 region.

2013) and in 2017 an implementation of the model in the form of a Jupyter notebook also became available (Van Schaik, 2017).

The take-home message for phoneticians is that there is now a freely available computational model of the cochlea that shows outstanding agreement with experimental results. This suggests the potential for future cross-disciplinary collaboration with auditory researchers to use this model to study the properties of particular speech sounds. The result of such work would be acoustic measurements that incorporate such important perceptual effects as forward masking and the upward spread of masking, and richer representations of the time-domain information in specific frequency channels of particular speech events such as the burst. This in turn could lead to more psychologically plausible theories in phonetics of the primary cues underpinning specific phonemic contrasts, theories that would be more adequately integrated with what is known about the auditory system.

Appendices

Appendix 1: Material

The following appendix contains the material that was presented to the speakers in the main study. The material consists of two parts: a sentence-reading task followed by a picture-description task. In the sentence-reading task there were 84 sentences in total, presented in the order given below. Each sentence was displayed on a separate slide in a slideshow. The break between lines 42 and 43 in the list below indicates the break given to speakers at the half-way point. The aim was to have as many plosives in each sentence without the sentences feeling like tongue twisters or otherwise artificial. As detailed in Chapter 4 (4.2.3), the material from the picture-description task was not utilized in the present study.

Task 1: Sentence Reading

1. Dave and Clarissa called at two and stayed until ten.
2. To me the guy with the black cap only looked about twenty-two.
3. Chris and Amber drove to Bordeaux during the Easter holidays.
4. Steph and Adele walked up above the cliff until they came to a wide stream.
5. The bone in Peter's left hand appeared to be fractured.
6. Provided we go soon, we should be there before late afternoon.
7. The best time Gwen had in Italy was when she and her cousin went to the colosseum.
8. Denise told me she's tired of seeing cat videos on Facebook.
9. Grace said it was definitely the hardest exam she's had to date.
10. Jackie played football for his local team until his mid-thirties.
11. Tamara asked Rita for a blue pen, some paper, and a few stickers.
12. Buying presents for his girlfriend was certainly not Brian's greatest talent.
13. The drip of the tap used to keep Abigail awake at night.
14. Tim's not a fan of rom-coms but he said that one was actually quite good.
15. Keith and Penny were not impressed by the secretary, who they found curt and abrupt.
16. Cal and Nikki bought two cute speckled puppies from Bernard.
17. I don't know for sure but I think it was in 1992 that the Olympics took place in Barcelona.
18. One of the dullest tasks the group faced was correcting the data in the Excel file.
19. It didn't occur to me that I had left without paying until I saw the stern look on the security guard.
20. Gail and Spencer had been jogging for two miles when they had to stop for lack of water.
21. The car stalled on the roundabout while Bill was changing gear.
22. Mandy said the scariest part of the roller coaster was the sudden dip at the end.
23. Scott took a photo but the grove of tall trees obscured the view of the dales.
24. Seb and Zac are two of the best pool players in the north east, and both have won prizes.

25. Dora looked up the term in both the index and the glossary but couldn't find it.
26. The baby cooed contentedly in the pram as he looked up at the beautiful blue sky.
27. Brixton was embarrassed when nobody wanted to clap at the end of his talk.
28. Rob logged in by guessing the password correctly, a pure fluke.
29. Luckily, Ken and Pam came back with sandbags before the flooding started.
30. The solar eclipse was one of the freakiest things Glenn had ever witnessed.
31. Derek lobbed the rugby ball to me but it was slippery and slid out of my grasp.
32. Caroline and a few of her co-workers complained the tyrannical supervisor to their boss.
33. Gary was on the cusp of tripping over the box when I grabbed him by the shoulder.
34. Adam said his chat with the pompous principal was quite possibly the most awkward of his life.
35. Megan caught wind of the story and was absolutely livid.
36. Kyle and Andy were reputed to be two of the toughest kids in the entire school.
37. Steve and Karen were angry when the tickets expired before they could make use of them.
38. Mr Tilbury asked all six of the accused pupils to stay after class.
39. The bouncer at the nightclub was staring at Dale with a sinister grin.
40. It was about ten past two when I heard a knock on the door from the police.
41. Doing homework at twelve o'clock the night before is definitely not the brightest idea.
42. Ted was annoyed that Stan and Claire turned up at his birthday party forty minutes late.

43. The leg of the girl's doll broke off in the playground.
44. By the time we came back the snow was already eight inches deep on the footpath.
45. Maggie said she was astounded that the firefighters escaped from the blaze unhurt.
46. Thinking back on it now, there was probably a good reason why Deane avoided Glenda like that.
47. Deriving equations from scratch is probably one of the most tedious and difficult jobs imaginable.
48. Bethany's counsellor told her to relax and not to dwell on it.
49. Some people are saying this is the stormiest winter on record but such claims are old hat to me.
50. Mr Kennedy's shop was burgled recently even though he had locked all the doors and windows.
51. Brad felt a rush of blood to the head as the pony began to gallop wildly.
52. Margaret thought the modernist art in the gallery was far too pretentious for her tastes.
53. Tony's neighbours were fed up with the constant barking of his hungry dogs.
54. Betty said her daughter is getting good grades despite being absent for three weeks.
55. Despite their best efforts, Paul and Cassie weren't able to pull the bike out of the muddy barn.
56. Tina is going to host a surprise birthday bash for Kevin this Saturday.
57. Some people say Kelly is snobbish but I think that's just nasty gossip.
58. Fred untied the rope on the gate and two big Labradors sprinted up excitedly.
59. Darren bought a green A4 pad and some pencils before the sketching trip.

60. Bernadette was curious why I hadn't attended music class in a fortnight, but I avoided answering.
61. When Blake looked in his magnifying glass, he could see the little beetles scuttling away.
62. Only twice in December did I get the time to go Christmas shopping.
63. Every Tuesday at two, Mikey taught the boys and girls art.
64. Doug and Cathy went to collect the newspaper and groceries.
65. Eddie used to grow carrots, parsnips, turnips, cabbages and cauliflowers.
66. The wedding didn't quite go to plan, as the marquee's canopy collapsed from the wind.
67. Teaching trigonometry was not a simple task for Agnes, and she had to re-study it beforehand.
68. Tiffany and Gladys cycled to Durham last Tuesday. It took them over two hours.
69. Texting in class has to be one of the most annoying habits for any teacher to deal with.
70. Given her poor grades, Meg felt carpentry was not a skill she was likely to improve.
71. The blind cat uses his whiskers and keen sense of smell to walk around.
72. Tara was bored to death by the grand prix and was glad there were only two more laps left.
73. Brenda said the creative writing class was terrific and well worth the time and effort.
74. During his gap year, Percy volunteered to work for a charity in Cambodia and found it rewarding.
75. Tom said *The Secret* is probably the worst book he's read in his entire life.
76. Pippa says her daughter is as daft as a brush but that's acceptable at her age.
77. Declan said winning his first golf tournament was the happiest day of his career.
78. Nicole and Tanya are thinking of going bowling this weekend; would you like to come too?
79. I doubt we'll see robots that can drive cars or play sports in my lifetime.
80. We asked him if being a night-time security guard was boring. He nodded his head vigorously.
81. Ben brought Keelan to the doctor, who prescribed her some painkillers for her backache.
82. I saw the weirdest bird today: it had a black bill, green and blue stripes, and the loudest squawk.
83. Luke did not appear to be a promising apprentice as he struggled to pay attention for long.
84. I must say the back garden looks better than ever this spring.

Task 2: Picture Description

Picture 1:

Picture 2:

Picture 3:

Picture 4:

Picture 5:

Appendix 2: Key to Transcription System

As mentioned in Chapter 4, the allophone tier of the TextGrid consists of a narrow phonetic transcription. To type this tier rapidly, certain substitutions of the IPA symbols were made, as detailed in the following table.

Symbol	Meaning	IPA symbol			
£	devoiced	◌̥	A	open back unrounded	ɑ
{	lefthand word boundary	[none]	D	voiced dental fricative	ð
}	righthand word boundary	[none]	DZ	voiced postalveolar affricate	dʒ
)	unreleased	◌̚			
~	primary stress	ˈ	E	open-mid front unrounded	ɛ
&	secondary stress	ˌ	N	voiced velar nasal	ŋ
$	syllabic	◌̩			
’	ejective	’	O	open-mid back rounded	ɔ
”	velarization	~			
€	approximant	[none]	r+	voiced labiodental approximant	ʋ
?	voiceless glottal plosive	ʔ			
%	voiced	◌̬	S	voiceless postalveolar fricative	ʃ
#	preceding or following pause	[none]			
,	lowered	◌̞	T	voiceless dental fricative	θ
-	retraction	◌̠	TS	voiceless postalveolar affricate	tʃ
+	advanced	◌̟			
=	unaspirated	[none]			
:	centralization	◌̈	Y	close-mid central (un)rounded	ɘ
*	affricated	◌ˢ			
^	open-mid central unrounded	ɜ	Z	voiced postalveolar fricative	ʒ
@	mid central	ə			

Some of the symbols used above turned out to cause difficulties in the Excel file, e.g. “?” cannot be found with the search facility. As a result some of the symbols were changed after the data had been extracted into Excel. These were as follows:

TextGrid	**Excel**
?	q
?p ?t ?k	cp ct ck
£	h
€	¬

Appendix 3: Comment Tier

As mentioned in Chapter 4 (4.3.4), the comment tier was used for a variety of purposes, including to describe various observed properties of the stops and surrounding segments. Most of these labels are short for speed of transcription and so need further elaboration.

Comment	**Meaning**
bimodal	Indicates a velar burst spectrum in which a peak appears in the high-frequency region of the Bark-phon image whose amplitude is equal or slightly greater than the mid-frequency ('typical') velar peak
diffuse	Indicates an alveolar burst spectrum in which there is a lack of a clear peak in the high-frequency region. Instead the amplitude appears to be approximately the same as that of the mid-frequency region.
fairlydiffuse	Indicates an alveolar burst spectrum in which there is a peak in the high-frequency region (as is typical for an alveolar) but the peak nevertheless appears to be less prominent than is typical for alveolars.
fairlyhigh	Indicates a velar spectrum whose burst peak is relatively high, potentially high enough to be misclassified as alveolar. The following segment in such cases was typically /j i/.
fairlylow	Indicates an alveolar spectrum whose burst peak is relatively low, potentially low enough to be misclassified as velar.
fairlyweak	Indicates a burst that was quieter than typical for its place of articulation.
good	Used to indicate that the burst of a velar or alveolar appeared typical (from spectrographic inspections and sometimes also a Bark-phon slice) for its place of articulation. In the case of alveolars this is defined as the peak being above ca. 3,500 Hz. In the case of velars it is defined as the peak being reasonably similar in frequency to the onset F2 of the following segment (in the case of /w r/ and back vowels) or the onset F3 of the following segment (in the case of /j/ and front vowels).
goodbilabial	Used to indicate that the burst of a bilabial was audible and that the shape of the burst (from spectrographic inspection and sometimes also a Bark-phon slice) appeared to be typical for a bilabial in lacking a well-defined peak.
goodbimodal	Indicates a velar burst spectrum in which a peak appears in the high-frequency region of the Bark-phon spectral slice whose amplitude is almost as great as the one in the mid-frequency ('typical') velar peak

veryweak	The burst, though visible on the spectrogram, does not seem to be audible. The audibility of a burst was determined by selecting the syllable with the burst, listening to it, and then selecting the syllable without the burst and listening again and trying to hear if there was a difference. Most such cases were bilabial.
weak	Indicates a burst that seemed to be audible but nevertheless appeared low in amplitude. Most such cases were bilabial.

As detailed in Chapter 4 (4.3.4), the comment tier was also used to indicate on a five-point scale how visible F2 and F3 were in the burst and aspiration on the spectrogram.

tfloud	'third formant loud' – the highest rating for the visibility of F3 on the spectrogram. F3 appeared as a dark streak.
htf	'has third formant' – the second highest rating for the visibility of F3 on the spectrogram. F3 appeared to be unambiguously present, though not as dark in colour as tfloud.
ptf	'probably has third formant' – third highest rating for the visibility of F3 on the spectrogram. F3 appeared to be present but did not stand out sharply enough from the surrounding frequencies to say this with certainty.
ktf	'kind of has third formant' – second lowest rating for the visibility of F3 on the spectrogram. There appeared to be some formant-like pattern in the F3 region but too ambiguous to be counted.
ntf	'no third formant' – the lowest rating for the visibility of F3 on the spectrogram. There appeared to be no formant-like pattern in the F3 region.

And here are the equivalent abbreviations and rankings for F2:

sfloud	'second formant loud' – the highest rating for the visibility of F2 on the spectrogram. F2 appeared as a dark streak.
sf	'has second formant' – the second highest rating for the visibility of F2 on the spectrogram. F2 appeared to be unambiguously present, though not as dark in colour as sfloud.
psf	'probably has second formant' – second highest rating for the visibility of F2 on the spectrogram. F2 appeared to be present but did not stand out sharply enough from the surrounding frequencies to say this with certainty.

ksf	‘kind of has second formant’ – second lowest rating for the visibility of F2 on the spectrogram. There appeared to be some formant-like pattern in the F2 region but too ambiguous to be counted.
nsf	‘no second formant’ – the lowest rating for the visibility of F2 on the spectrogram. There appeared to be no formant-like pattern in the F2 region.

Appendix 4: Transcription

The phonetic and phonemic annotation were outlined briefly above. Vowels are of secondary importance in the present study since this is a study of plosives' place of articulation. As a result, the important aspect of the vowel in the present study is its backness, since it is the F2 transition (and F3) that matters most for place of articulation. Furthermore, vowel F2 was treated in parts of the present study (e.g. Section 5.4) as a continuous variable, which bypassed the necessity of relying on an observer's transcriptions.

Nevertheless, the transcription system used in the present study warrants some further comments. The IPA vowel chart contains no less than 28 vowel symbols. When the diacritics ˖ ˗ ˔ ˕ ¨ are added to the list the result is a system with a large number of choices for how to transcribe vowels. For example, a vowel between Cardinal Vowels 3 and 4 can be transcribed as [ɛ̝ æ ɐ̟] with only modest (if any) audible difference in quality between the three. The result is that the same vowel sound can be transcribed in multiple ways. Another example is that a vowel centralized from Cardinal 2 can be transcribed as [ɪ̟] or [ë] or [ɵ̟] or [ɪ̈].

It was decided to mitigate such vagaries beforehand by restricting the number of ways of narrowly transcribing the vowels by avoiding the use of certain vowel letters and diacritics. Figure 1 shows the system followed.

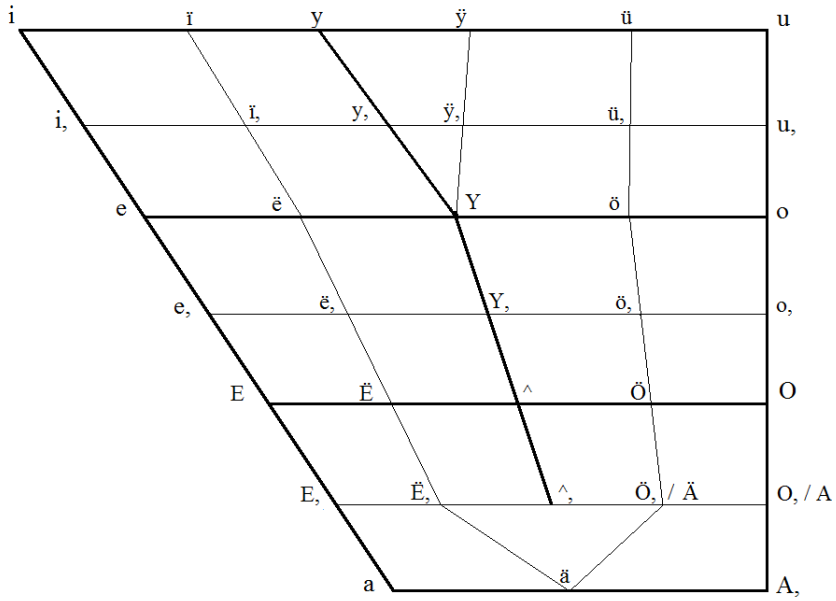


Figure 1: Schematic illustration of the narrow transcription for vowels used in the present study.

When vowels are being analysed by listening to a recording rather than looking at a speaker, only auditory criteria are available to guide the listener. It is unclear how one could reliably distinguish some of the symbols used on the full IPA chart without also having visual

criteria. For example, how is a researcher supposed to distinguish between [œ], [ɞ], and [ɜ] based on auditory criteria alone? The same can be said of [ø], [ɵ], and [ɘ]. In any case, is there any human language that phonemically contrasts six open-mid vowels? Or six close-mid vowels? Or six close vowels for that matter? If a difference between two vowel timbres of the same height or backness is not phonemic in any human language, as appears to be the case with [ø ɵ] and [œ ɞ], would the difference between the two timbres (assuming it is audible) not be better represented using a diacritic, since one principle of the IPA is to use separate letters for phonemic contrasts and diacritics for non-phonemic ones? This is why Figure 1 utilizes just 11 of the IPA letters. (Schwa, which is not on the chart, was also utilized, yielding a total of 12 out of 28 symbols utilized.)

The guiding principle was to only have as many vowel letters as is warranted by cross-linguistic phonemic contrasts. For example, the largest number of open-mid vowels phonemically contrasted by a human language appears to be three, e.g. French (Fougeron and Smith, 1999: 78). Therefore, only three vowel symbols were utilized for this dimension: [E] for a vowel that sounds front, [O] for back, and [^] for vowel qualities that are approximately halfway between [E] and [O], i.e. what would conventionally be transcribed as [ɜ]. The largest number of fully open vowels contrasted by a language is two. Therefore, only two symbols were used for this dimension: [a] for front, [A] for back. (As Lindsey (2012) has pointed out, the current RP realization of the LOT vowel is not fully open, being more similar to an open-mid vowel in terms of its F1 value. The 20 speakers in the present study usually realized it as such, hence it was transcribed [O] or [O,].)

Admittedly the current approach (in its incarnation in the present study for English) does not follow this cross-linguistic typological principle as rigorously as it could do. For example, the maximum number of close vowels contrasted phonemically in the world's languages is four, e.g. Swedish (Engstrand, 1999: 140), whereas the present system only has three letters for this dimension ([i] for front, [u] for back, and [y] for the intermediate range – which is where the GOOSE vowel for most speakers in the present study is realized). Similarly, there is evidence from Korean (Bok Lee, 1999: 121) suggesting that up to four degrees of phonemic contrast along the front-back dimension is possible among close-mid vowels – if correct this would mean that the present study's approach is lacking a symbol (the system is [e] for front unrounded, [o] for back, [Y] for intermediate qualities such as front rounded and central rounded/unrounded). For the purposes of transcribing English, however, this is not important.

In terms of choice of symbols, the use of ':' to represent vowel centralization rather than what it ordinarily represents – lengthening – runs the risk of causing confusion for those not familiar with the key (see Appendix 3). This usage was followed for the sake of speed of typing

and is an ad-hoc choice specific to the present study of plosives, in which the transcription of vowel length was not relevant. Naturally for a study in which the transcription of vowel duration *is* necessary, a different quick keyboard symbol for vowel centralization would have to be found instead.

One final limitation of the approach in its current guise concerns '[O,]' and '[A]': I cannot be confident that I was able to reliably distinguish the two qualities in this pair, so they should be taken as synonymous. Wells (1982) transcribes mainstream RP PALM-BATH-START as [ɑ̟], indicating his view that this vowel quality is not as back as Cardinal 5.

Nevertheless, the overarching approach seems principled: use only as many vowel letters as is justified by the cross-linguistic evidence on the maximum number of phonemic contrasts in a given degree of vowel height. Adoption of such an approach in the classroom would probably also spare the student from much of the confusion caused by the raft of transcription choices available in the full IPA chart (e.g. 'how am I supposed to hear the difference between [œ] and [ɜ] without looking at the speaker's lips?').

Another principle followed in the current transcription was not to apply one's phonetic abilities beyond what they can reliably deliver. For example, it is certainly true that one's auditory system can hear more than one vowel timbre between Cardinals 2 and 3. However, I do not believe it possible for a phonetician to transcribe consistently the difference between [e̞] and [ɛ̝]. Therefore in the present study's transcription there was only one permitted way of transcribing a front-vowel quality that sounds different from both Cardinal 2 and Cardinal 3, viz. [e̞] (which was typed as 'e,'). This convention is followed even when the vowel in question might sound somewhat closer to Cardinal 3 than Cardinal 2 – the point is that if a vowel quality is sufficiently different from both Cardinals 2 and 3 to be worth transcribing differently, then there should be only one way of doing so. The alternative is to invite inconsistency of the kind where a researcher transcribes a vowel quality as [e̞] on one occasion and [ɛ̝] on another occasion, even though the difference in quality is trivially small or non-existent. And if a vowel feels much closer to Cardinal 3 than Cardinal 2, then the wisest thing seems to be to transcribe it without any diacritics, viz. [ɛ].

Although the above discussion focused on the difference between Cardinals 2 and 3, the same point can be made about any continuum between two vowels in the closeness dimension.

Here is a table showing typical (narrow) transcriptions of f03's vowels, a speaker of RP, using Wells's (1982) lexical sets:

KIT	e:	FLEECE	i	PRICE	a:e,
DRESS	E	FACE	e,i,	CHOICE	o,e
TRAP	a	PALM	A,	MOUTH	A,o,
LOT	O	THOUGHT	o	*pool*	u
STRUT	^	GOAT	Y,y,	*goal*	o,u,
FOOT	Y	GOOSE	y	*comm*A	@

Table 1: Some typical narrow transcriptions of vowels from an RP speaker in the present study.

References

Abdelatty Ali, A. M., Van Der Spiegel, J., & Mueller, P. (2001). Acoustic-phonetic features for the automatic classification of stop consonants. *IEEE Transactions on Speech and Audio Processing*, *9*(8), 833–841.

Adank, P., Smits, R., & Van Hout, R. (2004). A comparison of vowel normalization procedures for language variation research. *Journal of the Acoustical Society of America*, *116*(5), 3099–3107.

Addison, P. S. (2002). *The Illustrated Wavelet Transform Handbook: Introductory Theory and Applications in Science, Engineering, Medicine and Finance*. Abingdon: Taylor & Francis.

Adelson, E. H. (1995). Checker-Shadow Illusion. Retrieved August 27, 2018, from http://persci.mit.edu/gallery/checkershadow.

Al-Tamimi, J. (2004). L'équation du locus comme mesure de la coarticulation VC et CV : Étude préliminaire en Arabe Dialectal Jordanien. In *Actes 25èmes Journées d'Études sur la Parole, Fès* (pp. 9–12). Fez, Morocco.

Al-Tamimi, J. (2007). *Indices dynamiques et perception des voyelles: Étude translinguistique en arabe dialectal et en français*. PhD thesis, Université Lumiere Lyon 2.

Al-Tamimi, J. (2017). Revisiting acoustic correlates of pharyngealization in Jordanian and Moroccan Arabic: Implications for formal representations. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, *8*(1), 1–40.

Allen, J. B. (2005). *Articulation and Intelligibility*. San Rafael: Morgan & Claypool.

Allen, J. B. (2005). Consonant recognition and the articulation index. *Journal of the Acoustical Society of America*, *117*(4 Pt 1), 2212–2223.

Allen, J. B, & Han, W. (2011). Sources of decoding errors of the perceptual cues, in normal and hearing impaired ears. *Speech Perception and Auditory Disorders*, 495–508.

Alwan, A. (1992). Modeling speech perception in noise: The stop consonants as a case study. *RLE Technical Report*, (569), 1–136.

Alwan, A., Jiang, J., & Chen, W. (2011). Perception of place of articulation for plosives and fricatives in noise. *Speech Communication*, *53*(2), 195–209.

Baken, R. J., & Orlikoff, R. F. (2000). *Clinical Measurement of Speech and Voice* (2nd ed.). San Diego, CA: Singular Thomson Learning.

Bakran, J., & Mildner, V. (1995). Effect of speech rate and coarticulation strategies on the locus equation determination. In *Proceedings of the XIII International Congress of Phonetic Sciences* (pp. 26–29).

Bates, D., Maechler, M., Bolker, B., & Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Bellman, R. E. (1961). *Adaptive Control Processes: A Guided Tour*. Princeton: Princeton University Press.

Blumstein, S. E., & Stevens, K. N. (1979). Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants. *Journal of the Acoustical Society of America*, *66*(4), 1001–1017.

Boersma, P. (1998). *Functional Phonology*. The Hague: Holland Academic Graphics.

Boersma, P., & Weenink, D. (2014). Praat: doing phonetics by computer. Retrieved February 15, 2015, from http://praat.org.

Bok Lee, H. (1999). Korean. In *Handbook of the International Phonetic Association* (pp. 120–123). Cambridge: Cambridge University Press.

Brancazio, L. (1998). Linear correlates in the speech signal: The orderly output constraint. *Behavioral and Brain Sciences*, *21*, 261.

Brancazio, L., & Fowler, C. A. (1998). On the relevance of locus equations for production and perception of stop consonants. *Perception & Psychophysics*, *60*(1), 24–50.

Bregman, A. (1990). *Auditory Scene Analysis: The Perceptual Organization of Sound*. Cambridge, MA: MIT Press.

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32.

Carroll, J. D., & Wish, M. (1974). Models and methods for three-way multidimensional scaling. In D. H. Krantz, R. C. Atkinson, R. D. Luce, & P. Suppes (Eds.), *Contemporary Developments in Mathematical Psychology* (pp. 57–105). San Francisco: Freeman.

Chennoukh, S., Carré, R., & Lindblom, B. (1997). Locus equations in the light of articulatory modeling. *Journal of the Acoustical Society of America*, *102*(4), 2380–2389.

Clark, J. E., Yallop, C., & Fletcher, J. (2007). *An Introduction to Phonetics and Phonology* (3rd ed.). Oxford: Blackwell.

Cole, R A, & Scott, B. (1974). Toward a theory of speech perception. *Psychological Review*, *81*(4), 348–374.

Cole, Ronald A., & Scott, B. (1974). The phantom in the phoneme: Invariant cues for stop consonants. *Perception & Psychophysics*, *15*(1), 101–107.

Cooper, F. S., Liberman, A. M., & Borst, J. M. (1951). The interconversion of audible and visible patterns as a basis for research in the perception of speech. *Proceedings of the National Academy of Sciences, USA*, *37*(5), 318–325.

Cramer, D., & Howitt, D. (2004). *The SAGE Dictionary of Statistics*. London: SAGE.

Crowther, C. S. (1994). Modelling coarticulation and place of articulation using locus

equations. *UCLA Working Papers in Phonetics*, *88*, 127–148.

Cvengros, R. M. (2011). *A Verification Experiment of the Second Formant Transition Feature as a Perceptual Cue in Natural Speech*. Msc thesis, University of Illinois at Urbana-Champaign.

Dannemiller, J. L. (1989). Computational approaches to color constancy: Adaptive and ontogenetic considerations. *Psychological Review*, *96*(2), 255–266.

Dannemiller, J. L. (1998). Color constancy and color vision during infancy: Methodological and empirical issues. In V. Walsh & J. J. Kulikowski (Eds.), *Perceptual Constancy* (pp. 229–261). Cambridge: Cambridge University Press.

Davidson, L. (2016). Variability in the Implementation of Voicing in American English Obstruents. *Journal of Phonetics*, *54*, 35–50.

Davis, S. B., & Mermelstein, P. (1980). Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *28*(4), 357–366.

Delattre, P. C., Liberman, A. M., & Cooper, F. S. (1955). Acoustic loci and transitional cues for consonants. *Journal of the Acoustical Society of America*, *27*(4), 769–773.

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised learning algorithms. *Neural Computation*, *10*(7), 1895–1923.

Docherty, G. J. (1992). *The Timing of Voicing in British English Obstruents*. Berlin: Walter de Gruyter.

Dorman, M. F., Studdert-Kennedy, M., & Raphael, L. J. (1977). Stop-consonant recognition: Release bursts and formant transitions as functionally equivalent, context-dependent cues. *Perception*, *22*(2), 109–122.

Duez, D. (1992). Second-formant locus patterns: An investigation of spontaneous French speech. *Speech Communication*, *11*, 417–427.

Duifhuis, H. (1973). Consequences of peripheral frequency selectivity for nonsimultaneous masking. *Journal of the Acoustical Society of America*, *54*, 1471–1488.

Engstrand, O. (1999). Swedish. In *Handbook of the International Phonetic Association* (pp. 140–142). Cambridge: Cambridge University Press.

Fagerland, M. W., Lydersen, S., & Laake, P. (2013a). The McNemar test for binary matched-pairs data: mid-p and asymptotic are better than exact conditional. *BMC Medical Research Methodology*, *13*(91), 1–8.

Fagerland, M. W., Lydersen, S., & Laake, P. (2013b). How to calculate the McNemar mid-p test (supplementary materials). *BMC Medical Research Methodology*, *13*(91), 1–8.

Fant, G. (1960). *Acoustic Theory of Speech Production*. The Hague: Mouton.

Fant, G. (1973). *Speech Sounds and Features*. Cambridge, MA: MIT Press.

Fant, G. (1975). Non-uniform vowel normalization. *Speech Transmission Laboratory Quarterly Progress and Status Reports*, *16*, 1–19.

Fant, G, Ishizaka, K., Lindqvist-Gauffin, J., & Sundberg, J. (1972). Subglottal Formants. *Speech Transmission Laboratory Quarterly Progress and Status Reports*, *13*(1), 1–12.

Fastl, H., & Zwicker, E. (2007). *Psychoacoustics: Facts and Models* (3rd ed.). Berlin: Springer.

Fischer-Jorgensen, E. (1954). Acoustic analysis of stop consonants. *Miscellanea Phonetica*, *2*, 42–59.

Fitch, W. T., & Hauser, M. D. (1998). Linear correlates in the speech signal: The orderly output constraint. *Behavioral and Brain Sciences*, *21*(2), 264–265.

Fletcher, H., & Munson, W. A. (1933). Loudness, Its Definition, Measurement and Calculation. *Journal of the Acoustical Society of America*, *5*(2), 82–108. https://doi.org/10.1121/1.1915637

Flynn, N. E. J. (2012). *A Sociophonetic Study of Nottingham Speakers*. University of York.

Forrest, K., Weismer, G., Milenkovic, P., & Dougall, R. (1988). Statistical analysis of word-initial voiceless obstruents: Preliminary data. *Journal of the Acoustical Society of America*, *84*, 115–123.

Fougeron, C., & Smith, C. L. (1999). French. In *Handbook of the International Phonetic Association* (pp. 78–81). Cambridge: Cambridge University Press.

Foulkes, P., Docherty, G., & Jones, M. J. (2011). Analyzing Stops. In *Sociophonetics: A Student's Guide* (pp. 58–71). Abingdon: Routledge.

Fowler, C. A. (1990). Calling a mirage a mirage: direct perception of speech produced without a tongue. *Journal of Phonetics*, *18*, 529–541.

Fowler, C. A. (1994). Invariants, specifiers, cues: An investigation of locus equations as informaion for place of articulation, *55*(6), 597–610.

Fowler, C. A. (1996). Listeners do hear sounds, not tongues. *Journal of the Acoustical Society of America*, *99*(3), 1730–1741. https://doi.org/10.1121/1.415237

Fowler, C. A. (1998). Linear correlates in the speech signal: The orderly output constrain. *Behavioral and Brain Sciences*, *21*, 265–266.

Fruchter, D., & Sussman, H. M. (1997). The perceptual relevance of locus equations. *J Acoust Soc Am*, 2997–3008.

Gerstman, L. J. (1968). Classification of self-normalized vowels. *IEEE Transactions on Audio and Electroacoustics*, *AU-16*(1), 78–80.

Glasberg, B. R., & Moore, B. C. J. (1990). Derivation of auditory filter shapes from notched-

noise data. *Hearing Research*, *47*(1–2), 103–138.

Gobl, C., & Ní Chasaide, A. (2010). Voice source variation and its communicative functions. In W. J. Hardcastle, J. Laver, & F. E. Gibbon (Eds.), *The Handbook of Phonetic Sciences* (2nd ed., pp. 378–423). Chichester: Wiley-Blackwell.

Google. (2013). The Cascade of Asymmetric Resonators with Fast-Acting Compression. Website: https://github.com/google/carfac (retrieved August 22, 2018).

Greenberg, J. H. (1978). Some Generalizations Concerning Initial and Final Consonant Clusters. In J. H. Greenberg, C. A. Ferguson, & E. A. Moravcsik (Eds.), *Universals of Human Language: Volume 2* (pp. 243–280). Stanford, CA: Stanford University Press.

Greenwood, D. D. (1961). Critical Bandwidth and the Frequency Coordinates of the Basilar Membrane. *Journal of the Acoustical Society of America*, *33*(10), 1344–1356.

Halle, M., Hughes, G. W., & Radley, J. -P. A. (1957). Acoustic properties of stop consonants. *Journal of the Acoustical Society of America*, *29*(1), 107–116.

Harrington, J. (2010). *Phonetic Analysis of Speech Corpora*. Chichester: Wiley-Blackwell.

Harrington, J., & Cassidy, S. (1999). *Techniques in Speech Acoustics*. Dordrecht: Springer.

Harris, J. (1994). *English Sound Structure*. Oxford: Blackwell.

Harrison, P. T. (2013). *Making Accurate Formant Measurements: An Empirical Investigation of the Influence of Measurement Tool, Analysis Settings and Speaker on Formant Measurements*. PhD thesis, University of York.

Hasegawa-Johnson, M. A. (1996). *Formant and Burst Spectral Measurements with Quantitative Error Models for Speech Sound Classification*. PhD thesis, Massachusetts Institute of Technology.

Hedrick, M., & Younger, M. (2007). Perceptual weighting of stop consonant cues by normal and impaired listeners in reverberation versus noise. *Journal of Speech, Language, and Hearing Research*, *50*(2), 254–269.

Hermansky, H. (1990). Perceptual linear predictive (PLP) analysis of speech. *Journal of the Acoustical Society of America*, *87*(4), 1738–1752.

Hermansky, H. (2011). Speech recognition from spectral dynamics. *Sadhana - Academy Proceedings in Engineering Sciences*, *36*(5), 729–744.

Hothorn, T., Bühlmann, P., Dudoit, S., Molinaro, A., & Van Der Laan, M. J. (2006). Survival ensembles. *Biostatistics*, *7*(3), 355–373.

Houde, R. (1967). *A study of tongue motion during selected speech sounds*. Speech Communication Research Laboratory, Santa Barbara.

Huckvale, M. (1996). Learning from the experience of building automatic speech recognition systems. *UCL Working Papers in Speech, Hearing and Language*, *9*, 1–14.

Huckvale, M. (2013). An introduction to phonetic technology. In *The Bloomsbury Companion to Phonetics* (pp. 208–226). London: Bloomsbury Academic.

IBM. (2013). Statistical Package for the Social Sciences. Armonk, NY: IBM Corporation.

ISO226. (2003). Acoustics - Normal Equal-Loudness Contours. Geneva: International Organization for Standardization.

Jakobson, R., Fant, G., & Halle, M. (1952). *Preliminaries to speech analysis: The distinctive features and their correlates*. Cambridge, MA: MIT Press.

James, G., Witen, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R*. London: Springer.

Jannedy, S., & Weirich, M. (2017). Spectral moments vs discrete cosine transformation coefficients: Evaluation of acoustic measures distinguishing two merging German fricatives. *Journal of the Acoustical Society of America, 142*(1), 395–405.

Johnson, K. (2012). *Acoustic and Auditory Phonetics* (3rd ed.). Malden, MA: Wiley-Blackwell.

Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika, 32*, 241–254.

Johnson, A. A., Reidy, P. F., and Edwards, J. R. (2018). Quantifying robustness of the /t/-/k/ contrast using a single, static spectral feature. *Journal of the Acoustical Society of America, 144*(2), 105-111.

Kapoor, A. (2010). *Perceptual Effects of Plosive Feature Modification*. MSc thesis, University of Illinois at Urbana-Champaign.

Kapoor, A., & Allen, J. B. (2012). Perceptual effects of plosive feature modification. *Journal of the Acoustical Society of America*, *131*(1), 478–491.

Kewley-Port, D. (1984). Time-varying features as correlates of place of articulation in stop consonants. *Journal of the Acoustical Society of America*, *73*(1), 322–335.

Kewley-Port, Diane, & Zheng, Y. (1999). Vowel formant discrimination: Towards more ordinary listening conditions. *Journal of the Acoustical Society of America*, *106*(5), 2945–2958.

Kingsbury, B. (2009). Automatic Speech Recognition: A Whirlwind Tour. Retrieved August 26, 2018, from https://vimeo.com/38843119.

Koenig, L. L., Shadle, C. H., Preston, J. L., Christine, R., & Campus, B. (2013). Toward improved spectral measures of /s/: Results from adolescents. *Journal of Speech Language, and Hearing Research*, *56*(4), 1175–1189.

Krull, D. (1987). Second formant locus patterns as a measure of consonant-vowel coarticulation. *Phonetic Experimental Research at the Institute of Linguistics, Stockholm University, V*, 43–61.

Krull, D. (1989). Second-formant locus patterns and consonant-vowel coarticulation in spontaneous speech. *Phonetic Experimental Research at the Institute of Linguistics, Stockholm University*, *X*, 87–101.

Ladefoged, P. (1996). *Elements of Acoustic Phoneticcs* (2nd ed.). Chicago: Chicago University Press.

Ladefoged, P. (2003). *Phonetic Data Analysis: An Introduction to Fieldwork and Instrumental Techniques*. Oxford: Blackwell.

Ladefoged, P., & Johnson, K. (2011). *A Course in Phonetics* (6th ed.). Boston, MA: Cengage Learning.

Ladefoged, P., & Maddieson, I. (1996). *The Sounds of the World's Languages*. Oxford: Blackwell.

Lahiri, A., Gewirth, L., & Blumstein, S. E. (1984). A reconsideration of acoustic invariance for place of articulation in diffuse stop consonants: Evidence from a cross-language study. *Journal of the Acoustical Society of America*, *76*(2), 391–404.

Laver, J. (1980). *The Phonetic Description of Voice Quality*. Cambridge: Cambridge University Press.

Laver, J. (1994). *Principles of Phonetics*. Cambridge: Cambridge University Press.

Lehiste, I., & Peterson, G. E. (1961). Transitions, Glides, and Diphthongs. *Journal of the Acoustical Society of America*, *33*(3), 268–277.

Li, F., Menon, A., & Allen, J. B. (2010). A psychoacoustic method to find the perceptual cues of stop consonants in natural speech. *Journal of the Acoustical Society of America*, *127*(4), 2599–2610.

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R News*, *2*(3), 18–22.

Liberman, A. M. (1996). *Speech: A Special Code*. Cambridge, MA: MIT Press.

Liberman, A. M., Delattre, P. C., Cooper, F. S., & Gerstman, L. J. (1954). The role of consonant-vowel transitions in the perception of the stop and nasal consonants. *Psychological Monographs*, *68*(8), 1–13.

Liberman, A. M., Delattre, P., & Cooper, F. S. (1952). The role of selected stimulus variables in the perception of unvoiced stop consonants. *American Journal of Psychology*, *65*(4), 497–516.

Liberman, A., & Mattingly, I. (1985). The Motor Theory of Speech Perception Revisited. *Cognition*, *21*, 1–36.

Lindblom, B. E. F., & Sundberg, J. E. F. (1971). Acoustical Consequences of Lip, Tongue, Jaw, and Larynx Movement. *Journal of the Acoustical Society of America*, *50*(4B),

1166–1179.

Lindblom, B. (1963). Spectroraphic Study of Vowel Reduction. *Journal of the Acoustical Society of America*, *35*(11), 1773–1781.

Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H&H theory. In *Speech Production and Speech Modelling* (pp. 403–439).

Lindblom, B. (1983). Economy of speech gestures. In P. F. McNeilage (Ed.), *The Production of Speech* (pp. 217–245). New York: Springer-Verlag.

Lindblom, B. (1996). Role of articulation in speech perception: Clues from production. *Journal of the Acoustical Society of America*, *99*(3), 1683–1692.

Lindblom, Björn, & Sussman, H. M. (2012). Dissecting coarticulation: How locus equations happen. *Journal of Phonetics*, *40*(1), 1–19.

Lindsey, G. (2012). The British English Vowel System. Retrieved August 26, 2018, from http://englishspeechservices.com/blog/british-vowels/.

Lisker, L., & Abramson, A. (1964). A Cross-Language Study of Voicing in Initial Stops: Acoustical Measurements. *Word*, *20*(3), 384–422.

Lobanov, B. M. (1971). Classification of Russian Vowels Spoken by Different Speakers. *Journal of the Acoustical Society of America*, *49*, 606–608.

Lobdell, B., & Allen, J. B. (2006). An information theoretic tool for investigating speech perception. In *Ninth International Conference on Spoken Language Processing* (pp. 853–856).

Lovitt, A., & Allen, J. (2006). 50 Years Late: Repeating Miller-Nicely 1955. *Ninth International Conference on Spoken Language Processing*, 2154–2157.

Lyon, R. F. (2017). *Human and Machine Hearing*. Cambridge: Cambridge University Press.

Malécot, A. (1958). The role of releases in the identification of released final stops: A series of tape-cutting experiments. *Language*, *34*(3), 370–380.

Mann, V. A. (1980). Influence of preceding liquid on stop-consonant perception. *Perception & Psychophysics*, *28*(5), 407–412.

McNeese, B. (2016). Are the Skewness and Kurtosis Useful Statistics? Retrieved August 25, 2018, from https://www.spcforexcel.com/publications/are-skewness-and-kurtosis-useful-statistics/mobile/index.html?doc=5399F2F40BC29BF5908624D55521252A

Miller, G., & Nicely, P. (1955). An analysis of perceptual confusions among some English consonant. *Journal of the Acoustical Society of America*, *27*(2), 338–352.

Miyazaki, K., & Sasaki, T. (1984). Pure-tone masking patterns in nonsimultaneous masking conditions. *Japanese Psychological Research*, *26*, 157–167.

Modarresi, G., Sussman, H., Lindblom, B., & Burlingame, E. (2004). An acoustic analysis of

the bidirectionality of coarticulation in VCV utterances. *Journal of Phonetics*, *32*, 291–312.

Modarresi, G., Sussman, H. M., Lindblom, B., & Burlingame, E. (2005). Locus equation encoding of stop place: revisiting the voicing / VOT issue, *Journal of Phonetics*, *33*, 101–113.

Moore, B. C .J. (2012). *An Introduction to the Psychology of Hearing* (6th ed.). Bingley: Emerald.

Moore, B. C. J. (2014). Development and current status of the "Cambridge" loudness models. *Trends in Hearing*, *18*, 1–29.

Moore, B. C. J., & Glasberg, B. R. (1983). Growth of forward masking for sinusoidal and noise maskers as a function of signal delay: Implications for suppression in noise. *Journal of the Acoustical Society of America*, *73*, 1249–1259.

Moore, B. C. J., & Glasberg, B. R. (1983). Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. *Journal of the Acoustical Society of America*, *74*(3), 750–753.

Morgan, N., Wegmann, S., & Cohen, J. (2013). What's Wrong With Automatic Speech Recognition (ASR) and How Can We Fix It? Berkeley, CA: International Computer Science Institute.

Nearey, T M. (1978). *Phonetic Feature Systems for Vowels*. PhD thesis, Indiana University.

Nearey, Terrance M, & Shammass, S. (1987). Formant transitions as partly distinctive invariant properties in the identification of voiced stops. *Canadian Acoustics*, *15*, 17–24.

Nossair, Z. B., & Zahorian, S. A. (1991). Dynamic spectral shape features as acoustic correlates for initial stop consonants. *Journal of the Acoustical Society of America*, *89*(6), 2978–2991.

Ohala, J. (1971). The role of physiological and acoustic models in explaining the direction of sound change. *Project on Linguistic Analysis Reports*, *15*, 25–40.

Öhman, S. E. G. (1966). Coarticulation in VCV utterances: Spectrographic measurements. *Journal of the Acoustical Society of America*, *39*(1), 151–168.

Oppenheim, J. N., & Magnasco, M. O. (2013). Human time-frequency acuity beats the Fourier uncertainty principle. *Physical Review Letters*, *110*(4), 1–5.

Oshiro, T. M., Perez, P. S., & Baranauskas, J. A. (2012). How Many Trees in a Random Forest? In P. Perner (Ed.), *Machine Learning and Data Mining in Pattern Recognition* (pp. 154–168). Berlin, Heidelberg: Springer Berlin Heidelberg.

Oxenham, A. J., & Moore, B. C. J. (1994). Modeling the additivity of nonsimultaneous masking. *Hearing Research*, *80*, 105–118.

Palmer, S. E. (1999). *Vision Science: Photons to Phenomenology*. Cambridge, MA: MIT Press.

Patterson, R. D. (1976). Auditory filter shapes derived with noise stimuli. *Journal of the Acoustical Society of America, 59*(3), 640–654.

Phatak, S. A., Lovitt, A., & Allen, J. B. (2008). Consonant confusions in white noise. *Journal of the Acoustical Society of America*, *124*(2) 1220-1233.

Pickett, J. M. (1999). *The Acoustics of Speech Communication: Fundamentals, Speech Perception, and Technology*. Boston: Allyn and Benton.

Pinker, S. (1997). *How the Mind Works*. London: Penguin.

Pintér, I. (1996). Perceptual wavelet-representation of speech signals and its application to speech enhancement. *Computer, Speech & Language*, *10*(1), 1–22.

Plomp, R., Pols, L. C. W., & van de Geer, J. P. (1967). Dimensional analysis of vowel spectra. *Journal of the Acoustical Society of America*, *41*(3), 707–712.

Potter, R. K., Kopp, G. A., & Green, H. C. (1947). *Visible Speech*. New York: D. Van Nostrand Company, Inc.

Psutka, J., Müller, L., & Psutka, J. (2001). Comparison of MFCC and PLP parameterizations in the speaker independent continuous speech recognition task. *Interspeech*, 5–8.

Régnier, M. S., & Allen, J. B. (2008). A method to identify noise-robust perceptual features: Application for consonant /t/. *Journal of the Acoustical Society of America*, *123*(5), 2801–2814.

Reidy, P. F. (2015). A comparison of spectral estimation methods for the analysis of sibilant fricatives. *Journal of the Acoustical Society of America*, *137*(4), 248-254.

Reidy, P. F. (2016). Spectral dynamics of sibilant fricatives are contrastive and language specific. *Journal of the Acoustical Society of America*, *140(*4), 2518-2529.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, *12*(1), 77.

Robinson, D. W., & Dadson, R. S. (1956). A redetermination of the equal-loudness relations for pure tones. *British Journal of Applied Physics*, *7*, 166–181.

Saremi, A., Beutelmann, R., Dietz, M., Ashida, G., Kretzberg, J., & Verhulst, S. (2016). A comparative study of seven human cochlear filter models. *Journal of the Acoustical Society of America*, *140*(3), 1618–1634.

Saremi, A., & Lyon, R. F. (2018). Quadratic distortion in a nonlinear cascade model of the human cochlea. *Journal of the Acoustical Society of America*, *143*(5), 418–424.

Schnupp, J., Nelken, I., & King, A. (2011). *Auditory Neuroscience: Making Sense of Sound*.

Cambridge, MA: MIT Press.

Schroeder, M. R. (1977). Recognition of complex acoustic signals. In T. H. Bullock (Ed.), *Life Sciences Research Report 5* (p. 324). Berlin: Abakon Verlag.

Sen, D., & Allen, J. B. (2006). Funcationality of cochlear micromechanics - as elucidated by the upward spread of masking and two-tone suppression. *Acoustics Australia*, *34*, 43–51.

Sengpiel, E. (n.d.). Forum für Mikrofonaufnahme Technik und Tonstudiotechnik. Retrieved September 26, 2018, from http://www.sengpielaudio.com/calculatorSonephon.htm.

Shanmugam, K. S. (1975). Comments on "Discrete Cosine Transform." *IEEE Transactions on Computers*, *100*, 759–759.

Shepard, R. (1972). Psychological representation of speech sounds. In E. E. David (Ed.), *Human Communciation: A Unified View* (pp. 67–113). New York: McGraw Hill.

Soli, S. D., Arabie, P., & Carroll, J. D. (1986). Discrete representation of perceptual structure underlying consonant confusions. *Journal of the Acoustical Society of America*, *79*, 826–837.

Stevens, K. N., & Blumstein, S. E. (1978). Invariant cues for place of articulation in stop consonants. *Journal of the Acoustical Society of America*, *64*(5), 1358–1368.

Stevens, K. N. (1998). *Acoustic Phonetics*. Cambridge, MA: MIT Press.

Stevens, K. N., Manuel, S. Y., & Matthies, M. (1999). Revisiting place of articulation measures for stop consonants: Implications for models of consonant production. *Proceedings of the 15th International Congress of Phonetic Sciences*, 1117-1120.

Stevens, S. S. (1957). On the psychophysical law. *Psychological Review*, *64*, 153–181.

Stevens, S. S. (1972). Perceived level of noise by mark VII and decibels (*E*). *Journal of the Acoustical Society of America*, *51*(2B), 575–601.

Stevens, S. S., Volkmann, J., & Newman, E. B. (1937). A scale for the measurement of the psychological magnitude pitch. *Psychological Review*, *8*, 185–190.

Strobl, C, Boulesteix, A.-L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, *8*(1), 1–25.

Strobl, Carolin, Malley, J., & Gerhard Tutz. (2009). Characteristics of classification and regression trees, bagging and random forests. *Psychological Methods*, *14*(4), 323–348.

Suchato, A. (2004). *Classification of Stop Consonant Place of Articulation*. PhD thesis, Massachusetts Institute of Technology.

Sussman, H. M. (1989). Neural coding of relational invariance in speech: Human language analogs to the barnowl. *Psychological Review*, *96*, 631–642.

Sussman, H. M., Fruchter, D., & Cable, A. (1995). Locus equations derived from

compensatory articulation. *Journal of the Acoustical Society of America*, *97*, 3112–3124.

Sussman, H. M., Fruchter, D., Hilbert, J., & Sirosh, J. (1998). Linear correlates in the speech signal: The orderly output constraint. *Behavioral and Brain Sciences*, *21*(2), 241–259.

Sussman, H. M., McCaffrey, H. A., & Matthews, S. A. (1991). An investigation of locus equations as a source of relational invariance for stop place categorization. *Journal of the Acoustical Society of America*, *90*(3), 1309–1325.

Team, R. C. (2018). R: A Language and Environment for Statistical Computing.

Team, Rs. (2018). RStudio: Integrated Development for R. Boston, MA.

Thomson, D. J. (2012). Spectrum estimation and harmonic analysis. *Proceedings of IEEE*, 70, 1055-1096.

Titze, I. R. (1994). *Principles of Voice Production*. Englewood Cliffs, NJ: Prentice-Hall.

van Schaik, A. (2017). CARFAC: Cochlear Modelling Jupyter Notebooks. Website: https://github.com/vschaik/CARFAC (retrieved August 22, 2018).

Viemeister, N. F., & Wakefield, G. H. (1991). Temporal integration and multiple looks. *Journal of the Acoustical Society of America*, *90*(2), 858–865.

Wada, T., Yasumoto, M., Ikeoka, N., Fujiki, Y., & Yoshinaga, R. (1970). An approach for the cinefluorographic study of articulatory movements. *Cleft Palate Journal*, *7*, 506–522.

Walsh, V., & Kulikowski, J. J. (1998). *Perceptual Constancy: Why Things Look as They Do*. Cambridge: Cambridge University Press.

Wang, D., & Brown, G. (2006). *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Hoboken, NJ: Wiley.

Wells, J. C. (1982). *Accents of English*. Cambridge: Cambridge University Press.

Wells, J. C. (1995). Computer-Coding the IPA: A Proposed Extension of SAMPA. Retrieved September 15, 2018, from https://www.phon.ucl.ac.uk/home/sampa/ipasam-x.pdf

Wheeler, D. J. (2004). *Advanced Topics in Statistical Process Control: The Power of Shewhart's Charts* (2nd ed.). Knoxville, TN.

Winter, B. (2013). Linear models and linear mixed effects models in R with linguistic applications. *ArXiv*, 1–42.

Xie, S. (2013). *Removing Redundancy in Speech by Modeling Forward Masking*. MSc thesis, University of Illinois at Urbana-Champaign.

Xu, Y., Thakur, C. S., Singh, R. K., Hamilton, T. J., Wang, R. M., & Van Schaik, A. (2018). A FPGA implementation of the CAR-FAC cochlear model. *Frontiers in Neuroscience*, *12*(APR), 1–14.

Yang, H., Vuuren, S. van, & Hermansky, H. (1999). Relevancy of time-frequency features for phonetic classification measured by mutual information. In *1999 IEEE International*

Conference on Acoustics, Speech, and Signal Processing. Proceedings. (Vol. 1, pp. 225–228).

Zue, V. (1976). *Acoustic Characteristics of Stop Consonants: A Controlled Study*. PhD thesis, Massachusetts Institute of Technology.

Zwicker, E. (1961). Subdivision of the Audible Frequency Range into Critical Bands (Frequenzgruppen). *Journal of the Acoustical Society of America*, *33*(2), 248.