VERIFICATION OF FEATURE REGIONS FOR STOPS AND FRICATIVES IN NATURAL SPEECH

BY

ANJALI INDUCHOODAN MENON

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Electrical and Computer Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2010

Urbana, Illinois

Adviser:

Associate Professor Jont B. Allen

ABSTRACT

The presence of acoustic cues and their importance in speech perception have long remained debatable topics. In spite of several studies that exist in this field, very little is known about what exactly humans perceive in speech. This research takes a novel approach towards understanding speech perception. A new method, named *three-dimensional deep search (3DDS)*, was developed to explore the perceptual cues of 16 consonant-vowel (CV) syllables, namely /pa/, /ta/, /ka/, /ba/, /da/, /ga/, /fa/, /θa/, /sa/, /ʃa/, /va/, /ða/, /za/, /ʒa/, from naturally produced speech. A verification experiment was then conducted to further verify the findings of the 3DDS method. For this purpose, the time-frequency coordinate that defines each CV was filtered out using the short-time Fourier transform (STFT), and perceptual tests were then conducted. A comparison between unmodified speech sounds and those without the acoustic cues was made. In most of the cases, the scores dropped from 100% to chance levels even at 12 dB SNR. This clearly emphasizes the importance of features in identifying each CV. The results confirm earlier findings that stops are characterized by a short-duration burst preceding the vowel by $\approx$ 10 cs in the unvoiced case, and appearing almost coincident with the vowel in the voiced case. As has been previously hypothesized, we confirmed that the F_2 transition plays no significant role in consonant identification. 3DDS analysis labels the /sa/ and /za/ perceptual features as an intense frication noise around 4 kHz, preceding the vowel by $\approx$ 15–20 cs, with the /za/ feature being around 5 cs shorter in duration than that of /sa/; the /ʃa/ and /ʒa/ events are found to be frication energy near $\approx$ 2 kHz, preceding the vowel by $\approx$ 17–20 cs. /fa/ has a relatively weak burst and frication energy over a wide-band including 2–6 kHz, while /va/ has a cue in the 1.5 kHz mid-frequency region preceding the vowel by $\approx$ 7–10 cs. New information is established regarding /ða/ and /θa/, especially with regards to the nature of their significant confusions.

To my parents, for inspiring me every day

ACKNOWLEDGMENTS

The author would like to express her appreciation to Prof. Jont Allen, Feipeng Li, Bryce E. Lobdell, Abhinauv Kapoor, Len Pan, Riya Singh, Andrea Trevino, Roger Serwy, and other members of the HSR group at University of Illinois, Urbana. The support of Etymotic Research, Phonak, and LDC for providing the database, is acknowledged. This research was supported, in part, by Grant R21-RDC009277A from the National Institute of Health.

The author would like to express special thanks to Ajay Menon and Ankit Chandra for their love and support.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

Speech is, by far, the most important mode of human communication, and for this reason there has been considerable research towards understanding how it works. However, these efforts have not been successful in getting to the crux of the matter leading to several conflicting theories. In analyzing speech, a widely recognized key problem is its large natural variability. This variability takes many forms, such as speaker gender (e.g., pitch), age, and accent, that would typically need to be considered and controlled, thereby greatly complicating any research.

Following the 1930–1940 development of the speech "vocoder" at Bell Labs, speech synthesis has been the hallmark of speech research. Starting at Haskins Laboratories in the 1950s, almost all the classical studies used vocoded speech or speech synthesis methods as a way of controlling for the speech features. While this use of synthetic speech controls for the desired dialect, talker, and feature control, a major disadvantage is that it fails to address the feature identification issues. One must first make assumptions about the speech features to synthesize and then use these prebuilt cues for perceptual tests. Obviously, one cannot generalize the analysis of synthetic speech to establish the various cues in natural speech.

Another serious disadvantage with the use of synthetic speech is the low quality and barely intelligible nature of the resulting speech. For example, in many of the classic studies, formants were replaced by tones, producing "sine-wave speech". Since perceptual tests have been used as the means of feature verification, speech quality is of key importance to this research.

Many studies also look at the problem from the distinctive feature standpoint. However, this study has a completely different approach. While listening to speech, humans do not usually analyze it in terms of production techniques. What this research aims to find is what it is in the speech waveform that uniquely identifies a sound.

The study of speech in noise is an area of great impact, ranging from hearing aids and telecommunications to consumer goods. This research meticulously studies the impact that noise has on speech, but using varying levels and type of noise in all the psychoacoustic experiments conducted.

A new methodology, named 3DDS, is introduced in this study. This method, reported first in [23], uses three independent experiments to find the feature regions for CV syllables. This method works well for different sounds across talkers and vowels. In this study, the 3DDS method has been used to find the feature regions for 14 CV pairs including stop and fricative consonants. Of these, a pictorial description of the feature regions with the vowel /a/, is presented in Fig. 1.1. These features have then been verified by another experiment, also reported in this study.

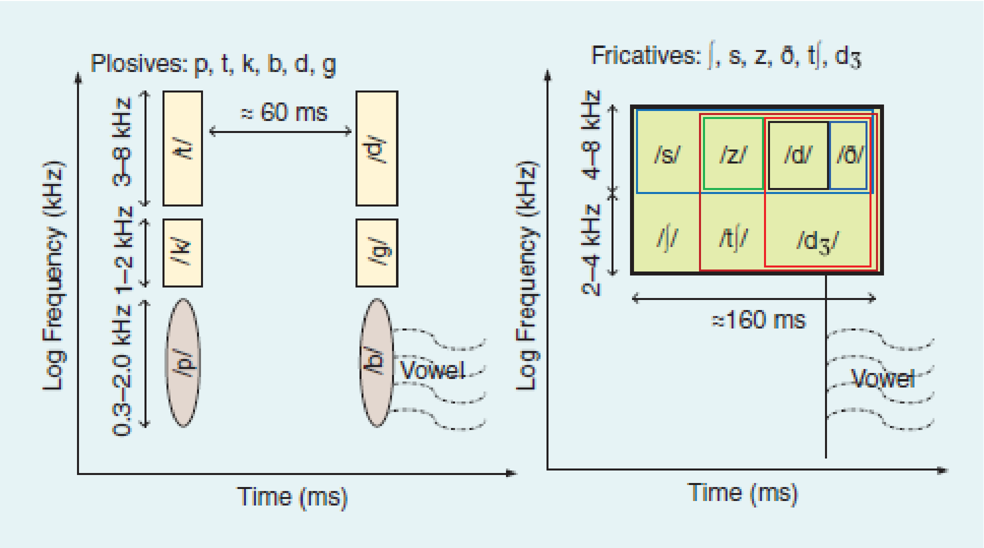


Figure 1.1: Summary of events for stops and fricatives with the vowel /a/.

CHAPTER 2

HISTORY

Speech sounds are characterized by time-varying spectral patterns called acoustic cues. When a speech wave propagates on the basilar membrane (BM), unique perceptual cues (named *events*), which define the basic units for speech perception, become resolved. The relationship between the acoustic cues and perceptual units has been a key research problem for speech perception [3, 5, 14].

The first search for acoustic cues dates back to 1940s at Bell Labs, when Potter et al. (1966) [28] began their *visible speech* project, with the goal of training the hearing-impaired to read spectrograms. Five normal hearing (NH) listeners and one hearing-impaired (HI) listener participated in the study. Following a series of lectures on the spectrogram and its use on isolated syllables and continuous speech, the subjects were successfully trained to "read" speech spectrograms. Even though the acoustic cues identified by visual inspection were not very accurate, this pioneering work laid a solid foundation for subsequent quantitative analysis.

2.1 Stop consonants

Cooper et al. (1952) [11], along with other researchers at the Haskins Laboratories over the following decade, conducted a series of landmark studies on the acoustic cues of consonant sounds. A speech synthesis system, called the *Pattern Playback*, was created to convert a spectrogram into (low-quality) speech sound. Based on the spectrograms of real speech, it was postulated that stop consonants are characterized by a initial burst, followed by a consonant-vowel transition. In this study, the authors investigated the effect of center frequencies of the burst and the second formant (F_2) transition on the percept of unvoiced stop consonants, by using a set of "nonsense" synthetic consonant-vowel (CV) speech sounds synthesized from 12 bursts followed by seven F_2 formant frequencies. The subjects were instructed to identify the stimulus as /p/, /t/, or /k/ (a closed-set task). Results show that most people hear /t/ when the burst frequency is higher than the F_2 frequency; when the two frequencies are close, most listeners report /k/; otherwise they hear /p/. In a following study by Delattre et al. (1955) [12], the authors dropped the burst and examined the effect of F_2 transition alone on the percept of stop consonants. It was found that stimuli with rising F_2 transition were identified as /b/; those with F_2 emanating from 1.8 kHz were associated with /d/; and those with a falling transition were reported as /g/.

Liberman et al.'s study [24] had a major impact on the research of speech perception. Since their study, speech synthesis has become a standard method for feature analysis. It was used in the search for acoustic correlates for stops [9], fricatives [19, 20], nasals [25, 29], as well as for distinctive and articulatory features [7, 8, 35]. Remez et al. (1981) [31] took a similar approach to generate highly unintelligible "sine-wave" speech and then concluded that the traditional cues, such as bursts and transitions, are not required for speech perception. The status quo is extremely confusing in that many people strongly believe that the stop consonants are defined by the bursts and transitions [11, 12], yet still argue that modulation is the key to understanding speech perception [13, 34], without realizing that the two arguments are actually in conflict.

The argument in favor of the speech synthesis method is that the features can be carefully controlled. However, the major disadvantage of synthetic speech is that it requires prior knowledge of the cues being sought. This incomplete and inaccurate knowledge about the acoustic cues has often led to synthetic speech of low quality, and it is common that such speech sounds are unnatural and barely intelligible, which by itself is a strong evidence that the critical cues for the perception of target speech sound are poorly represented. In all of these cases, it is necessary to study the acoustic cues of naturally produced speech, rather than artificially synthesized speech, to really understand what causes one sound to morph into another or what makes one sound more robust than another when presented in noise.

The use of confusion matrices to study the perception of speech sounds was started by Campbell (1914), and then taken up by Miller and Nicely (1955) [26], with a mutual information analysis of five distinctive features. Wang and Bilger (1973) [38] used the Sequential Information Analysis (SINFA) method to partition transmitted information of distinctive features to characterize confusion matrices. Accordingly, each phoneme was assigned a weight according to a set of distinctive features, such as voicing or nasality. Thereafter, information transmitted for each such feature, when held constant, was calculated until all features were accounted for. The information transmitted [33] represents the contingency between the joint feature categories of the stimulus and the joint feature categories of the response. Since all of these features are not truly independent, this leads to some amount of redundancy. The redundancy was then calculated, the hypothesis being that once the internal redundancy of the feature systems is taken into account, some articulatory and phonological features account for information transmitted better than others.

A related classical study is that of Cole and Scott (1974) [10], who studied three types of speech cues: (1) the transitional (frequency glides occurring as a result of the vowel following the consonant); (2) invariant (features that do not change irrespective of the following vowel); and (3) envelope cues. They conclude that the fricatives /s,ʃ,z,ʒ/ and the affricates /ȼ,ɟ/ are characterized by invariant cues, which also help discriminate phoneme pairs /m,n/, /f,θ/, and /v,ð/, from other sounds. However, transitional cues are needed to discriminate between the members of each pair. For stop consonants, the voiced/voiceless discrimination was based on invariant cues, while transitional cues were needed to identify each individual stop consonant.

Blumstein and Stevens made significant contributions to the understanding of stop consonants (especially /b, d, g/). Blumstein et al. (1977) [9] investigated the role of initial bursts and transitions, to identify the place of articulation (POA) of stop consonants. A continuum of formant frequencies (those of /b/ to /d/ to /g/) were used. Stevens and Blumstein (1978) [35] used the stops /b, d, g/ with the vowels /a,u,i/ via a Klatt synthesizer [21], and postulated that the place of articulation for a syllable-initial consonant could be identified on the basis of the gross shape of the spectrum sampled at the consonantal release. This was reported to be vowel independent. In particular, velar sounds had a prominent mid-frequency spectral peak, alveolars had a diffuse rising spectrum and labials had a diffuse falling spectrum. This hypothesis was tested in Blumstein and Stevens (1979) [7]. A series of templates were designed to reflect the spectral properties mentioned, namely diffuse-rising, diffuse-falling and compact. Using these templates, the average classification score for all the stop consonant data that the study used was above 80%. This was interpreted as strong support for the theory of acoustic invariance. Blumstein and Stevens (1980) [8] reported a series of four experiments. The first experiment studied the brief onset portion of CVs for correlations with the POA to establish the minimum burst vowel time. A second experiment was conducted to determine whether the onset information provided by the formant transitions (without the burst) could cue the place of articulation. Experiment three aimed at studying the effects on perception of POA when the intensity of the spectral cues was varied. The fourth experiment asked whether listeners could derive vowel as well as consonant information from the stimuli. It was seen that listeners could, in fact, identify the consonant and vowel information from as little as 1 glottal pulse of information.

Van Tassel et al. (1987) [37] modulated noise with the speech envelopes for 19 /aCa/ natural speech nonsense syllables, along with three sets of low-pass filter cutoffs. Multidimensional scaling (SINDSCAL) was used for an analysis of the consonant confusions. Their experiments showed that, compared to unprocessed sounds, subjects identified the speech envelope noise poorly, but well above chance. The multidimensional analysis revealed three waveform envelope features: voicing, amplitude, and wideband burst envelope. When the consonants were divided into envelope feature groups (envemes) and visually distinctive feature groups (visemes), nearly 95% scores were achieved with just these features and with no additional spectral cues.

More recently, Hazan and Simpson (1998) [18] worked with enhancing features of sounds in two separate experiments, one with just VCVs and the other with semantically unpredictable sentences. In the case of VCVs, they used four different modes of enhancement namely enhancement, of the burst only and of the burst and the transition region, and these two were then repeated with filtering done to change the spectral content of perceptually important regions to make them more discriminable. It was seen that the highest mean increase in speech intelligibility was 12% at -5 and 6% at 0 dB SNR for the last type of enhancement.

2.2 Fricative consonants

Fricatives also have been studied in great detail, and they have cues very different in nature compared to stops. Hughes and Halle (1956) [20] studied the fricative spectra in the context of meaningful words. Natural speech was used for the study and, as expected, they found large variation in the spectra across different speakers for the same sound. They reported that the differences among the different classes of fricatives were fairly consistent and went on to develop a recognizer to segregate fricatives on the basis of the energy at different frequencies. While this worked well for most fricatives, the fricative /f/ had the greatest ambiguity.

Heinz and Stevens (1961) [19] reviewed the acoustical theory behind the production of voiceless fricatives and developed an electrical production model. The stimuli generated by this model were then presented to listeners, and the results of the perceptual tests were consistent with their acoustic analysis.

Stevens et al. (1992) [36] studied the distinguishing factors between voiced and unvoiced fricatives. They found that listeners based their voicing judgments of intervocalic fricatives on the time interval duration for which there was no glottal vibration. If this time interval was greater than 60 ms, the fricative was usually judged as voiceless.

The Ph.D. dissertation of Alwan (1992) [6] was focused on the CV confusions between the stop consonants /b/ and /d/ in the context of vowels /a/ and /ε/. Using synthesized speech, she studied the role of formant trajectories to differentiate between the two in the presence of noise.

Hasegawa-Johnson (2000) [17] defined the *infogram* as an estimate of the mutual information between the value of a distinctive feature and the amplitude of each point in time-frequency plane, relative to an acoustic landmark. He showed that manner features were easier to identify, based on a well-defined point in the time-frequency plane, than place features.

Other significant studies include Rhebergen et al. (2005) [32], which used a Speech Intelligibility Index (SII-based model to account for speech reception thresholds (SRT) in stationary noise, fluctuating noise, interrupted noise, and multiple-talker (babble) noise.

CHAPTER 3

THE 3D DEEP SEARCH METHOD

3.1 Introduction

Speech sounds are characterized in three dimensions: time, frequency and intensity. Event identification involves isolating the speech cues along these three dimensions. In past studies, confusion tests on nonsense syllables has long been used for the exploration of speech features. For example, Fletcher and his colleagues investigated the contribution of different frequency bands to speech intelligibility using high-pass and low-pass filtered CV syllables [14, 15], resulting in the *articulation index* (AI) model. Furui [16] examined the relationship between dynamic features and the identification of Japanese syllables modified by initial and final truncation. More often masking noise was used to study consonant [26, 38] and vowel [27] recognition. Regnier and Allen [30] successfully combined the results of time truncation and noise masking experiments, for the identification of /ta/ events.

The 3DDS method requires three independent experiments for each CV utterance. This method was developed by Feipeng Li. The *first* experiment (TR07) determines the contribution of various time intervals by truncating the consonant into multiple segments of 5, 10 or 20 ms per frame from the front, depending on the sound and its duration. The *second* experiment (HL07) divides the full band into multiple bands of equal length along the BM and measures the score in different frequency bands by using high-pass/low-pass filtered speech as the stimuli. Once the time-frequency coordinates of the event have been identified, a *third* experiment (MN05) assesses the strength of the speech event by masking the speech at various signal-to-noise ratios. To reduce the length of the experiments, the three dimensions, i.e., time, frequency and intensity, are assumed to be independent. The identified events are verified by a special software designed for the manipulation of acoustic cues, based on the short-time Fourier transform (STFT) [1,4].

3.2 Methods

The detailed methods of the three experiments are described next.

3.2.1 Subjects

In total, sixty-two listeners were enrolled in these three studies, of which nineteen subjects participated in experiment HL07, another nineteen subjects participated in experiment TR07, one participated in both the experiments, while the remaining 24 subjects were assigned to experiment MN05. All subjects self-reported no history of speech or hearing disorder. Except for two listeners, all the subjects were born in the United States. with their first or primary language being English. The subjects were paid for their participation. IRB approval was obtained prior to the experiment.

3.2.2 Speech stimuli

Sixteen CVs: /pa, ta, ka, fa, θa, sa, ʃa, ba, da, ɡa, va, ða, za, ʒa, ma, na/ chosen from the University of Pennsylvania's Linguistic Data Consortium (LDC-2005S22, aka *the Fletcher AI corpus*) were used as the common test material for the three experiments. The speech sounds were sampled at 16 kHz. Each CV had 20 talkers. Experiment MN05 used 18 talkers. For the other two experiments (TR07 and HL07), to reduce the total time, only 6 utterances (half male and half female) were chosen for the test. The 6 utterances were selected such that they were representative of the speech material in terms of confusion patterns and articulation score, based on the results of the MN05 speech perception experiment. For this reason of balance, a percentage of low-scoring sounds were included in the utterance set. There were thus a total of 96 utterances used (16 sounds × 6 utterances per sound). The speech sounds were presented diotically (both ears) through Sennheiser HD-280 PRO circumaural headphone, adjusted in level at the listener's *Most Comfortable Level* (MCL), i.e., ≈ 70–80 dB SPL. All experiments were conducted in a single-walled IAC sound-proof booth, situated in a lab with no windows, with the lab outer door shut.

3.2.3 Conditions

Experiment TR07 assesses the temporal distribution of events. For each utterance, truncation starts before the beginning of the consonant and stops after the end of the consonant. The truncation times were chosen such that the duration of the consonant was divided into non-overlapping intervals of 5, 10, or 20 ms. An adaptive scheme was applied for the calculation of the sample points. The basic idea is to assign more points where the speech scores change rapidly. Starting from the end of the consonant, where the consonant-vowel transition is located, it allocates eight truncation times (frames) of 5 ms, followed by twelve frames of 10 ms, and as many 20 ms frames as needed until it covers the entire interval of the consonant part. To make the truncated speech sounds more natural, white noise was added following truncation, to mask the speech stimuli at a signal-to-noise ratio of 12 dB.

Experiment HL07 investigates the frequency distribution of events. It has 19 filtering conditions, including one full-band (0.25–8 kHz), nine high-pass, and nine low-pass conditions. The cutoff frequencies were calculated using Greenwood's inverse function, so that the full-band frequency range was divided into 12 bands, each having an equal length along the basilar membrane. The cutoff frequencies of the high-pass filtering were 6185, 4775, 3678, 2826, 2164, 1649, 1250, 939, and 697 Hz, with the upper limit being fixed at 8000 Hz. The cutoff frequencies of the low-pass filtering were 3678, 2826, 2164, 1649, 1250, 939, 697, 509, and 363 Hz, with the lower limit being fixed at 250 Hz. Note how the high-pass and low-pass filtering share seven cut-off frequencies (3678–697 Hz) over the middle of the frequency range. The filters were implemented via a sixth-order elliptic filter having a stopband attenuation of 60 dB. Again, white noise (12 dB SNR) was added to the modified speech in order to make it sound more natural. Note that for most CV sounds, 12 dB SNR does not affect the scores [27].

Experiment MN05 measures the strength of the event in terms of robustness to masking white noise. Besides the quiet condition, speech sound were masked at eight different signal-to-noise ratios $[-21, -18, -15, -12, -6, 0, 6, 12]$. Details may be found in [27].

All three experiments included a common control condition, i.e., full-band speech at 12 dB SNR. The recognition scores for this common control condition should be the same across the three experiments.

3.2.4 Procedure

The three experiments (TR07, HL07, MN16R) used nearly identical procedures. A Matlab program was created for the stimulus presentation and data collection. A mandatory practice session with feedback was given at the beginning of each experiment. Speech tokens were randomized across the talkers, conditions, and consonants. Following each stimulus presentation, subjects responded by clicking on a button labeled with the CV that they thought they heard. In case the CV was completely masked by the noise, or the processed token did not sound like any of the 16 consonants, the subject was instructed to click on a "Noise Only" button. Frequent breaks were encouraged to prevent fatigue. Subjects were allowed to play each token up to 3 times, after which the token was pushed to the end of the list. The waveform was played via a SoundBlaster 24-bit sound card in a PC Intel computer, running Matlab under Ubuntu Linux.

CHAPTER 4

RESULTS FOR THE 3DDS METHOD

The results are organized into unvoiced/voiced pairs of stops and fricatives /ta/–/da/, /ka/–/ga/, /pa/–/ba/, /ʃa/–/ʒa/, /sa/–/za/, /fa/–/va/, and /θa/–/ða/. The sounds have been thus paired in each figure, and discussed together, to more effectively highlight their similarities and differences. Each of these four figures (4.1–4.4) consists of two sub-figures. The left figure (a) shows the 3DDS method applied to the unvoiced stop/fricative and the right figure (b) to the corresponding voiced stop/fricative.

Each of the sub-figures [(a), (b)] contains 5 panels labeled panel $\boxed{1}$–$\boxed{5}$. Details about the utterance (i.e., the talker ID and gender) are given in panel $\boxed{1}$ of the sub-figure. For each sub-figure [i.e., (a) or (b)], panel $\boxed{1}$ (center-left), shows the AI-gram of the original sound at 18 dB SNR. The AI-gram is a spectrogram made with Fletcher critical bands normalized to the background noise. Panel $\boxed{2}$ (top-left) shows the TR07 truncation scores, panel $\boxed{3}$ (top-right) shows the MN05 noise masking scores, while panel $\boxed{4}$ (center-right) gives the HL07 high-pass and low-pass filtering scores. Panel $\boxed{5}$ (bottom) provides an AI-gram summary of the sound, at six different SNR values from -12 to $+18$ dB, in 6 dB steps.

To identify each sound event it is necessary to align the time and frequency coordinates from each of the three experiments. The perceptual scores are a function of the truncation times t_n, cutoff frequencies f_k and signal-to-noise ratios SNR_k. Accordingly the truncation scores of panel $\boxed{2}$ are aligned to the time axis of the AI-gram panel $\boxed{1}$; similarly the high-pass and low-pass frequencies of panel $\boxed{4}$ are aligned to the frequency axis of the AI-gram panel $\boxed{1}$.

For the truncation data in panel $\boxed{2}$, the scores (i.e., probability correct) on the ordinate are denoted by $c_{h|s}(t_n)$. In each of these cases, the first subscript (h) for the consonant score c specifies the consonant heard given that consonant identified by the second subscript (s) was spoken; e.g. $c_{p|k}(t_n)$

gives the score c for subjects reporting a h =/pa/ when a s =/ka/ was the spoken stimulus, as a function of the truncation time t_n.

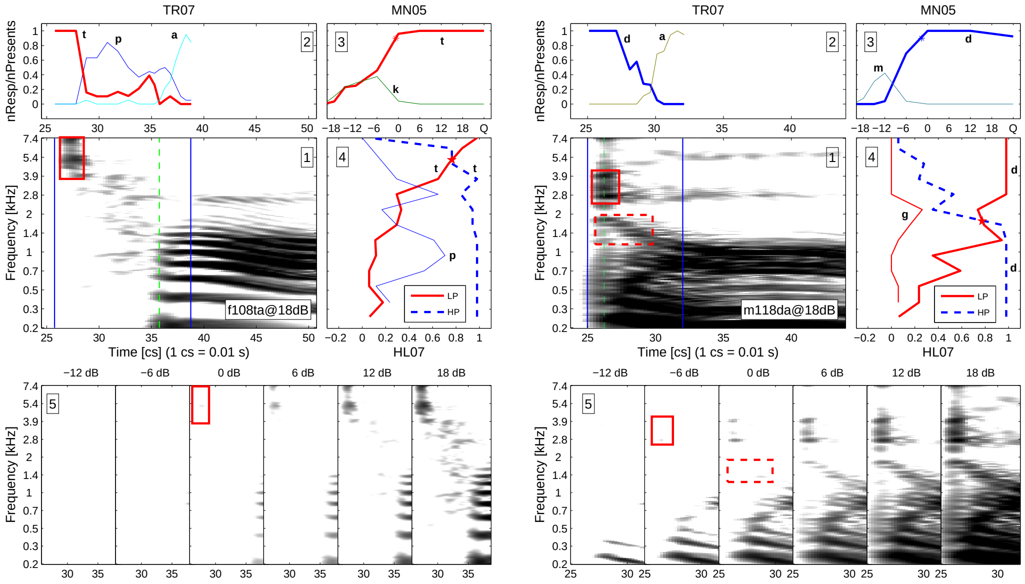
The low-pass and high-pass filtering scores are $c^L_{h|s}(f_k)$ and $c^H_{h|s}(f_k)$, on the horizontal axis, since the graph is rotated. The scores and confusions for the high-pass data are indicated by dashed lines, and those for the low-pass data are indicated by solid lines in panel $\boxed{4}$ of each sub-figure. The noise masking scores $c_{h|s}(\mathrm{SNR}_k)$ of panel $\boxed{3}$ are on the vertical axis, while the horizontal axis is the SNR. In all three cases, only the significant confusions have been plotted, along with the target scores.

4.1 Results for stops

4.1.1 /ta/ and /da/

Results of the three experiments (TR07, HL07, and MN05) clearly indicates that the /ta/ event (refer to Fig. 4.1(a) for a /ta/ from talker f105) is a high-frequency burst above 3 kHz, 1.5 cs in duration and 5–7 cs prior to the vowel. Panel $\boxed{1}$ shows the AI-gram of the sound at 18 dB SNR in white noise with the hypothetical /ta/ event being highlighted by a rectangular frame. On top of it, panel $\boxed{2}$ depicts the results of the time-truncation experiment. When the burst is completely removed at 28 cs, the score for the time-truncated /t/ drops dramatically from 1 to chance level and listeners start reporting a /pa/, suggesting that the high-frequency burst is critical for /ta/ perception. This is in total agreement with the high-pass and low-pass data, as shown in panel $\boxed{4}$. Once the high-frequency burst has been removed by the low-pass filtering (solid curve), the /ta/ score $c^L_{t|t}(f_k)$ drops dramatically and the confusion with /pa/ increases significantly. The intersection of the high-pass and the low-pass perceptual scores (indicated by the $\star$) is at around 5 kHz, showing the dominant cue to be the high-frequency burst. These results are confirmed by the noise-masking experiment. From the AI-grams in panel $\boxed{5}$ we see that the high-frequency burst becomes inaudible when the SNR is lower than 0 dB; as a consequence the recognition score drops sharply at -1 dB SNR (labeled by a $*$ in panel $\boxed{3}$), meaning that the perception of /ta/ is dominated by the high-frequency burst.

Of the six /ta/ sounds, five morphed to /pa/ once the /ta/ burst was



(a) Unvoiced /ta/ by talker f108

(b) Voiced /da/ by talker m118

Figure 4.1: Hypothetical events for high-frequency stop consonants /ta/ and /da/. The multiple panels in each sub-figure are: Panel $\boxed{1}$: AI-gram. A dashed vertical line labels the onset of voicing (sonorance), indicating the start of the vowel. The solid and dashed boxes indicate the dominant and minor events respectively. Panel $\boxed{2}$: CPs as a function of truncation time t_n. Panel $\boxed{3}$: CPs as a function of $\mathrm{SNR_k}$. Panel $\boxed{4}$: CPs as a function of cutoff frequency f_k. Panel $\boxed{5}$: AI-grams of the consonant region [defined by the solid vertical lines on panel $\boxed{1}$], at $-12, -6, 0, 6, 12, 18$ dB SNR.

truncated, while one morphed to /ka/ (m112ta). For this latter sound, it was seen that the /ta/ burst preceded the vowel by only around 2 cs as opposed to 5–7 cs which is the case for a normally articulated /ta/. This timing cue is especially important for the perception of /pa/, as we will show later in the results section.

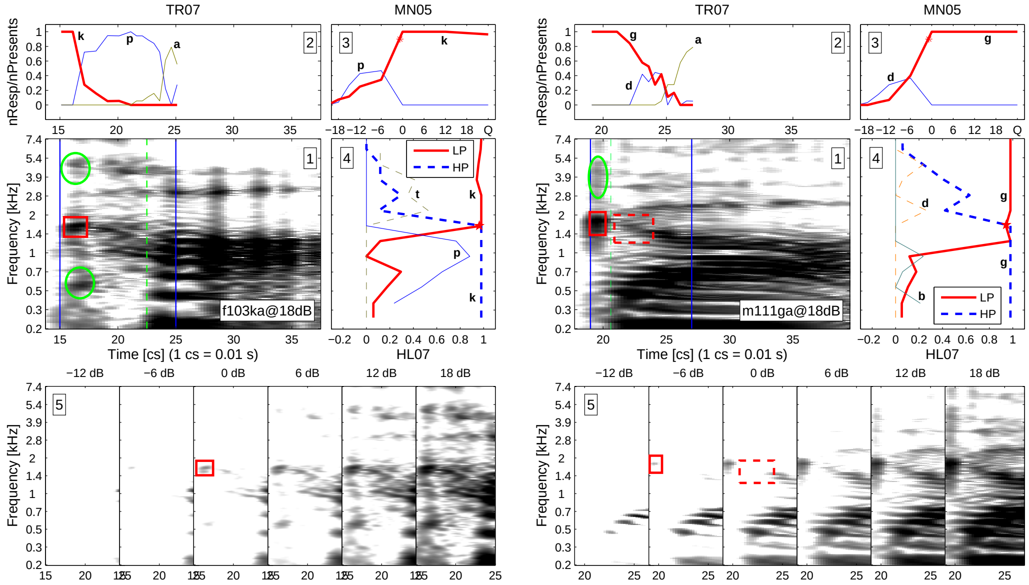
Consonant /da/ (Fig. 4.1(b)) is the voiced counterpart of /ta/. It is characterized by a high-frequency burst above 4 kHz and an $\mathrm{F_2}$ transition near 1.5 kHz, as shown in panel $\boxed{1}$. Truncation of the high-frequency burst (panel $\boxed{2}$) leads to an immediate drop in the score for /da/ from 100% at $t_n = 27$ cs to about 70% at $t_n = 27.5$ cs. The recognition score keeps decreasing until the $\mathrm{F_2}$ transition is removed completely at 30 cs. From the high-pass and low-pass data (panel $\boxed{4}$), it is seem that subjects need to hear both the $\mathrm{F_2}$ transition and the high-frequency burst to get a full score of 100%. Lack of the burst usually leads to the /da/→/ga/ confusion, as shown by the low-pass confusion of $c^L_{g|d}(f_k)$ = 30% at f_k = 2 kHz (solid curve labeled "g" in panel $\boxed{4}$), meaning that both the high-frequency burst and the F_2 transition are important for the identification of a high quality /da/. This is confirmed by the results of the noise-masking experiment. From the AI-grams (panel $\boxed{5}$) the F_2 transition becomes masked by noise at 0 dB SNR; accordingly the /da/ score in panel $\boxed{3}$ drops quickly at the same SNR. When the remnant of the high-frequency burst is finally gone at −6 dB SNR, the /da/ score decreases even faster, until the /d/ and /m/ scores are equal.

Some of the /da/s are much more robust to noise than others. For example, the SNR_{90}, defined as the SNR where the listeners begin to lose the sound (P_c = 0.90), is −6 dB for /da/-m104, and +12 dB for /da/-m111. The variability over the six utterances is impressive, yet the story seems totally consistent with the requirement that both the burst and the F_2 transition need to be heard.

4.1.2 /ka/ and /ga/

Analysis of Fig. 4.2(a) reveals that the event of /ka/ is a mid-frequency burst around 1.6 kHz, articulated 5 − 7 cs before the vowel, as highlighted by the rectangular boxes in panels 1 and 5. The truncation data (panel $\boxed{2}$) show that once the mid-frequency burst is truncated at 16.5 cs, the recognition score for /ka/ jumps from 100% to chance level within 1-2 cs. At the same time, most listeners begin to hear /pa/. The high-pass score $c^H_{k|k}(f_k)$ and the low-pass score $c^L_{k|k}(f_k)$ (panel $\boxed{4}$) cross each other at 1.4 kHz. Both curves have a sharp dive around the intersection point, suggesting that the perception of /ka/ is dominated by the mid-frequency burst. Based on the AI-grams (panel $\boxed{5}$), the mid-frequency burst is just above its detection threshold at 0 dB SNR; accordingly the recognition score of /ka/ (panel $\boxed{3}$) drops dramatically at 0 dB SNR. Thus the results of the three experiments are in perfect agreement in identifying the mid-frequency burst around 1.6 kHz as the single dominant cue of /ka/.

Not all of the six sounds strongly morph to /pa/ once the /ka/ burst was truncated, as is seen in Fig. 4.2(a). Two out of six had no morphs, just remained a very weak /ka/ once the onset-burst was removed (m114ka,



(a) Unvoiced /ka/ by talker f103

(b) Voiced /ga/ by talker m111

Figure 4.2: Hypothetical events for mid-frequency stop consonants /ka/ and /ga/. The multiple panels in each sub-figure are: Panel $\boxed{1}$: AI-gram. A dashed vertical line labels the onset of voicing (sonorance), indicating the start of the vowel. The solid and dashed boxes indicate the dominant and minor events respectively. Panel $\boxed{2}$: CPs as a function of truncation time t_n. Panel $\boxed{3}$: CPs as a function of SNR_k. Panel $\boxed{4}$: CPs as a function of cutoff frequency f_k. Panel $\boxed{5}$: AI-grams of the consonant region [defined by the solid vertical lines on panel $\boxed{1}$], at $-12, -6, 0, 6, 12, 18$ dB SNR.

f119ka). Again, these scores are consistent with guessing.

Consonant /ga/ (Fig. 4.2(b)) is the voiced counterpart of /ka/. It is represented by a mid-frequency burst from 1.4 to 2 kHz, followed by an F_2 transition between 1 and 2 kHz, as highlighted with boxes in panel $\boxed{1}$. According to the truncation data (panel $\boxed{2}$), the recognition score of /ga/ starts to drop when the mid-frequency burst is truncated beyond $t_n = 22$ cs. At the same time the /ga/→/da/ confusion appears, with the score for /da/ being 40% at $t_n = 23$ cs. The high-pass score and low-pass score (panel $\boxed{4}$) fully overlap at the frequency of 1.6 kHz, where both show a sharp decrease of more than 60%, which is consistent with the statements about /ga/ events. Based on the AI-grams in panel $\boxed{5}$, the F_2 transition is masked by 0 dB SNR, corresponding to the turning point of the /ga/ score, labeled by a $*$ in panel $\boxed{3}$. When the mid-frequency burst gets masked at -6 dB SNR,

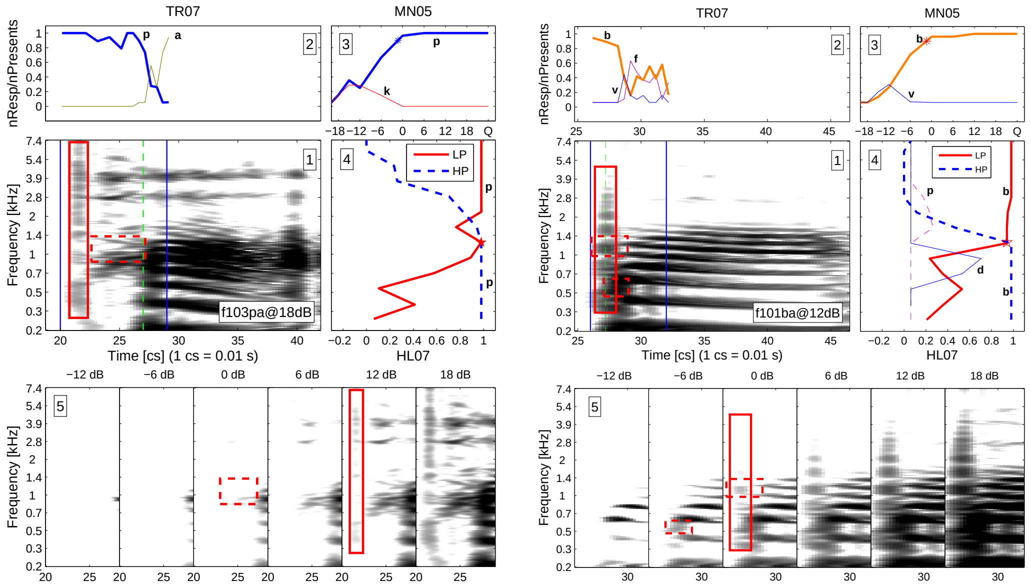
/ga/ becomes confused with /da/, suggesting that the perception of /ga/ is dominated by the mid-frequency burst.

All six /ga/ sounds have well defined bursts between 1.4 and 2 kHz. Most of the /ga/s (m111, f119, m104, m112) have a perfect score of /ga/ is 100% at 0 dB SNR. The other two /ga/s (f109, f108) are relatively weaker.

It is interesting to see that these two mid-frequency sounds all have conflicting cues that are characteristic of competing sounds. For example, the /ka/ sound (Fig. 4.2(a)) also contains a high-frequency burst around 5 kHz and a low-frequency burst around 0.5 kHz, which can be used as the perception cues of /ta/ and /pa/, respectively. As a consequence, listeners hear /ta/ when the high-pass cutoff frequency is higher than the upper limit of the /ka/ burst (2 kHz). In the low-pass experiment, people hear /pa/ when the low-pass cutoff frequency is smaller than 1.2 kHz, the lower limit of the /ka/ cue. Similarly the /ga/ sound also contains a high-frequency burst above 4 kHz that promotes confusion with /da/.

4.1.3 /pa/ and /ba/

Figure 4.3(a) for /pa/ spoken by female talker f103 (LDC file `s_f103_pa.wav`) reveals that there may be two different events: (1) a wide-band click running from 0.3 to 7.4 kHz, maskable by white noise at 12 dB SNR; and (2) a formant transition at 1–1.4 kHz, maskable by white noise at 0 dB SNR. Panel $\boxed{2}$ shows the truncated /p/ score as a function of the truncation time t_n. It starts at 100% from the beginning. When the wid- band click, which includes the low-frequency burst, is truncated at around 23 cs, the score is seen to drop. It drops to the chance level (1/16) when the transition is removed at $t_n = 27$ cs. Thus both the wide-band click and the F_2 transition contribute to the perception of /pa/. The low-pass and high-pass scores, as depicted in panel $\boxed{4}$, start at 100% at each end of the spectrum, and they begin to drop only near the intersection point, close to 1.3 kHz. This intersection (indicated by a $\star$) appears to be a clear indicator of the center frequency of the dominant perceptual cue, which is the F_2 region running from 22 to 26 cs. The recognition score of the noise masking experiment (panel $\boxed{3}$) drops dramatically at 0 dB SNR (denoted by a $\ast$). From the six AI grams (panel $\boxed{5}$), we can see that the audible threshold for the F_2 transition is at 0 dB



(a) Unvoiced /pa/ by talker f103

(b) Voiced /ba/ by talker f101

Figure 4.3: Hypothetical events for low-frequency stop consonants /pa/ and /ba/. The multiple panels in each sub-figure are: Panel $\boxed{1}$: AI-gram. A dashed vertical line labels the onset of voicing (sonorance), indicating the start of the vowel. The solid and dashed boxes indicate the dominant and minor events respectively. Panel $\boxed{2}$: CPs as a function of truncation time t_n. Panel $\boxed{3}$: CPs as a function of $\mathrm{SNR_k}$. Panel $\boxed{4}$: CPs as a function of cutoff frequency f_k. Panel $\boxed{5}$: AI-grams of the consonant region [defined by the solid vertical lines on panel $\boxed{1}$], at $-12, -6, 0, 6, 12, 18$ dB SNR.

SNR, the same as the turning point ($*$) in panel $\boxed{3}$, where the listeners begin to lose the sound, giving credence to the energy of F_2 sticking out in front of the sonorant portion of the vowel as the main cue for the /pa/ event.

The stop consonant /pa/ is characterized as having a wide-band click, which is seen in this /pa/ example, but not in the five others we have studied. For most /pa/s, the wide-band click diminishes into a low-frequency burst. The click does appear to contribute to the overall quality of /pa/ when it is present. The 3D displays of other five /pa/s are in basic agreement with that of Fig. 4.3(a), with the main difference being the existence of the wide-band burst at 22 cs for f103, and slightly different high-pass and low-pass intersection frequency, ranging from 0.7 to 1.4 kHz, for the other five sounds. The required duration of the F_2 energy before the onset of voicing (around $3-5$ cs) is very critical to the perception of /pa/. The existence of excitation of F_3 is evident in the AI-grams, but it does not appear to interfere with the identification of /pa/, unless F_2 has been removed by filtering (a minor effect for f103). Also /ta/ was identified in a few examples, as high as 40% when F_2 was masked.

The perceptual events for /ba/ are perhaps the most difficult of the six stops. For the 3DDS method to work well, high scores in quiet are essential. Among the six /ba/ sounds, only the one shown has 100% scores at 12 dB SNR and above. Based on the analysis of Fig. 4.3(b), the hypothetical features for /ba/ include: (1) a wide-band click in the range of 0.3 to 4.5 kHz; (2) a low-frequency burst at around 0.4 kHz; and (3) an F_2 transition around 1.2 kHz. When the wide-band click is completely truncated at t_n =28 cs, the /ba/ score [Fig. 4.3(b)] drops dramatically from 80% to chance level, at the same time the /ba/→/va/ confusion /ba/→/fa/ confusion increase quickly, indicating that the wide-band click is important for the distinguishing /ba/ from the two fricatives /va/ and /fa/. However, since the three events overlap on the time axis, it is hard to tell which event plays the major role. Panel $\boxed{4}$ shows that the high-pass score $c^H_{b|b}(f_k)$ and low-pass score $c^L_{b|b}(f_k)$ cross each other at 1.3 kHz, both change fast within 1–2 kHz, indicating that the F_2 transition, centered around 1.3 kHz, is very important. Without the F_2 transition, as we see in the low-pass data while f_k <1 kHz, most listeners guess /da/ instead of /ba/. Besides, the small jump in the low-pass score $c^L_{b|b}(f_k)$ around 0.4 kHz suggests that the low-frequency burst may also play a role in /ba/ perception. From the AI-grams in panel $\boxed{5}$, the F_2 transition and wide-band click become masked by the noise somewhere below 0 dB SNR. Accordingly the listeners begin to lose the /ba/ sound in the masking experiment around the same SNR, as represented by a $*$ in panel $\boxed{3}$. Once the wide-band click has been masked, the confusions with /va/ increase, and become equal to /ba/ at −12 dB SNR with a score of 40%.

There are the only three LDC /ba/ sounds out of 18 with 100% scores at and above 12 dB SNR, i.e., /ba/ from f101 shown here and /ba/ from f109, which has a 20% /va/ error rate for SNR ≤ -10 dB SNR. The remaining 18 /ba/ utterances have /va/ confusions between 5 and 20%, in quiet. We do not know whether it is the recordings in the LDC database that are responsible for these low scores, or whether /ba/ is inherently difficult. Low-quality consonants with error rates greater than 20% were also observed in the LDC study by Phatak and Allen (2007) [27]. From unpublished research, we have found that in order to achieve a high-quality /ba/ (defined as 100% identification in quiet), the wide-band burst must exist over a wide frequency range. For example, a well-defined 3 cs burst from 0.3 to 8 kHz will give a strong percept of /ba/, which, if the burst is missing or removed, may likely be heard as /va/ or /fa/. These very low starting (quiet) scores are part of our difficulty in identifying the /ba/ event with certainty, since the 3DDS method requires high scores in quiet for its proper operation.

In all the speech perception tests, /pa, ta, ka/ commonly form a confusion group. This can be explained by the fact that the stop consonants share the same type of event patterns. The relative timing for these three unvoiced sounds is nearly the same. The major difference lies in the center frequencies of the bursts, with the /pa/ burst in the low frequency, the /ka/ burst in the mid-frequency, and the /ta/ burst in the high frequency. Similar confusions are observed for the voiced stop consonants /da/ and /ga/. The bilabial sound /ba/ is more likely to be confused with /va/.

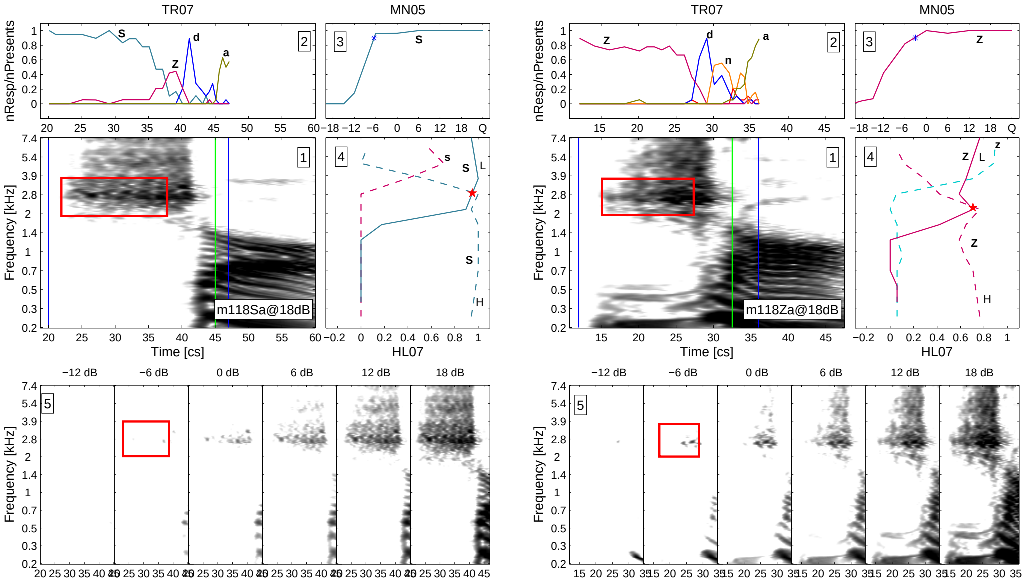
4.2 Results for fricatives

4.2.1 /ʃa/ AND /ʒa/

The dominant perceptual cue for /ʃa/ is summarized in Fig. 4.4(a) panel $\boxed{1}$ by the solid box, as determined by the high-pass and low-pass data of panel $\boxed{4}$ (between 2 and 2.8 kHz) and spanned by ≈ 15–20 cs before the vowel (panel $\boxed{2}$). From panel $\boxed{3}$, the perceptual scores for /ʃa/ are close to 100% at noise levels ≥ -7 dB SNR. The symbol $*$ in panel $\boxed{3}$ indicates SNR$_{90}$ ([i.e., the SNR at which the scores drops to 90% [30]). The *perceptual threshold* (SNR$_{90}$) is a measure of consonant robustness (strength). We say that the *utterance strength* is -7 dB SNR. This is confirmed in panel $\boxed{5}$, where the AI-grams from -12 to $+18$ dB SNR are shown, with a red box around the event region, at its threshold (at -7 dB).

For utterance m118 shown in Fig. 4.4(b), panels $\boxed{1}$ and $\boxed{4}$ indicate the /ʒa/ perceptual cue is at ≈ 2.4 kHz, and from panel $\boxed{2}$, the duration is defined as 15 cs before the vowel (panel $\boxed{2}$) with a strength (panel $\boxed{3}$) of SNR$_{90} = -3$ dB, confirmed by panel $\boxed{5}$.

As seen in panel $\boxed{2}$ of Fig. 4.4(a), $c_{ʃ|ʃ}(t_n)$ remains at $\geq$ 90% for t$_n$ <30



(a) Unvoiced /ʃa/ by m118

(b) Voiced /ʒa/ by m118

Figure 4.4: Hypothetical events for fricatives /ʃa/ and /ʒa/. The multiple panels in each sub-figure are: Panel $\boxed{1}$: AI-gram. A dashed vertical line labels the onset of voicing (sonorance), indicating the start of the vowel. The solid and dashed boxes indicate the dominant and minor events respectively. Panel $\boxed{2}$: CPs as a function of truncation time t_n. Panel $\boxed{3}$: CPs as a function of SNR$_\text{k}$. Panel $\boxed{4}$: CPs as a function of cutoff frequency f_k. Panel $\boxed{5}$: AI-grams of the consonant region [defined by the solid vertical lines on panel $\boxed{1}$], at $-12, -6, 0, 6, 12, 18$ dB SNR. Due to the lack of IPA symbols in Matlab figures, /ʃa/ and /ʒa/ have been denoted by S and Z respectively in the figure.

cs, after which the score drops to $\approx$ 70% for $t_n \leq 35$ cs, and then drops to 0 at 40 cs. Thus, the duration of the /ʃa/ feature is $\approx$ 15 cs. Between 37 and 40 cs, the confusions with /ʒa/ increase to $\approx$ 40%. At 40 cs, truncation of the frication energy causes most listeners (i.e. 80%) to report hearing /da/. Further truncation leads to perception of the vowel.

Panel $\boxed{2}$ of Fig. 4.4(b) shows a small drop in the /ʒa/ score to 75% around 16 cs. As the frication energy is further truncated, the change becomes steep around 25 cs. Once a majority of the frication energy is truncated, only a short-duration high-frequency burst remains, resulting in the strong perception (i.e., 90%) of /da/, at 27 cs [23].

Once the entire frication is removed (30–32 cs), some confusions with /na/

(voiced) are also seen. One possible explanation for these confusions is that the voicing energy in /ʒa/ gives the nasal cue needed to perceive /na/. Confusions with /na/ are not seen for /ʃa/, which is unvoiced, even though the duration and bandwidth of the /ʃa/ and /ʒa/ cue are virtually the same.

The low-pass filtering score for /ʃa/ (solid line labeled L) in panel $\boxed{4}$ of Fig. 4.4(a) shows a sharp increase for $\mathrm{f}_k > 2$ kHz. For the high-pass filtering (dashed line labeled H), for cutoff frequencies above below 4 kHz the score increases significantly, and below 2 kHz, the score reaches its maximum. These results suggest that the /ʃa/ perceptual feature lies between 2 and 4 kHz. Above 2.8 kHz the confusion score for /sa/ [i.e., $c^{H}_{s|ʃ}(f_k)$] steadily increases and at cutoff frequencies of $\approx$ 4 kHz, it reaches 80%. This is in agreement with the /sa/ feature to be discussed in the next subsection, which lies in the range of 4–7.4 kHz

For /ʒa/ (Fig. 4.4(b) panel $\boxed{4}$), the low-pass score touches 80% above cutoff frequencies of 2 kHz. The high-pass score is 100% for cutoff frequencies $\leq$ 2 kHz. This is also where both the high-pass and low-pass curves intersect, indicated by $\star$, thus showing the presence of a feature region. There are high confusions with /za/ at frequencies $\geq$ 4 kHz for the high-pass filtering experiment, which is in agreement with the /za/ feature to be discussed in the next subsection. However, it is significant that in the high-pass filtering data, in the absence of any voicing information, the major confusions are limited to /za/ and /ʒa/ . Since /ʃa/ has a cue similar to /ʒa/, one would have expected confusions between /ʃa/ and /ʒa/ in the high-pass filtering data.

In Fig.4.4(a) panel $\boxed{4}$, there is a sharp drop in score $c^{L}_{ʃ\,|\,ʃ}(f_k)$ at f_k =2 kHz. Since the /ʃa/ cue lies between 2 and 4 kHz, the score would be expected to rise gradually in that range and peak at 100% at f_k >4 kHz for the low-pass filtering experiment. Since this is not the case, it is possible that the sharp transition in spectral energy of the frication portion, and not the entire bandwidth, is the critical feature for /ʃa/ perception. This *low-frequency frication-edge hypothesis* stands true for /ʒa/ as well, although the change is score for /ʒa/ is not as dramatic as for /ʃa/.

The /ʃa/ utterance has $\mathrm{SNR}_{90} = -7$ dB, above which $c_{ʃ\,|\,ʃ}(\mathrm{SNR}_k)$ remains at 100%. Based on the AI-grams of Fig.4.4(a) panel $\boxed{5}$, at 0 dB there is only a weak frication energy, and below -6 dB SNR, it is totally masked.

The /ʒa/ utterance has $\mathrm{SNR}_{90} = -3$ dB for talker m118. panel $\boxed{5}$ of

Fig. 4.4(b) shows that at 0 dB SNR the frication energy at 2.8 kHz is masked. The score reaches chance at −12 dB SNR.

For the remaining five /ʃa/ utterances, three (f109, f103, and m111) have similar confusion patterns for both the truncation and the high-pass and low-pass filtering experiment. Token m115 differs, since the frication energy spans 0.2–1 kHz and is slightly shorter in duration. Although the duration of the utterance f106 is typical, it has relatively weak intensity, explaining its low scores $\geq$ 12 dB SNR.

The remaining five /ʒa/ utterances are consistent with regard to feature region and confusions. Talker f108 is an exception, since even the score in quiet for this utterance is 40%, due to the weak voicing and the frication region. Like f108, m107 is a low-scoring sound due to a weak frication region.

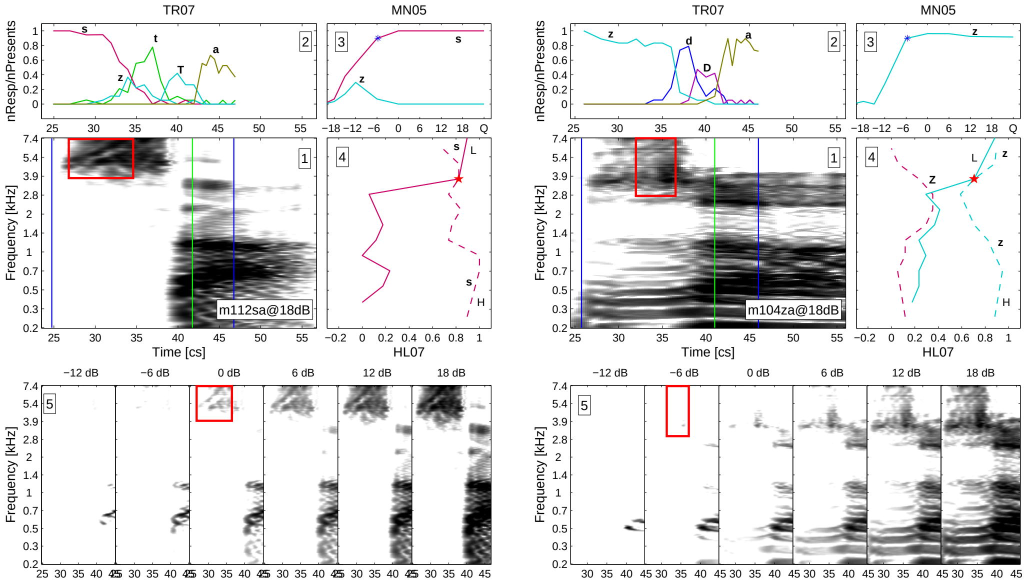
4.2.2 /sa/ and /za/

The AI-gram of utterance /sa/ for male talker 112 is shown in Fig. 4.5(a) panel $\boxed{1}$. The dominant perceptual cue is between 4 and 7.5 kHz and spans ≈10 cs before the start of the vowel. The SNR$_{90}$ is −6 dB for this sound.

The /za/ feature in Fig. 4.5(b) panel $\boxed{1}$ is between 3 and 7.5 kHz and spans for ≈ 7 cs before the vowel. The SNR$_{90}$ is −6 dB SNR.

In panel $\boxed{2}$ of Fig. 4.5(a), $c_{s|s}(t_n)$ remains above 90% for $t_n < 32$ cs after which it steadily drops, and by 37 cs is at 0. As indicated by the box in panel $\boxed{1}$ of 4.5(a), the region before 32 cs is critical to the perception of /sa/. There are minor confusions with /za/ as the score drops, with $c_{z|s}(t_n)$ = 40% at t_n= 35 cs. Also, in agreement with the findings of [23,30], since /ta/ has a high-frequency feature occurring ≈ 5–7 cs before the vocalic portion, confusions with /ta/ are seen at ≈ 37 cs. This is because by then /sa/ has been sufficiently truncated to have a burstlike quality. Thereafter, there are minor (<40%) confusions with /θa/, after which most subjects report only hearing the vowel.

In Fig. 4.5(b) panel $\boxed{2}$, $c_{z|z}(t_n)$ drops slightly at the beginning of the truncation and then remains constant at around 80% until 35 cs, where it begins to drop. Thus, high-frequency frication energy before 35 cs is critical for /za/ perception. Once the /za/ feature is truncated listeners report /da/ (at 37 cs), and then /ða/ (at 40 cs). Beyond 42 cs, only the vowel is reported.



(a) Unvoiced /sa/ by m112

(b) Voiced /za/ by m104

Figure 4.5: Hypothetical events for fricatives /sa/ and /za/. The multiple panels in each sub-figure are: Panel $\boxed{1}$: AI-gram. A dashed vertical line labels the onset of voicing (sonorance), indicating the start of the vowel. The solid and dashed boxes indicate the dominant and minor events respectively. Panel $\boxed{2}$: CPs as a function of truncation time t_n. Panel $\boxed{3}$: CPs as a function of SNR$_\mathrm{k}$. Panel $\boxed{4}$: CPs as a function of cutoff frequency f_k. Panel $\boxed{5}$: AI-grams of the consonant region [defined by the solid vertical lines on panel $\boxed{1}$], at $-12, -6, 0, 6, 12, 18$ dB SNR. Due to the lack of IPA symbols in Matlab figures, /ʒa/, /θa/, and /ða/ have been denoted by Z, T, and D, respectively in the figure.

This is consistent with the results reported in [23] wherein the /da/ cue was found to be a burst of energy at frequencies above 4 kHz.

It is interesting to note that a few listeners (30%) report /za/ when they hear /sa/ at $t_n = 34$ cs. Thus, a truncated /sa/ is reported as /za/, showing that the /za/ and /sa/ differ in the duration of the frication energy. This is confirmed by the fact that the /za/ truncation score drops only after 37 cs as opposed to the /sa/ scores, which start dropping at ≈ 30 cs. The frication region begins at around 27 cs for both /sa/ and /za/.

The /sa/ low-pass filtering experiment data (solid lines) in Fig. 4.5(a) panel $\boxed{4}$ shows that once the cutoff frequency goes above 3 kHz, $c^L_{s|s}(f_k)$ abruptly rises to 80% and touches a maximum score of 90% at full bandwidth. Since

/sa/ has a high-frequency cue, for the high-pass filtering experiment, $c^H_{s|s}(f_k)$ always remains above 80%. The scores peak at 1 kHz to 100%, after which they dip again slightly. The high-pass and low-pass curves intersect at $\approx$ 4 kHz which is where the frication energy of /sa/ lies. This high-frequency region is clearly critical to the perception of /sa/.

For /za/, the low-pass filtering score $c^L_{z|z}(f_k)$ in Fig. 4.5(b) panel $\boxed{4}$ rises at cutoff frequencies above 4 kHz. The high-pass filtering score remains quite high all through. There is a brief dip in the score from around 1 to 4 kHz, which is an indication of an interfering cue, in this case that of /ʒa/. This is because, as seen previously, /ʒa/ also has a high-frequency cue slightly lower in frequency compared to /za/. The low-pass and high-pass curves are close to one another at $\approx$ 4 kHz and 5.5 kHz, and this defines the region where the feature is present.

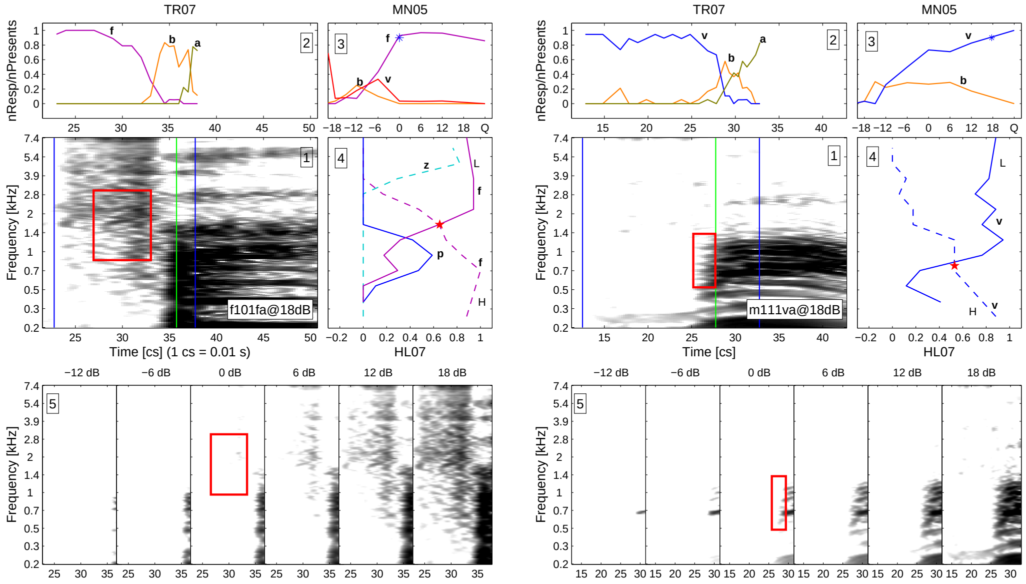
The low-frequency frication-edge hypothesis holds based on the abrupt drop in /sa/ and /za/ low pass scores around 3.9 kHz.

In Fig. 4.5(a) panel $\boxed{3}$, the /sa/ feature has $\text{SNR}_{90} = -6$ dB. Between -6 and -18 dB SNR, $c_{s|s}(\text{SNR}_k)$ drops from 90% to chance. At -12 dB, there are voicing confusions with /za/ and $c_{z|s}(\text{SNR}_k)$ goes up to 30%. The sound /za/ is actually just a voiced equivalent of the fricative /sa/, but shorter in duration, which explains this confusion.

In Fig. 4.5(b) panel $\boxed{3}$, $c_{z|z}(\text{SNR}_k)$ remains above 90% for white noise levels above -6 dB SNR. The AI-grams in Fig. 4.5(b) panel $\boxed{5}$ clearly show that at noise levels below -6 dB SNR, most of the high-frequency energy in the utterance is masked, leading to the steep drop in score below -6 dB.

For /sa/, the utterances f108, f109 and f113 have very similar scores and confusions as that of m112 discussed in this section. Utterance m111 deviates from this due to its frication region, which spans 1–7.4 kHz. Utterance m117, on the other hand, is consistently reported as /θa/. This can be attributed to its unusually short duration and narrow band frication region.

The /za/ feature is quite robust to white noise to levels as high as -12 dB-SNR with the exception of utterance f109, which has a weak high-frequency narrow band (6.5–7.4 kHz) frication region, explaining its low scores in all three experiments.



(a) Unvoiced /fa/ by f101

(b) Voiced /va/ by m111

Figure 4.6: Hypothetical events for fricatives /fa/ and /va/. The multiple panels in each sub-figure are: Panel $\boxed{1}$: AI-gram. A dashed vertical line labels the onset of voicing (sonorance), indicating the start of the vowel. The solid and dashed boxes indicate the dominant and minor events respectively. Panel $\boxed{2}$: CPs as a function of truncation time t_n. Panel $\boxed{3}$: CPs as a function of $\mathrm{SNR_k}$. Panel $\boxed{4}$: CPs as a function of cutoff frequency f_k. Panel $\boxed{5}$: AI-grams of the consonant region [defined by the solid vertical lines on panel $\boxed{1}$], at $-12, -6, 0, 6, 12, 18$ dB SNR.

4.2.3 /fa/ and /va/

For the utterance f101/fa/, an important feature region is between 0.9 and 2.8 kHz maskable by noise at 0 dB. This outlined portion of /fa/ is $\approx$ 7 cs before the vocalic portion.

The voiced /va/ feature is between 0.5 and 1.5 kHz, as highlighted in panel $\boxed{1}$ of Fig. 4.6(b). The /va/ sound is not robust to white noise masking with $\mathrm{SNR_{90}}$ = 18 dB, with scores of 100% only in quiet.

The fricatives /fa/ and /va/ are also characterized by a wide-band frication energy from 1 to 7.5 kHz, with /va/ being especially prone to even small amounts of masking noise. It is this that dramatically improves the perception of these sounds.

Figure 4.6(a) panel $\boxed{2}$ shows that the percent correct score $c_{f|f}(t_n)$ remains at 100% for $\mathrm{t}_n \leq 28$ cs. Further truncation leads to a gradual drop in scores for 28 cs $\leq$ t$_n$ $\leq$ 33 cs. When portions beyond this are truncated, scores drop gradually to 0 at $\approx$ 34 cs by which the frication energy spanning 0.7–7.4 kHz is completely truncated, leading to 80% of the subjects reporting /ba/. This is in line with the findings of [23] wherein /ba/ was seen have a low-frequency cue onset simultaneously with the vocalic portion. Accordingly, when utterance f101 is truncated to $\approx$ 37 cs, the /ba/ cue has been truncated and listeners stop reporting /ba/ and hear only the vowel.

In Fig. 4.6(b) panel $\boxed{2}$, except for a slight anomalous dip in $c_{v|v}(t_n)$ at $\approx$ t$_n$ =17 cs, the /va/ score remains around 90%. At 25 cs, where the vowel starts, the score begins to drop, and at 28 cs, $c_{v|v}(t_n) = 0$. Once the /va/ cue has been truncated, most subjects report /ba/, since it has a low-frequency cue. Beyond 30 cs, the vowel dominates.

Both /va/ and /fa/ are characterized by a high-frequency frication energy as well as a strong cue at mid-frequencies. Since the frication energy is highly susceptible to noise masking, the /fa/ and /va/ sounds are not as robust as the other fricatives.

In Fig. 4.6(a)panel $\boxed{4}$, for low-pass cutoff frequencies f_k greater than 1.2 kHz, $c^L_{f|f}(f_k)$ climbs steadily and it reaches 100% by $\approx$ 2.2 kHz. For $f_k <$ 1.2 kHz, there are strong confusions with the sound /pa/ with $c^L_{p/f}(f_k) =$ 60% at $f_k = 0.8$ kHz. This again, is in agreement with previous findings [23] wherein /pa/ was defined by a low-frequency cue and a wide-band click. Even with only the low-frequency cue, subjects report /pa/. In the high-pass experiment, the score $c^H_{f/f}(f_k)$ steadily increases for cutoff frequencies above 4 kHz and touches 100% at $\approx$ 0.8 kHz. For cutoff frequencies of greater than 2.8 kHz, most subjects report hearing a /za/. From these curves, we can conclude that the perceptual feature for /fa/ is in the range of 1–2.8 kHz where the high-pass and low-pass curves intersect.

In Fig. 4.6(b) panel $\boxed{4}$, $c^L_{v|v}(f_k)$ starts rising at f_k =0.5 kHz and by $f_k =$ 1.2 kHz the score stabilizes to around 87%. On the other hand, $c^H_{v|v}(f_k)$ steadily increases for $f_k >$ 1.7 kHz with a full-band score of around 90%. The intersection of the high-pass and low-pass curves is at $\approx$ 0.7 kHz. It is to be remembered that m111/va/ is a fricative that does not have a noise-robust frication energy in the high-frequency region, explaining the low scores in all three experiments.

In Fig. 4.6(a) panel $\boxed{3}$, the score $c_{f|f}(\text{SNR}_k) \approx 90\%$ for SNR$_k$ > 0 dB SNR, then drops sharply. From the AI-grams in Fig. 4.6(a) panel $\boxed{5}$, it is clear that by 0 dB SNR, the frication portion is completely masked, in agreement with he SNR_{90} threshold. Between -12 and -6 dB SNR, /fa/, /va/ and /ba/ have almost equal scores at $\approx 30\%$ each. It is interesting to note that /ba/, /va/, and /fa/ seem to form a confusion group. This is because both /va/ and /fa/ are characterized by long duration frication energy above 1 kHz and a mid-frequency cue at ≈ 0.9 kHz, that precedes the vocalic portion. Once these cues are masked, the low-frequency energy is identified as /ba/.

As mentioned above in Fig. 4.6(b) panel $\boxed{3}$, the /va/ cue is poorly articulated, $c_{v|v}(\mathrm{SNR}_k)$ drops at 18 dB SNR and by 0 dB SNR is 60%. Owing to the low scores even at high values of SNR, it is not possible to make strong conclusions about the /va/ feature.

/fa/, /ba/, and /va/ have confusions with each other to a large extent in all three experiments.

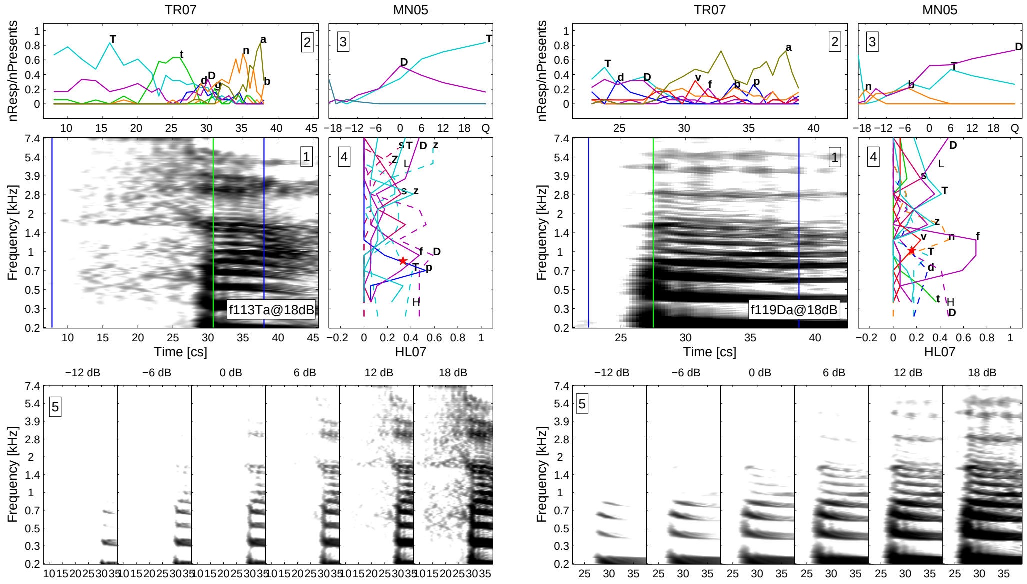
For /fa/, utterance m111 was similar to f101 discussed here. Utterance m112 was an aberration since it had low scores for all three experiments, owing to its weak frication region. Utterance m117 had absolutely no frication region at all and thus had scores $c_{f|f}(\mathrm{SNR}_k)$ as low as 60% at 12 dB SNR.

For /va/, the utterances f108 and m104 have confusions similar to m111 discussed here. The utterances f105 and f103 were identified as /fa/ showing that they are poorly articulated.

Looking at the effect of the frication region on the scores on both /fa/ and /va/, it seems that although the mid-frequency cue is enough for the discrimination of /fa/ or /va/, it is the high-frequency frication cue that leads to high perceptual scores. The release of the frication portion as a burst is especially salient. Moreover, the low scores, even in quiet conditions, are a major shortcoming. Future scope for confirming the /fa/ and /va/ features would be to design experiments using speech-weighted noise to judge the importance of the frication region, with higher SNRs.

4.2.4 /θa/ and /ða/

For /θa/, taking the case of talker f113, it has been impossible to ascertain any particular feature region. In Fig. 4.7(a) panel $\boxed{2}$, even with no truncation, $c_{\theta|\theta}(t_n)$ starts at 60% with a great deal of variation in the score. Even at its maximum it only reaches a score of 80%. Thereafter, once the frication



(a) Unvoiced /θa/ by f113

(b) Voiced /ða/ by f119

Figure 4.7: Hypothetical events for fricatives /θa/ and /ða/. The multiple panels in each sub-figure are: Panel $\boxed{1}$: AI-gram. A dashed vertical line labels the onset of voicing (sonorance), indicating the start of the vowel. The solid and dashed boxes indicate the dominant and minor events respectively. Panel $\boxed{2}$: CPs as a function of truncation time t_n. Panel $\boxed{3}$: CPs as a function of $\mathrm{SNR_k}$. Panel $\boxed{4}$: CPs as a function of cutoff frequency f_k. Panel $\boxed{5}$: AI-grams of the consonant region [defined by the solid vertical lines on panel $\boxed{1}$], at $-12, -6, 0, 6, 12, 18$ dB SNR. Due to the lack of IPA symbols in Matlab figures, /ʒa/, /θa/, and /ða/ have been denoted by Z, T, and D respectively in the figure.

portion has been truncated, the confusions spread out with high entropy. This is true even for the filtering data. The /θa/ scores are especially low for the filtering experiment. Even in the quiet condition $c_{\theta|\theta}(\mathrm{SNR_k}) = 80\%$. There is a significant confusion with /ða/ with $c_{D|\theta}(\mathrm{SNR}_k)$ around 50% at 0 dB SNR. Also, the variation of the confusions is significant across different utterances. This is true for the truncation and the filtering experiments. Owing to such low scores and huge variability, zeroing in on a feature region has not been possible.

Much like /θa/, $c_{D|D}(SNR_k)$ is low at high SNRs. The scores for /ða/ remain low even for the other two experiments. Both /θa/ and /ða/ are characterized by a wide-band noise burst at the onset of the consonant almost in line with the vocalic portion. Owing to this, chances of confusions or alterations are seen to be maximized in the case of these sounds. Again, it is difficult to make any sort of conjecture with /θ/ and /ð/ based on the 3DDS method.

CHAPTER 5

ROBUSTNESS OF THE FEATURES

The features discussed in the Results section are consistent across all the 96 utterances present in the whole study.

As masking noise is increased for a particular utterance, the score for the utterance remains unaffected till the primary perceptual cue is masked. Once this critical cue is masked, the score for the consonant drops abruptly to chance. The same was reported by Regnier and Allen (2008) [30] that a threshold of 90% correctness (SNR_{90}) is directly proportional to the threshold of the /t/ burst (SNR_e) based on the prediction of AI-gram. For the truncation experiment, this drop in score is much more abrupt in the case of the stop consonants than for the fricatives because stops are short. For a particular utterance (a point on the plot), the psychological threshold SNR_{90} is interpolated from the PI function, while the threshold of audibility for the dominant cue is estimated from the AI-gram plots [panel (5)] of Figs. 4.1(a)–4.2(b). The two thresholds are nicely correlated, indicating that the recognition of each stop consonants is mainly dependent on the audibility of the dominant cues. Speech sounds with stronger cues are easier to hear in noise than weaker cues because it takes more noise to mask them. When the dominant cue becomes masked by noise, the target sound is easily confused with other consonants within the same group.

Scatter plots showing the correlation between the threshold of consonant identification and the audible threshold of dominant cues for both stops and fricatives are given in Figures 5.1 and 5.2.

Figure 5.1 and 5.2 shows the high correlation between the threshold of the acoustic feature of the fricative and SNR_{90}. This plot clearly shows that the identification of a speech sound is determined by the audibility of the acoustic feature. When this primary feature is masked, confusions result. The more intense the perceptual cue, the more robust the sound is to masking noise. Typically, a 3–6 dB gap exists between the threshold of the acoustic cue

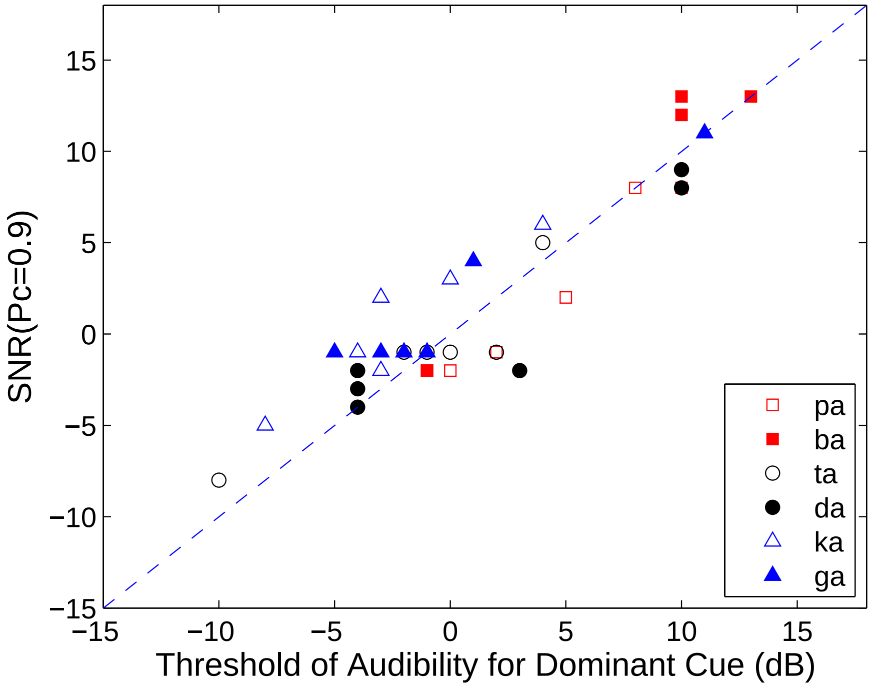


Figure 5.1: Correlation between SNR_e and SNR_{90} for the stops having scores greater than 90% in Quiet.

and the value of SNR_{90}. As an example, the utterance m118Sa, which was discussed in the previous chapter, is shown in the scatter plot for the fricatives (Fig. 5.2). As can be seen from Fig. 4.4(a) panel $\boxed{3}$, the SNR_{90} point for this utterance was around −7 dB. From the AI grams, it was observed that the threshold of audibility for the feature was ≈ -7 dB as well. This means that when the audibility of the feature region was affected by the masking noise, the score for m118Sa started to drop.

As expected, the sounds /va/ and /fa/ have SNR_e as high as ≈ 10 dB SNR. Once this feature is masked, the scores drop rapidly. On the other hand, /ʃa/ and /ʒa/, both have $\mathrm{SNR}_e \approx -8$ dB SNR.

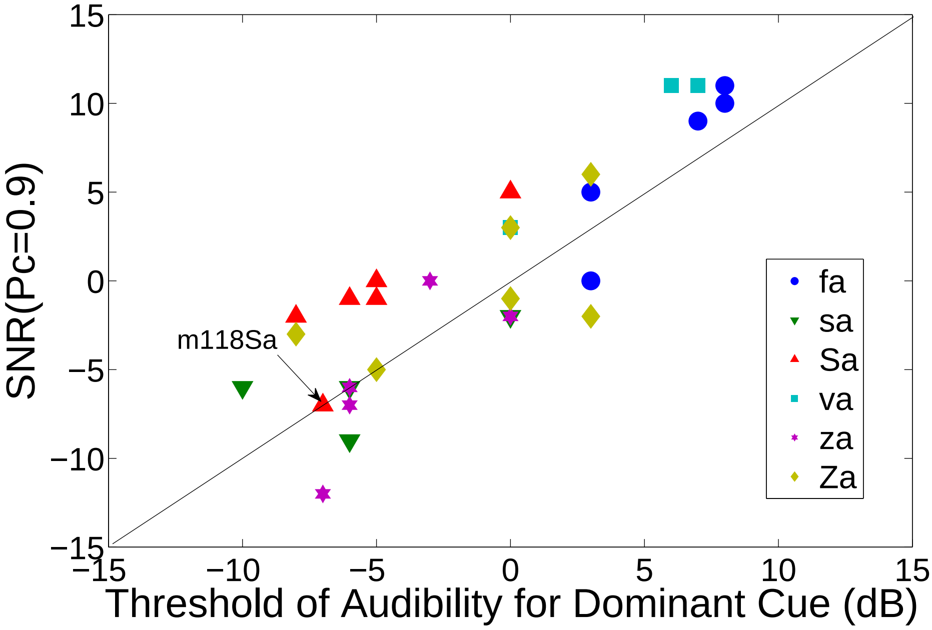


Figure 5.2: Correlation between SNR_e and SNR_{90} for the fricatives having scores greater than 90% in Quiet.

CHAPTER 6

THE VERIFICATION STUDY

6.1 Introduction

The 3DDS method identifies the time-frequency feature coordinate for each sound using three independent experiments [2, 23]. The truncation experiment gives information about the time coordinate while the high-pass/low-pass experiments decide the frequency co-ordinate of the feature. The noise masking experiment defines the threshold for the feature in the presence of masking noise. Once the feature regions for the different CVs were known, a verification study was needed to confirm the results of the 3DDS experiment.

The verification experiment is the unique contribution of the author. It verifies the role of the perceptual cues for the identification of a sound. To this end, sounds were modified so as to remove the feature region identified by the 3DDS method. The feature removal was done using an STFT based analysis-synthesis algorithm [1]. These modified sounds were then played as stimuli to subjects. The hypothesis is that when the feature region is removed, the perception of the sound should be completely altered.

6.2 Methods

A total of 23 subjects completed the study. All subjects self-reported no history of speech or hearing disorder. The first language for all the subjects was English. The subjects were paid for their participation. IRB approval was obtained prior to the experiment.

6.2.1 Speech stimuli

Twelve CVs, namely /p/, /t/, /k/, /b/, /d/, /g/, /s/, /ʃ/, /z/, /ʒ/, /f/, /v/, /followed by the vowels /a/, /e/, were chosen from the University of Pennsylvania's Linguistic Data Consortium (LDC-2005S22, aka *the Fletcher AI corpus*) for the verification experiment. The speech sounds were sampled at 16 kHz. A total of 18 talkers were used in the experiment. Twelve utterances were presented for each sound, of which 10% were unmodified and were used as controls. The speech sounds were presented diotically (both ears) through a Sennheiser HD-280 PRO circumaural headphone, adjusted in level at the listener's *Most Comfortable Level* (MCL), i.e., ≈ 70–80 dB SPL. All experiments were conducted in a single-walled IAC soundproof booth, situated in a lab with no windows, with the lab outer door shut. People in the lab were instructed to speak softly so as to not distract the subject under test.

6.2.2 Conditions

The motive of the verification experiment was to evaluate the results of the 3DDS method for the stop and fricative consonants. The feature regions found using the 3DDS method were removed using a software written in Matlab (Beren). There were 9–10 modified tokens for each CV pair. Each token was played to the subjects at 4 different SNRs namely, −6 dB, 0 dB, 6 dB, and 12 dB SNR. Speech-weighted noise was used to mask the speech sounds. The conditions of the verification experiment were similar to the study reported by [27], which was used as control data.

6.2.3 Procedure

A Matlab program was created for the data collection. A mandatory practice session with feedback was given at the beginning of each experiment. The purpose of this was to familiarize the subjects with the sounds for which one token per sound (at least) was played to the subjects. Speech tokens were randomized across the talkers, conditions and CV pairs. Even though the test set for the experiment had 12 consonants, the response set had all 16 consonants that were used in Phatak and Allen (2007) [27]. Following each stimulus presentation, subjects responded by clicking on a button labeled with the consonant that they thought they heard. In case the CV was completely masked by the noise, the subject was instructed to click on a "Noise Only" button. In case the subjects thought that what they heard was not on the screen, they could click on a button labeled "Other", which had some more commonly reported options. Subjects used the "Noise Only" option only 1% of the time. The "Other" button was used around 2.5% of the time. More than half of these were at higher SNRs such at 12 and 6 dB SNR. Subjects were allowed to play each token up to 3 times, after which the token was pushed to the end of the list. They were also encouraged to take frequent breaks. The waveform was played via a SoundBlaster 24-bit sound card in a PC Intel computer, running Matlab under Ubuntu Linux.

6.2.4 Control data

Owing to the large number of CVs used in the verification experiment, it would not have been possible to have unmodified counterparts for each modified utterance without affecting the length of the experiment dramatically. Moreover, a very similar study (named MN64) with unmodified sounds had already been conducted by Phatak and Allen (2007) [27]. For these reasons, only 10% of the sounds used in the verification study were unmodified as opposed to 100%. The purpose of these controls was to show that these sounds were similar to those used in MN64, so that the whole of the MN64 data could be used to gain information about unmodified sounds.

There were, however, several fundamental differences between MN64 and the verification experiment. Firstly, the SNRs at which the tests were conducted were different for the two experiments. Since MN64 aimed at identifying the perceptual thresholds of different speech sounds, the masking noise could go as low as -22 dB. The values used were $-22, -20, -16, -10, -2$ dB and Quiet. The verification experiment, on the other hand, aimed at modifying sounds and studying the effects of such modifications. It was essential that these modifications be heard. Thus, high SNRs, namely -6, 0, 6 and 12 dB were used.

The test procedure was also different for the two experiments. The verification study had a larger response set. Apart from the 16 CVs used in MN64, the verification study also had an "Other" option with commonly reported

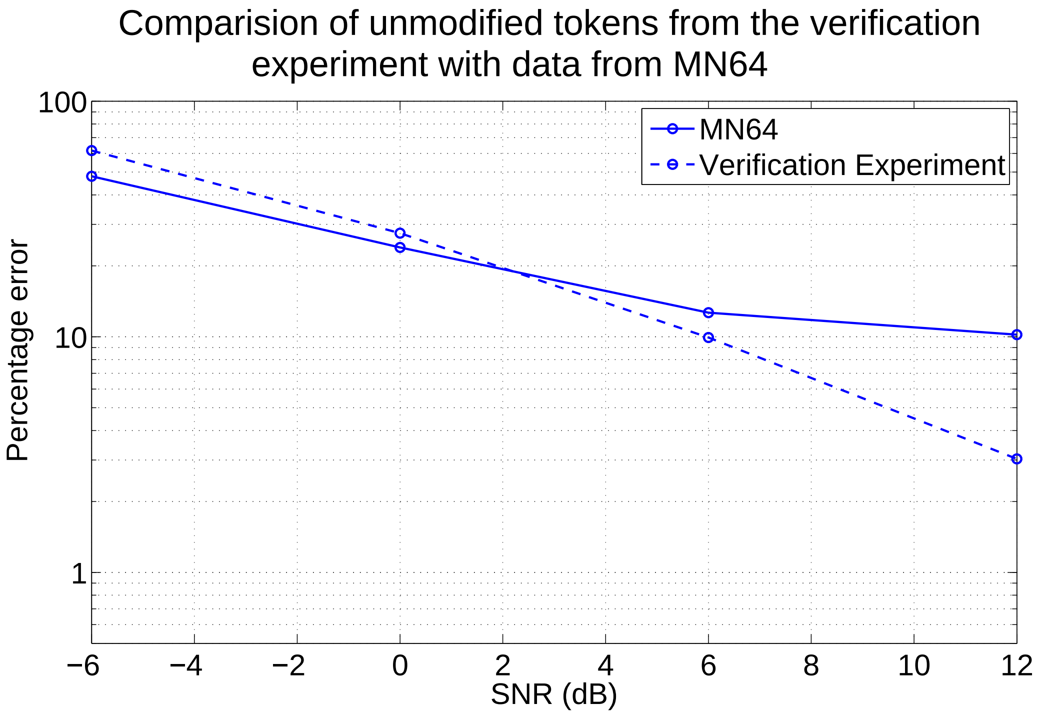


Figure 6.1: A comparison between the average error of MN64 (dashed line) and the verification experiment (solid line).

sounds such as /la/, /ra/ and /ha/. Moreover, the number of times a sound could be repeated was different for the two experiments.

The algorithms used for calculating the noise level in the two cases was also different. To make a careful comparison, the speech and noise RMS were computed for each utterance across each noise condition used in MN64, using the algorithms used in MN64 and the verification study. Thereafter, the true SNR was found by using the speech RMS from the verification study and the noise RMS from MN64. This analysis was done on an utterance to utterance basis and the errors were interpolated for −6, 0, 6 and 12 dB SNR. Fig. 6.1 shows a comparison of the two. A 3–10% difference is observed between the two curves.

Stimuli that had more than 20% error at 12 dB SNR, across all listeners, for both MN64 and the verification study were discarded [27]. Listeners with an average score of less than 80%, at 6 and 12 dB SNR, were not used for the analysis.

CHAPTER 7

RESULTS FOR THE VERIFICATION STUDY

The 3DDS experiment was conducted using CVs with the vowel /a/. Once the feature regions for these CVs were analyzed, those features for the vowel /e/ were deduced and then used as stimuli for the verification experiment. The results obtained were similar to that with the vowel /a/.

The features for /p/, /b/, /t/, and /d/ were largely invariant across both vowels /a/ and /e/. The sounds /pe/ and /be/ were characterized by low-frequency bursts, while /te/ and /de/ were characterized by high-frequency bursts. However, the duration for these bursts was altered significantly for /e/ compared to /a/. Since the burst location is largely determined by the location of the second formant [10, 11, 23, 24], the bursts for /ke/ and /ge/ were much higher in frequency compared to those for /a/.

The fricative cue is typically longer in duration than for a stop consonant. There seems to be no significant change in duration across vowels. The frication energy was more or less invariant for /a/ and /e/.

In the following discussion, for ease of comparison, the results have been paired by consonant. The results for the stops are presented first followed by the fricatives. Thus, /ta/ and /te/ have been explained first, followed by /ka/ and /ke/, and so on.

For the first CV (/ta/ and /te/), shown in Figures 7.1 and 7.2, a set of four panels are shown per CV (8 total). Panels $\boxed{1}$ and $\boxed{2}$ show the AI-grams of the unmodified and modified sounds, respectively. The unmodified sounds are basically data from the MN64 experiment, used as control data in this study. The portion highlighted by the solid box in panel $\boxed{2}$ denotes the feature region for the sound, which has been removed. Panel $\boxed{3}$ shows the comparison of the errors of the data from the unmodified tokens used in the verification experiment (solid line) to the data from MN64 (dashed line). Panel $\boxed{4}$ shows the confusion patterns for the sound. Only significant confusions are shown.

For all subsequent CV pairs, the top two and bottom two panels are merged to form two panels, as opposed to four. Thus, for each CV, the panel to the left shows the AI-gram of the sound, with the feature region to be removed, highlighted in the solid box while the right panel shows the total error for the sound in the Verification experiment (dashed line with diamond markers) and MN64 (dashed line with circle markers), as well as the scores for the confusion pattern (solid lines).

7.1 Results for stops

7.1.1 /ta/ and /te/

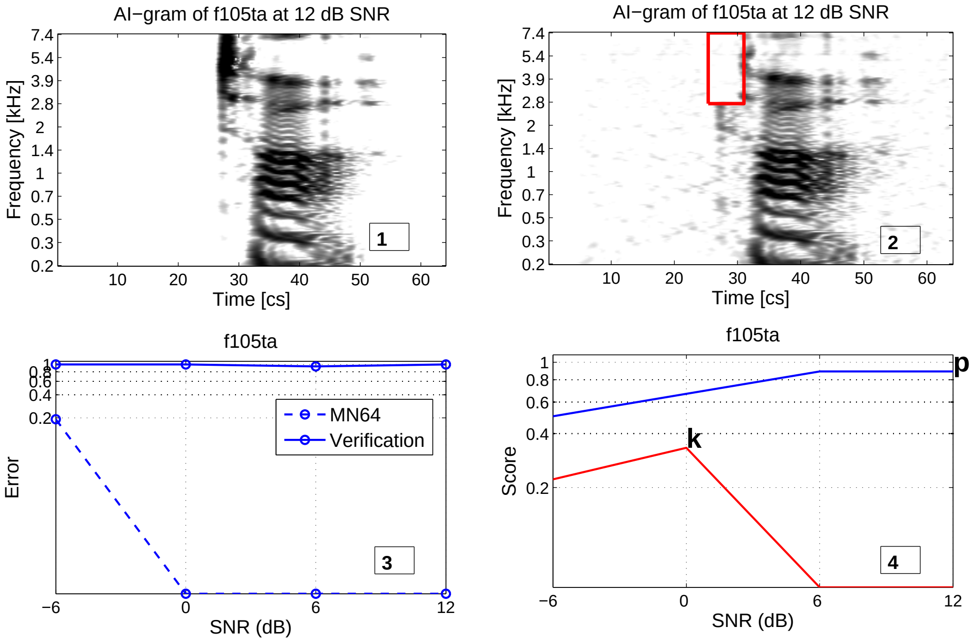


Figure 7.1: Panel $\boxed{1}$: AI-gram of the unmodified f105ta. Panel $\boxed{2}$: AI-gram for the modified sound with the removed feature region, highlighted with a box. Panel $\boxed{3}$: Comparison of the errors for f105ta between the verification experiment and the control data from MN64. Panel $\boxed{4}$: Confusion patterns for f105ta

Figures 7.1 and 7.2 show the results for the consonant /t/ with the vowels /a/ and /e/ respectively. The /ta/ sound has been studied in great detail in the past [23, 30]. The /ta/ feature is a high-frequency (above 4 kHz) burst preceding the vowel by ≈ 7 cs. When the /ta/ feature is removed, the error goes to 100% at all SNRs, as shown by the solid line in panel $\boxed{3}$ of Fig. 7.1. The control data (dotted lines in panel $\boxed{3}$ of Fig. 7.1) shows that the error for /ta/ is 0 when the high-frequency burst is present. We conclude that the burst is critical to /ta/ perception. In the absence of the /ta/ cue (Fig. 7.1 panel $\boxed{4}$), most subjects (90% at 12 and 6 dB SNR) perceive a /pa/ while 1/3 (around 35% at 0 dB SNR) report /ka/. The /pa,ta,ka/ confusions group is commonly seen [26, 27].

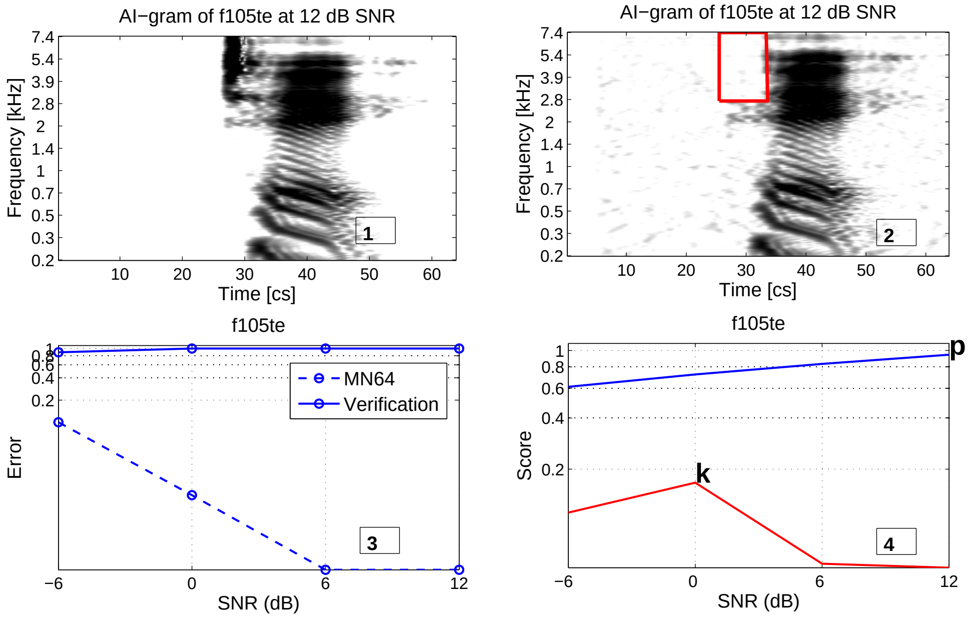


Figure 7.2: Panel $\boxed{1}$: AI-gram of the unmodified m117te. Panel $\boxed{2}$: AI-gram for the modified sound with the feature region that has been removed, highlighted with the solid box. Panel $\boxed{3}$: Comparison of the errors for m117te between the verification experiment and the control data from MN64. Panel $\boxed{4}$: Confusion patterns for m117te.

The /te/ feature, like /ta/, is a high-frequency burst preceding the vowel by ≈ 7 cs, as shown in panel $\boxed{2}$ of Fig. 7.2. When this feature is removed (Fig. 7.2 panel $\boxed{3}$, solid line), the error for /te/ goes to 100% across all SNRs. In the presence of the cue (Fig. 7.2 panel $\boxed{3}$, dashed line), the /te/ error is 0 at 12 dB SNR. This high-frequency burst is critical for /te/ perception.

The source of this confusion group seen in both Fig. 7.1 and 7.2 is the presence of *conflicting cues* [23]. As seen in panel $\boxed{1}$ of Fig. 7.1, the /ta/ utterance, apart from a high-frequency burst above 4 kHz, also has a mid-frequency burst at around 2 kHz and a low-frequency burst at 1 kHz. This leads to the perception of either a /ka/ or a /pa/, depending on which burst is stronger, in the absence of the /ta/ burst. As with the vowel /a/, the /te/ sound has /ke/ and /pe/ conflicting cues. Of these, the /pe/ cue is stronger. To be discussed later is the /ke/ feature, in a region that overlaps with the /te/ feature. A large part of the /ke/ cue has been masked while masking the /ta/ feature for the verification experiment. Owing to this, in the absence of the /te/ feature, most subjects report /pe/ across all SNRs. This is because, as in the case of /ta/, even /te/ has the conflicting /pe/ cue.

7.1.2 /ka/ and /ke/

The 3DDS method shows the /ka/ feature to be a burst at ≈ 2 kHz preceding the vowel by ≈ 7 cs (Fig. 4.2(a)). From Fig. 7.3 panel $\boxed{2}$, the dashed line with diamond markers show that when this cue is removed, the error for /ka/ goes to nearly 100% . The dashed lines with circle markers in panel $\boxed{2}$ of Fig. 7.3 show that when subjects hear the mid-frequency burst, the error for /ka/ is 0 up to 0 dB SNR. Like /ta/, /ka/ also has conflicting /pa/ and /ta/ cues. This is why, in the absence of the /ka/ cue, nearly 90% of the subjects report a /pa/ at 12 dB SNR while ≈ 20% report a /ta/ at 0 dB SNR. The presence of the burst at 1 kHz as seen in panel $\boxed{1}$ of Fig. 7.3, leads to the /pa/ perception that is observed.

Panel $\boxed{3}$ of Fig. 7.3 shows that the burst energy for /ke/ is not in the mid-frequency region as in the case of /ka/, but is present at much higher frequencies. The /ke/ feature location is dependent on the second formant location which is higher for the vowel /e/. Thus, the /ke/ feature is located as highlighted by the solid box in panel $\boxed{3}$ of Fig. 7.3. When this feature is removed, the errors for /ke/ go to 100% across all SNRs (Fig. 7.3 panel $\boxed{4}$ dashed line with diamond markers). When the feature is audible, the error for /ke/ remains under 20% even at 0 dB SNR. Since the /te/ conflicting cue doesn't exist as seen in panel $\boxed{3}$ of Fig. 7.3, /pe/ is heard ≈ 90% of the time. At −6 dB however, subjects report /pe/ and /ðe/ equally.

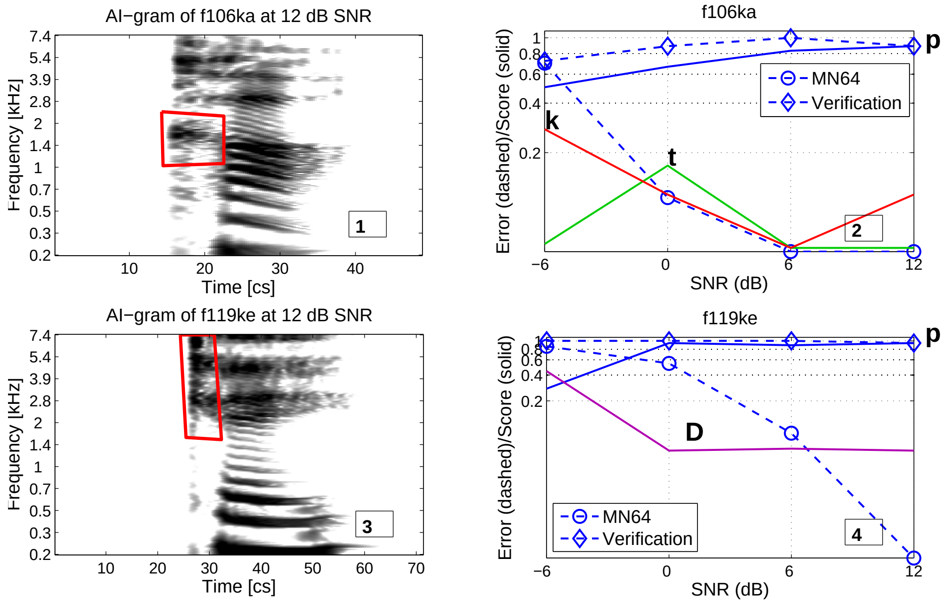


Figure 7.3: Panel $\boxed{1}$: AI-gram of the unmodified f106ka with feature region highlighted in the solid box. Panel $\boxed{2}$: Comparison of errors of the unmodified (dashed line with circle markers) and the modified (dashed line with diamond markers) sounds along with the scores of the confusion patterns (solid lines) for the modified sound. Panel $\boxed{3}$: AI-gram of the unmodified f119ke with feature region highlighted in the solid box. Panel $\boxed{4}$: Comparison of errors of the unmodified (dashed line with circle markers) and the modified (dashed line with diamond markers) sounds along with the scores of the confusion patterns (solid lines) for the modified sound. Due to the lack of IPA symbols in Matlab figures, /ða/ has been denoted by D in the figure.

7.1.3 /pa/ and /pe/

The feature region for /pa/ according to the 3DDS method, is a burst at around 1 kHz and a wide-band noise click [23]. However, removal of this feature region, as highlighted by the solid box in panel $\boxed{1}$, does not seem to alter the perception as drastically as is seen for other utterances. In the absence of the /pa/ cue, the error goes up to 30% at 6 dB SNR. The feature region for /pa/ needs to be further investigated to get a clear picture.

For /pe/, the cue is invariant and it is just a low-frequency burst and a wide-band noise click preceding the vowel by ≈ 7 cs. The feature region is highlighted in panel $\boxed{3}$ of Fig. 7.4. When the feature region is removed, there is an increase in the error from 0 to 60%. Even though the change

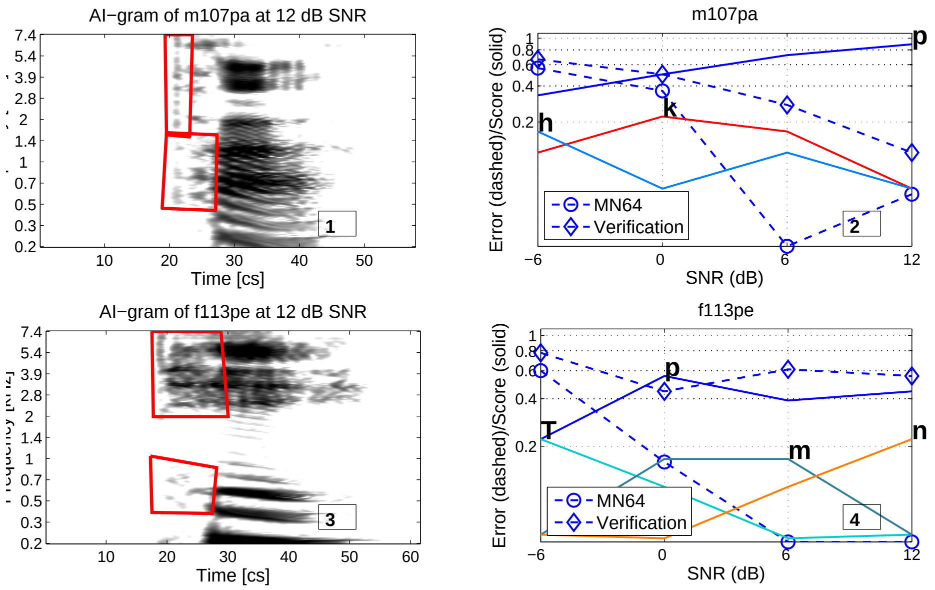


Figure 7.4: Panel $\boxed{1}$: AI-gram of the unmodified f109pa with feature region highlighted in the solid box. Panel $\boxed{2}$: Comparison of errors of the unmodified (dashed line with circle markers) and the modified (dashed line with diamond markers) sounds along with the scores of the confusion patterns (solid lines) for the modified sound. Panel $\boxed{3}$: AI-gram of the unmodified f113pe with feature region highlighted in the solid box. Panel $\boxed{4}$: Comparison of errors of the unmodified (dashed line with circle markers) and the modified (dashed line with diamond markers) sounds along with the scores of the confusion patterns (solid lines) for the modified sound. Due to the lack of IPA symbols in Matlab figures, /θa/ has been denoted by T in the figure.

in error is not as drastic as in the other cases, it still is quite high. The highlighted region is thus, important to /pe/ perception. Looking at the confusion patterns in panel $\boxed{4}$ of Fig. 7.4, there seem to be minor confusions with /me/ and /ne/ but they both have scores below 20%. More analysis may be necessary to come decisively to a conclusion about the /pe/ cue.

7.1.4 /da/ and /de/

A high-frequency burst that comes on concurrently with the vowel defines /da/ [23]. When this feature region, as highlighted by the solid box in panel $\boxed{1}$ of Fig. 7.5, is removed, the perceptual errors for/da/ go to 100% from 0 as seen in panel $\boxed{2}$ of Fig. 7.5 (dashed lines). With the increase in masking

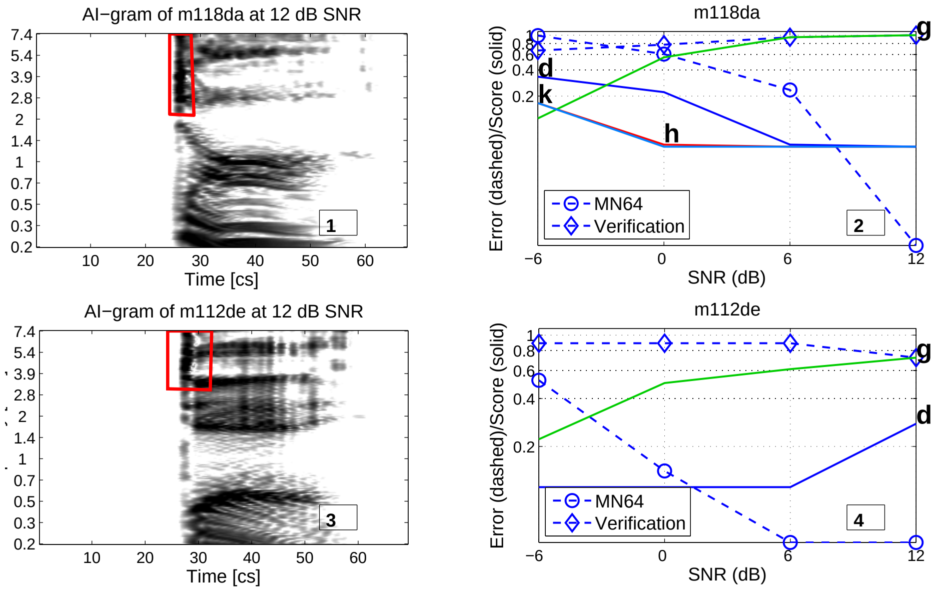


Figure 7.5: Panel $\boxed{1}$: AI-gram of the unmodified m118da with feature region highlighted in the solid box. Panel $\boxed{2}$: Comparison of errors of the unmodified (dashed line with circle markers) and the modified (dashed line with diamond markers) sounds along with the scores of the confusion patterns (solid lines) for the modified sound. Panel $\boxed{3}$: AI-gram of the unmodified m112de with feature region highlighted in the solid box. Panel $\boxed{4}$: Comparison of errors of the unmodified (dashed line with circle markers) and the modified (dashed line with diamond markers) sounds along with the scores of the confusion patterns (solid lines) for the modified sound.

noise however, there is an improvement in /da/ scores (scores up to 40%). Thus, the high-frequency burst is very important to /da/ perception. It is significant that at 12 dB SNR, the audibility of the second formant transition does not contribute to any /da/ perception. panel $\boxed{1}$ shows that apart from the high intensity burst above 4 kHz, the /da/ utterance also has a (somewhat weaker) mid-frequency burst as well. A mid-frequency cue almost coinciding with the vowel is actually a /ga/ cue according to the 3DDS analysis [23]. Thus, the /da/ sound has a conflicting /ga/ cue [22]. Owing to this, as soon as the /da/ feature is removed, all the subjects report a /ga/ more than 90% of the time at 12 dB and 6 dB SNR, as seen in panel $\boxed{2}$ of Fig. 7.5. With the increase in masking noise, however, the conflicting cue also gets masked. This explains the increase in score for /da/ with an increase in masking noise.

The /de/ cue, much like the /da/ cue, is a high-frequency burst almost in line with the vocalic portion of the sound. The feature region is highlighted by the solid box in panel $\boxed{3}$ of Fig. 7.5. When the /de/ feature is removed, the perceptual error for /de/ goes to 70% at 12 dB SNR as opposed to an error of 0 as shown by control data in panel $\boxed{4}$ of Fig. 7.5. In the absence of the high-frequency cue, the conflicting /ge/ cue becomes dominant. In the case of the vowel /e/, the /ge/ cue is pushed up in frequency. However, in this case, since frequencies above 3 kHz have been removed, some parts of the /ge/ cue are still audible. This explains the low score for /ge/ at 12 dB SNR, even in the absence of the /de/ cue.

7.1.5 /ga/ and /ge/

The 3DDS analysis shows a mid-frequency burst almost in line with the vowel, to be the /ga/ feature. Like /ka/, the position of the /ga/ feature is actually tied to the second formant frequency. When this mid-frequency energy at $\approx$ 2 kHz, highlighted by the solid box in panel $\boxed{1}$ of Fig. 7.6, is removed, the scores for /ga/ go to 0 from 100%, as seen in panel $\boxed{2}$ of Fig. 7.6. The score for /ga/ remains at 0 across all SNRs. This shows that the mid-frequency burst at $\approx$ 2 kHz is the /ga/ feature region. As in the case of /da/, this /ga/ utterance also has a conflicting high-frequency /da/ cue as seen in panel $\boxed{1}$ of Fig. 7.6. Apart from the mid-frequency burst at 2 kHz, a high-frequency burst above 4 kHz is also seen in panel $\boxed{1}$ of Fig. 7.6, which is the /da/ cue. This explains the high score for /da/ in the absence of the /ga/ cue as seen in panel $\boxed{2}$ of Fig. 7.6. At higher levels of masking, subjects also report /ta/ and /ka/ since the voicing information is masked. The presence of conflicting cues for /ga/ and /da/ is consistent across all the utterances used in this study.

The /ge/ feature is very different from the /ga/ feature as the /g/ burst depends on the second formant location. Since the /e/ vowel has a higher second formant, the /ge/ feature is a burst above 2 kHz coinciding with the vowel. It is highlighted by the solid box in panel $\boxed{1}$ of Fig. 7.6. As soon as the /ge/ cue is removed, the errors go to 100% across all SNRs. Subjects had an error of 0 when the /ge/ cue was audible. Moreover, since the /ge/ cue is a high-frequency phenomenon, removing it implies removal of the /de/ cue as well. For this reason, as with the vowel /a/, confusions as a result

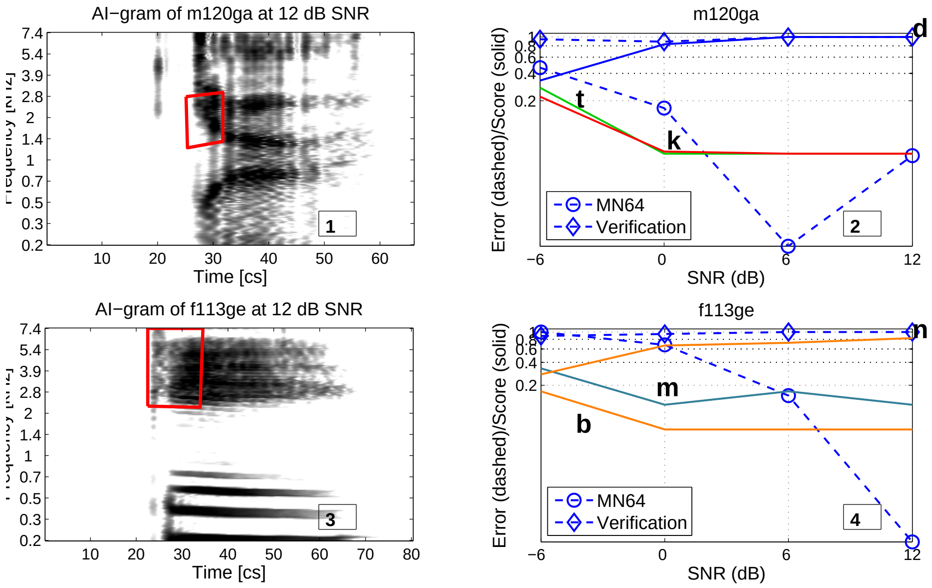


Figure 7.6: Panel $\boxed{1}$: AI-gram of the unmodified m120ga with feature region highlighted in the solid box. Panel $\boxed{2}$: Comparison of errors of the unmodified (dashed line with circle markers) and the modified (dashed line with diamond markers) sounds along with the scores of the confusion patterns (solid lines) for the modified sound. Panel $\boxed{3}$: AI-gram of the unmodified f113ge with feature region highlighted in the solid box. Panel $\boxed{4}$: Comparison of errors of the unmodified (dashed line with circle markers) and the modified (dashed line with diamond markers) sounds along with the scores of the confusion patterns (solid lines) for the modified sound.

of conflicting cues is not seen. For /ge/, a nasal quality is perceived in the absence of its cue. 80% of the subjects report /ne/ for the modified /ge/ at 12 dB SNR. At −6 dB SNR, /me/ and /ne/ are reported with equal probability.

7.1.6 /ba/ and /be/

Using the 3DDS method (Li et al., 2009), the feature region for /ba/ was identified as a burst almost coinciding with the vocalic portion of the sound, as highlighted by the solid box in panel $\boxed{1}$ of Fig. 7.7. As seen in panel $\boxed{2}$ of Fig. 7.7, removing the region highlighted by the solid box in panel $\boxed{1}$ leads to an error of almost 100%, even at 12 dB SNR. When subjects can listen to the feature region, the error is 0, as shown by the dotted line in panel $\boxed{2}$. This error of almost 100% remains consistent across all values of SNRs. Clearly, the low-frequency energy is critical to /ba/ perception. Panel $\boxed{2}$ shows that in the absence of the /ba/ cue, all subjects perceived a /pa/ sound at 12 dB SNR and the score for /pa/ was 70% even at 0 dB SNR. Since /pa/ too is characterized by a low-frequency energy and a wide-band click, this confusion is in agreement with the findings of Li et al. (2009) [23]. In the absence of the voicing information and the burst and in the presence of the wide-band click, a /pa/ sound is heard.

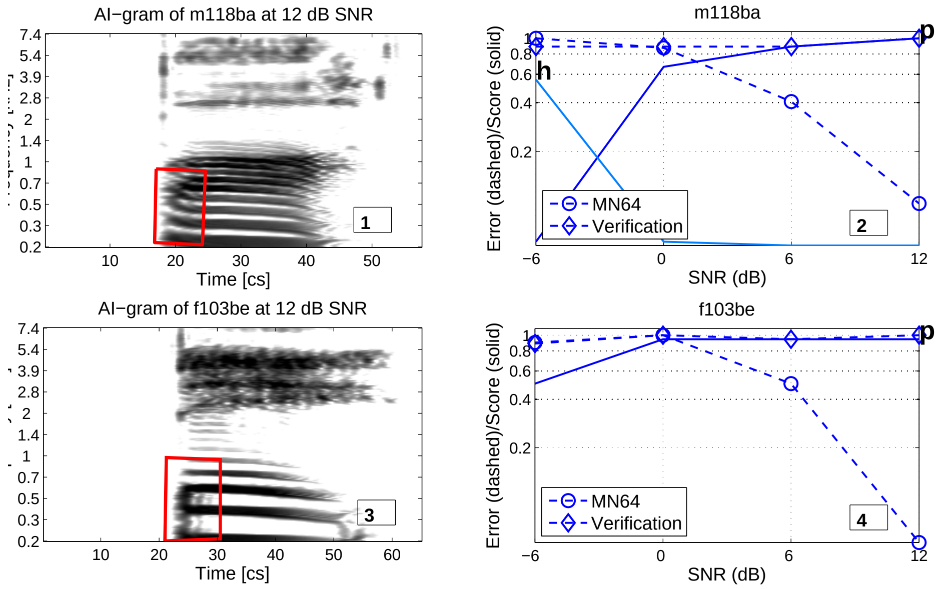


Figure 7.7: Panel $\boxed{1}$: AI-gram of the unmodified m118ba with feature region highlighted in the solid box. Panel $\boxed{2}$: Comparison of errors of the unmodified (dashed line with circle markers) and the modified (dashed line with diamond markers) sounds along with the scores of the confusion patterns (solid lines) for the modified sound. Panel $\boxed{3}$: AI-gram of the unmodified f103be with feature region highlighted in the solid box. Panel $\boxed{4}$: Comparison of errors of the unmodified (dashed line with circle markers) and the modified (dashed line with diamond markers) sounds along with the scores of the confusion patterns (solid lines) for the modified sound.

The region highlighted with the solid box in panel $\boxed{3}$ of Fig. 7.7 is the /be/ feature based on the features for the vowel /a/. This sound has a low-frequency cue that coincides with the vowel. When this region was removed, the errors for /be/ went to 100% (dashed lines with diamond markers in panel $\boxed{4}$ of Fig. 7.7) from 0 (dashed lines with circle markers in panel $\boxed{4}$

of Fig. 7.7) at 12 dB SNR. This error remains constant across all SNRs. Thus, the highlighted region in panel $\boxed{1}$ of Fig. 7.7 certainly is the /be/ cue. Moreover, with no voicing information, all subjects report a /pe/ even at 0 dB SNR. This is similar to /ba/, which also had most subjects reporting /pa/ in the absence of the /ba/ cue.

7.2 Results for fricatives

7.2.1 /sa/ and /se/

According to the results of the 3DDS method, as seen in Fig. 4.5(a), the /sa/ event is triggered by a high-frequency frication edge at around 4 kHz. This frication region is highlighted in the red box in panel $\boxed{1}$ of Fig. 7.8. When this region is removed, the frication edge cue is no longer available and this leads to a drastic difference in perception of /sa/ and most subjects report /θa/. When the sound m112sa was unmodified, subjects had a very low error peaking at −6 dB SNR with a value of 40% (shown by the dashed lines with circle markers in panel $\boxed{2}$ of Fig. 7.8). However, when the feature region is not heard, the error for the sound is consistently 100%, as denoted with the dashed lines with diamond markers in panel $\boxed{2}$ of Fig. 7.8. In the absence of the frication cue, almost 50% of the subjects report hearing a /θa/ and 30% report hearing a /ða/. This trend was also seen in the 3DDS method, with the truncation experiments.

The features for fricatives have very little dependence on the vowel. Owing to this, the /se/ cue highlighted in the solid box in panel $\boxed{3}$ of Fig. 7.8, is very similar to the one highlighted in the solid box in panel $\boxed{1}$ of Fig. 7.8, both in terms of duration and bandwidth. The unmodified sound has very low error, as denoted by the dashed lines with circle markers in panel $\boxed{4}$ of Fig. 7.8. However, when the frication cue is absent the error is 100% across all values of SNR (dashed line with diamond markers in panel $\boxed{4}$ of Fig. 7.8). The sounds reported in the absence of the feature are also consistent across vowels. /θa/ and /ða/ are the principal confusions seen. At higher levels of noise, the subjects have a greater number of confusions with several stop consonants.

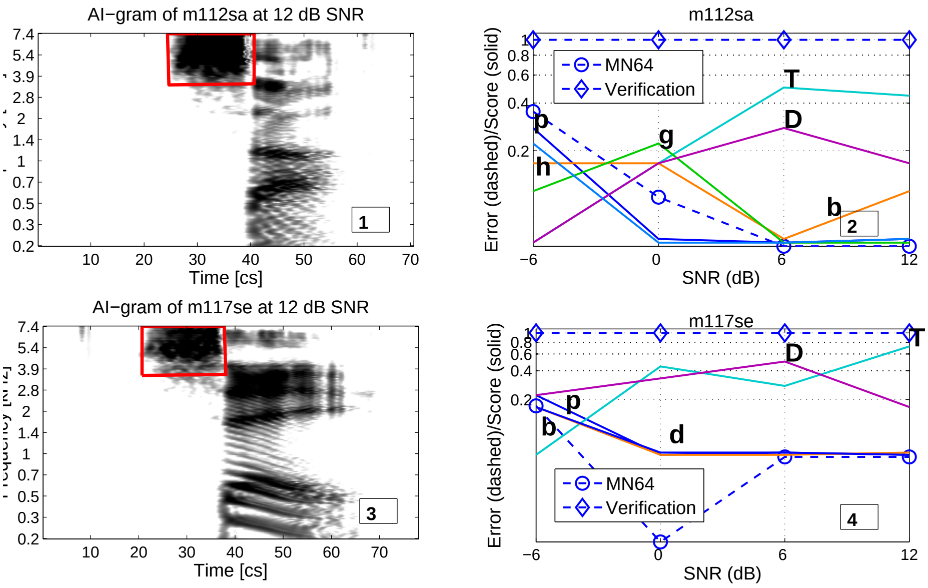


Figure 7.8: Panel $\boxed{1}$: AI-gram of the unmodified m112sa with feature region highlighted in the solid box. Panel $\boxed{2}$: Comparison of errors of the unmodified (dashed line with circle markers) and the modified (dashed line with diamond markers) sounds along with the scores of the confusion patterns (solid lines) for the modified sound. Panel $\boxed{3}$: AI-gram of the unmodified m117se with feature region highlighted in the solid box. Panel $\boxed{4}$: Comparison of errors of the unmodified (dashed line with circle markers) and the modified (dashed line with diamond markers) sounds along with the scores of the confusion patterns (solid lines) for the modified sound. Due to the lack of IPA symbols in Matlab figures, /θa/ and /ða/ have been denoted by T and D, respectively in the figure.

7.2.2 /fa/ and /fe/

According to the 3DDS method as seen in Fig. 4.6(a), the /fa/ cue is in the mid-frequency region. However, /fa/ is a fricative and thus has a noiselike high-frequency frication region as well. This frication region, is not robust as in the case of other fricatives such as /sa/ and /ʃa/. In the presence of white noise (as was the case with the 3DDS method), a large part of this region was also masked. This would explain the low scores that the unmodified /fa/ sounds had in the noise masking experiment for the 3DDS method. In panel $\boxed{1}$ of Fig. 7.9, the red box indicates the frication region for /fa/ spoken by talker f109. When subjects can hear the entire sound, the errors for/fa/ are quite low, remaining below 20% even at 0 dB SNR (dashed lines with circle markers in panel $\boxed{2}$ of Fig. 7.9). However, when the frication region is removed, the perception of the sound alters drastically. Clearly, this frication region is important to /fa/ perception. Most subjects report hearing a /ba/ sound in the absence of the frication region. The burst-like quality that /fa/ has as seen in panel $\boxed{1}$ of Fig. 7.9, gives rise to confusions with /ba/.

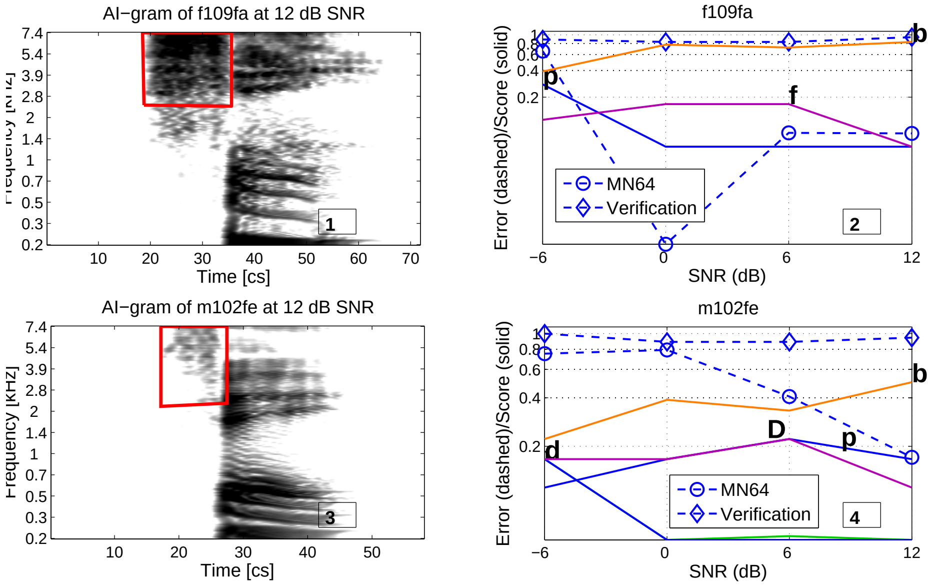


Figure 7.9: Panel $\boxed{1}$: AI-gram of the unmodified f109fa with feature region highlighted in the solid box. Panel $\boxed{2}$: Comparison of errors of the unmodified (dashed line with circle markers) and the modified (dashed line with diamond markers) sounds along with the scores of the confusion patterns (solid lines) for the modified sound. Panel $\boxed{3}$: AI-gram of the unmodified m102fe with feature region highlighted in the solid box. Panel $\boxed{4}$: Comparison of errors of the unmodified (dashed line with circle markers) and the modified (dashed line with diamond markers) sounds along with the scores of the confusion patterns (solid lines) for the modified sound. Due to the lack of IPA symbols in Matlab figures, /ða/ has been denoted by D in the figure.

As in the case of /fa/, panel $\boxed{3}$ of Fig. 7.9 shows the high-frequency frication region for /fe/. It can be seen that the frication region in this case is not as robust as it was for /fa/. For the unmodified sound, the error was significant even at 12 dB SNR (20% as seen in panel $\boxed{4}$ of Fig. 7.9). However, this error goes up to 100% when the frication energy is removed. This shows the contribution of the frication region to /fa/ perception. As in the case of /fa/, panel $\boxed{2}$ of Fig. 7.9 shows that in the absence of the frication region, most subjects report hearing a /be/ sound instead of /fe/.

7.2.3 /ʃa/ and /ʃe/

From the 3DDS method, as seen in Fig. 4.4(a), the high-frequency frication edge located at ≈ 2 kHz is critical to /ʃa/ perception. This is highlighted with the solid box in panel $\boxed{1}$ of Fig. 7.10. However, as seen in panel $\boxed{2}$ of Fig. 7.10, the error for /ʃa/ in the presence and absence of the region highlighted with the solid box is more or less the same. There is a score of ≈ 100% in both the cases. Keeping with the frication-edge theory, it is possible that for /ʃa/, the removal of the region highlighted in the box was not enough to cue the frication edge for the /sa/ sound. Owing to this, the syllable was still perceived as /ʃa/. Further experiments will need to be conducted to verify the frication-edge theory and to find the threshold for the frication edge for the different fricatives.

In the case of /ʃe/, panel $\boxed{4}$ of Fig. 7.10 shows that the error for the unmodified /ʃe/ sound was low, reaching a maximum of 40% only at -6 dB SNR (dashed lines with circle markers). However, in the absence of the feature region, the error was close to 100% across all SNRs (dashed lines with diamond markers in panel $\boxed{4}$ of Fig. 7.10). Since the 2 kHz frication edge is no longer detected and has instead moved to 4 kHz, as seen in panel $\boxed{3}$ of Fig. 7.10, most subjects report a /sa/ sound. This is in perfect agreement with the /sa/ cue as identified by the 3DDS method (Fig. 4.4(a)). Minor confusions with /za/ are also seen at -6 dB SNR, which is probably due to loss of voicing information due to masking noise.

7.2.4 /za/ and /ze/

Because /za/ is the voiced co-relate of /sa/, the /za/ cue too, is a high-frequency frication edge at around 4 kHz (Fig. 4.5(b)). The feature region for/za/ is highlighted with the solid box in panel $\boxed{1}$ of Fig. 7.11. When this feature region is removed, the error for /za/ remains at 100% across all values of SNR, as denoted by the dashed lines with diamond markers in panel $\boxed{2}$ in Fig. 7.11. Just as most subjects reported /θa/ when the /sa/ feature region was removed, /ða/ was reported in the absence of the /za/ feature. This

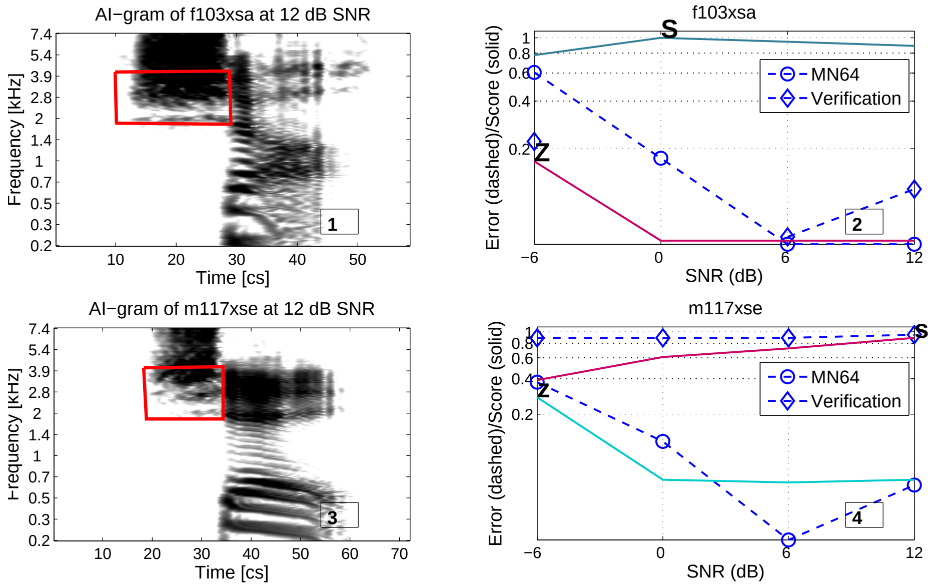


Figure 7.10: Panel $\boxed{1}$: AI-gram of the unmodified f103xsa with feature region highlighted in the solid box. Panel $\boxed{2}$: Comparison of errors of the unmodified (dashed line with circle markers) and the modified (dashed line with diamond markers) sounds along with the scores of the confusion patterns (solid lines) for the modified sound. Panel $\boxed{3}$: AI-gram of the unmodified m117xse with feature region highlighted in the solid box. Panel $\boxed{4}$: Comparison of errors of the unmodified (dashed line with circle markers) and the modified (dashed line with diamond markers) sounds along with the scores of the confusion patterns (solid lines) for the modified sound. Due to the lack of IPA symbols in Matlab figures, /ʃa/ and /ʒa/ have been denoted by S and Z, respectively in the figure.

makes sense since /ða/ is the voiced correlate of /θa/. Minor confusions with stop consonant /ba/ is also seen.

There is practically no change in the frication edge across vowels. panel $\boxed{3}$ of Fig. 7.11 shows the feature region for /ze/ highlighted in the solid box. When subjects heard the unmodified sound, there was no error at 6 and 12 dB SNR (dashed lines with circle markers in panel $\boxed{4}$ of Fig. 7.11). When the subjects hear the modified sound, with the feature region removed, the error for /ze/ went to 100% across all SNRs. As in the case of /za/, most subjects reported /ða/, in the absence of the feature region.

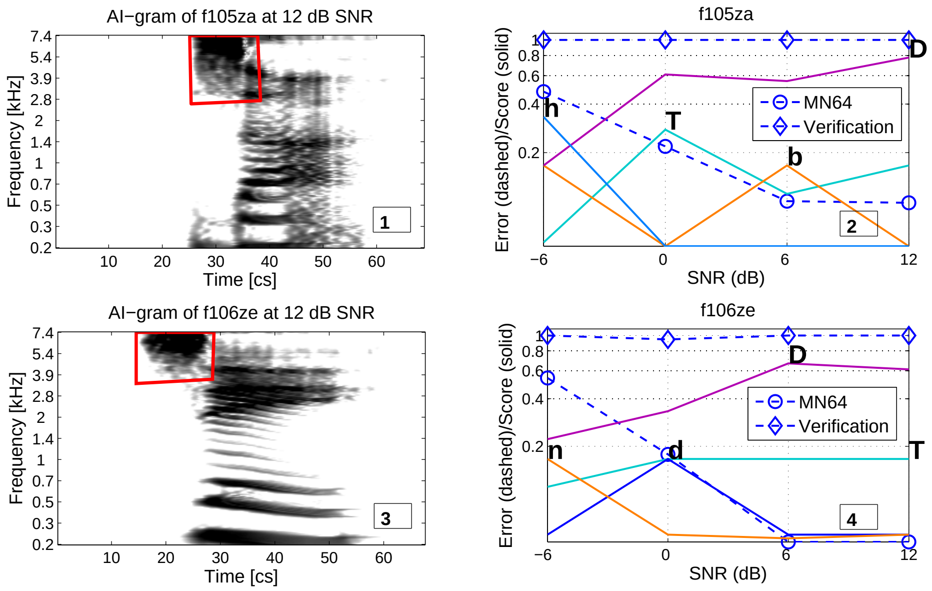


Figure 7.11: Panel $\boxed{1}$: AI-gram of the unmodified f105za with feature region highlighted in the solid box. Panel $\boxed{2}$: Comparison of errors of the unmodified (dashed line with circle markers) and the modified (dashed line with diamond markers) sounds along with the scores of the confusion patterns (solid lines) for the modified sound. Panel $\boxed{3}$: AI-gram of the unmodified f106ze with feature region highlighted in the solid box. Panel $\boxed{4}$: Comparison of errors of the unmodified (dashed line with circle markers) and the modified (dashed line with diamond markers) sounds along with the scores of the confusion patterns (solid lines) for the modified sound. Due to the lack of IPA symbols in Matlab figures, /θa/ and /ða/ have been denoted by T and D, respectively in the figure.

7.2.5 /va/ and /ve/

As in the case of /fa/, very few /va/ sounds used in the 3DDS study had very high scores in the noise masking experiment. The fact that they were masked by white noise that masked the weak frication energy that /va/ has could be a reason for this. According to the 3DDS method, the mid-frequency energy highlighted in the solid box in panel $\boxed{1}$ of Fig. 7.12 is the /va/ feature. Apart from that, since /va/ is essentially a fricative, the frication region is also highlighted. The errors when the sound was unmodified is shown with the dashed lines with the circle markers in panel $\boxed{2}$ of Fig. 7.12. In the absence of the features, the error for /va/ goes up to 80%. In the absence of the /va/ cue, most subjects report hearing a /ba/. There seems to be a

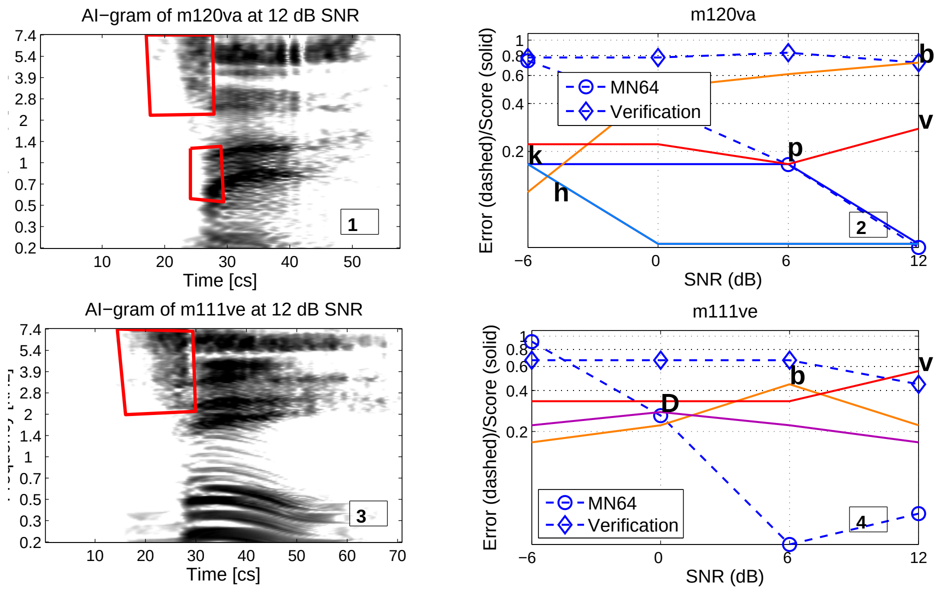


Figure 7.12: Panel $\boxed{1}$: AI-gram of the unmodified m120va with feature region highlighted in the solid box. Panel $\boxed{2}$: Comparison of errors of the unmodified (dashed line with circle markers) and the modified (dashed line with diamond markers) sounds along with the scores of the confusion patterns (solid lines) for the modified sound. Panel $\boxed{3}$: AI-gram of the unmodified m111ve with feature region highlighted in the solid box. Panel $\boxed{4}$: Comparison of errors of the unmodified (dashed line with circle markers) and the modified (dashed line with diamond markers) sounds along with the scores of the confusion patterns (solid lines) for the modified sound. Due to the lack of IPA symbols in Matlab figures, /ða/ has been denoted by D in the figure.

close link between the sounds /ba/, /va/, /fa/ as they seem to confused often within the set. This can be explained by the fact that /ba/ , being is stop consonant, is confused with /va/ and /fa/ as both of them have a burst-like energy below 1 kHz.

In this particular case of /ve/, since the frication region was somewhat more intense, that is highlighted with the solid box in panel $\boxed{2}$ of Fig. 7.12 as the feature region. Most /v/ tokens in the data base are not robust as can be seen in panel $\boxed{4}$ of Fig. 7.12. The change in error in this case (panel $\boxed{4}$ of Fig. 7.12) is not as drastic as has been the case in the other CVs tested. In the absence of the /ve/ cue, most subjects reported hearing /be/ or /ðe/ (panel $\boxed{4}$ of Fig. 7.12). For most of the fricatives, in the complete absence of the frication energy, a /θ/ or /ð/ sound is reported and the same trend is seen here.

7.2.6 /ʒa/ and /ʒe/

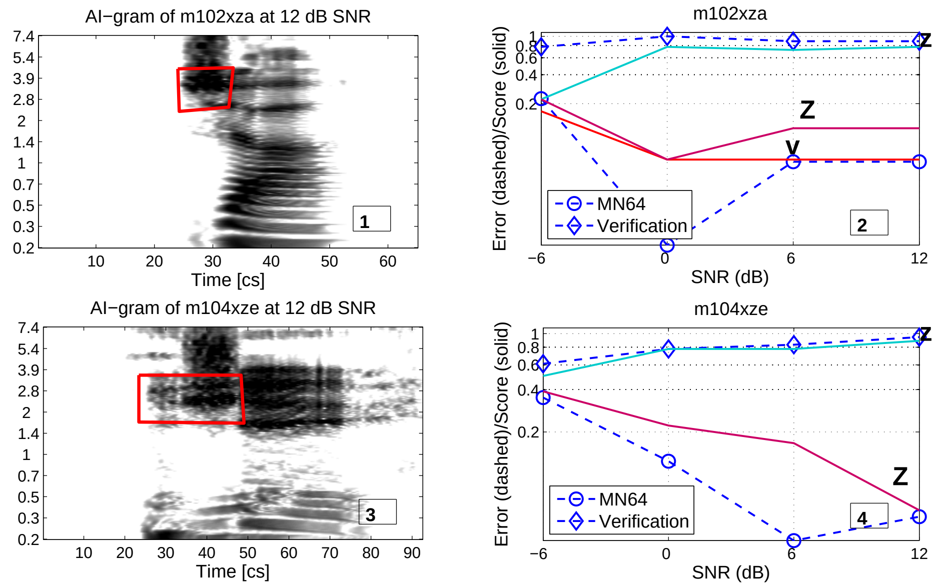


Figure 7.13: Panel $\boxed{1}$: AI-gram of the unmodified m102xza with feature region highlighted in the solid box. Panel $\boxed{2}$: Comparison of errors of the unmodified (dashed line with circle markers) and the modified (dashed line with diamond markers) sounds along with the scores of the confusion patterns (solid lines) for the modified sound. Panel $\boxed{3}$: AI-gram of the unmodified m104xze with feature region highlighted in the solid box. Panel $\boxed{4}$: Comparison of errors of the unmodified (dashed line with circle markers) and the modified (dashed line with diamond markers) sounds along with the scores of the confusion patterns (solid lines) for the modified sound. Due to the lack of IPA symbols in Matlab figures, /ʒa/ has been denoted by Z in the figure.

According to the 3DDS method, the /ʒa/ cue, much like the /ʃa/ cue, is a frication edge at around 2 kHz. The /ʒa/ cue is shorter in duration compared to /ʃa/ and is highlighted in the solid box in panel $\boxed{1}$ of Fig. 7.13. When this feature is heard, the error for /ʒa/ remains below 20% across all SNRs. However, in the absence of the cue, the error goes to 100%. Clearly, the frication edge needs to be heard to perceive /ʒa/ correctly. By removing the feature, the frication edge heard is at ≈ 4 kHz and this leads to perception of /za/. This is in agreement with the 3DDS results for the /za/ cue. A small percentage of the subjects do continue to report /ʒa/ but the numbers are too low to be significant.

The /ʒe/ cue, as highlighted in the solid box in panel $\boxed{3}$ of Fig. 7.13, is the same as that for /ʒa/. In the presence of the feature region, subjects had practically no error as shown by dashed lines with circle markers in panel $\boxed{4}$ of Fig. 7.13. However, when the feature region was removed, the errors rise to over 80% at 12 dB SNR and remains high consistently. As in the case of /ʒa/, on removal of the /ʒe/ cue, most subjects report a /ze/, since the frication edge at $\approx$ 4 kHz is now heard, which is the cue for /ze/.

CHAPTER 8

CONFUSIONS STUDY

As shown in the previous chapters, the effect of removing the feature region is consistent across all the sounds used in the verification study. The critical band of the features determines the scores for a particular sound as is seen with the strong co-relation between SNR_{90} and $\mathrm{SNR}_{\mathrm{e}}$. The results discussed in the previous chapter show support for the theory of acoustic invariance in speech cues [7]. Apart from this, it is also necessary to study the confusions with other sounds in the absence of the feature region. If the confusions were randomly spread across the response set, that would not lead to any insights about features. However, if there emerged a pattern in the confusions between different sounds, it would allow us to predict the behavior of these sounds in the presence of noise.

Significant information can be derived from the knowledge of what sounds can be heard in the absence of a feature for the modified sounds. This is especially exciting since it has been possible to reasonably predict confusions. This definitely accomplishes the goal of this study: to verify the feature regions of sounds. To accurately predict confusions, it is important that the sound be audible. For this reason, an analysis of the highest confusions for each CV, across all the utterances and all the listeners has been done at +12 dB SNR. Figures 8.1 and 8.2 show bar plots of the two highest confusions for a given CV stop consonant pair that have been modified so as to remove the feature region, for the vowel /a/ and /e/ while Fig. 8.3 and Fig. 8.4 do the same for the fricative CVs.

In the case of the stop consonants, the classic /p/, /t/, /k/ confusion group is seen which is in agreement with Miller and Nicely (1955) [26]. The reason for this is also quite clear due to the presence of conflicting cues. As observed by Li and Allen (2009) [22], each sound has latent cues of other sounds which are called conflicting cues. Each of /p,t,k/ actually has the conflicting cues of the other two sounds which leads to this confusion group.

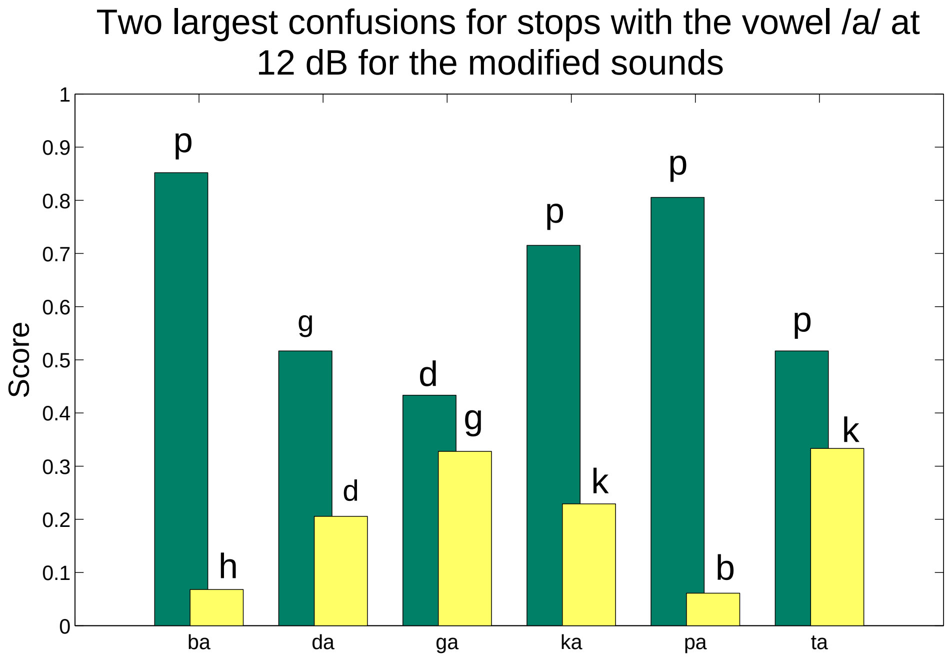


Figure 8.1: Bar plots of the two highest confusions for each stop consonant CV for the vowel /a/.

When the dominant feature is masked, the latent cues become audible, as for the case for /g,d/. Each of them contains the conflicting cue of the other, forming a confusion group. This phenomenon has been discussed in Chapter 7. Another common confusion was that of /b/ and /p/. Since the /ba/ feature is in the low frequencies, removing it gets rid of the voicing information too. This is what leads to the /b,p/ confusions. Although the confusions discussed remain more or less consistent across sounds, /g/ is one sound that had different confusions in the context of different vowels. With the vowel /e/, a nasal quality was observed with the /ge/ sound, owing to which the largest confusion is not /d/ as in the case of the vowel /a/, but /ne/. Minor confusions with /me/ are also seen in the case of the sound /pe/. The difference in the duration of the two vowels may be a factor in this dissimilar confusion.

Fricatives, /f/, /v/, /b/ also form a confusion group, more so because modified /v/ and /f/ sounds lead to a confusion with /b/. This is owing to the burst nature of /b/ which both /v/ and /f/ inherently share. Consonants /ʃ/ and /ʒ/ have the conflicting /s/ or /z/ cue, which can be obtained by high-pass filtering the sound. Thus, on removal of their feature region, subjects

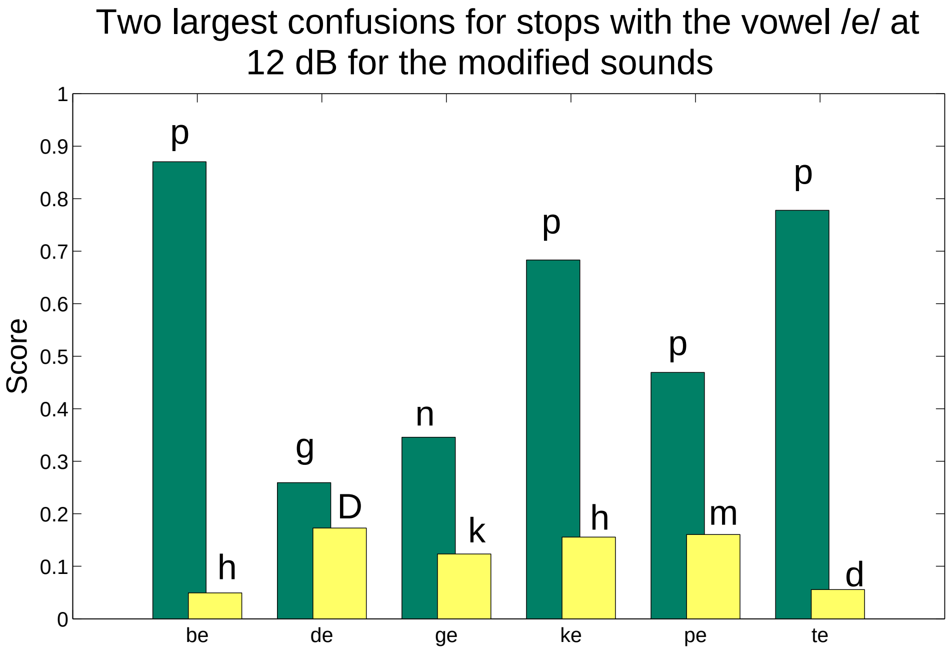


Figure 8.2: Bar plots of the two highest confusions for each stop consonant CV for the vowel /e/.

report /s/ or /z/ respectively. In the case of most fricatives, including /s/ and /z/, the complete removal of the frication energy consistently seem to trigger a /θ/ or /ð/ response.

It is interesting to note that this behavior is robust across noise and across vowels.

Figures 8.2 and 8.4 are similar to their counterparts with the vowel /a/ (Fig. 8.1 and 8.3). The major difference seen in Fig. 8.2, with reference to Fig. 8.1, is that in the confusions with vowel /e/, a nasal quality is seen leading to confusions with /me/ and /ne/.

With the fricatives, the feature removal of /ve/ for the verification experiment seems to have been insufficient. In spite of the feature removal, half the subjects continue to report /ve/. Around 20% of the subjects also report /ð/, keeping with the fricative characteristic to sound like a /θ/ or /ð/, in the absence of the frication region. For /ʃe/, the feature removal seems to have worked well, with most subjects reporting /s/. This is in keeping with the /sa/ feature seen in Fig. 4.5(a).

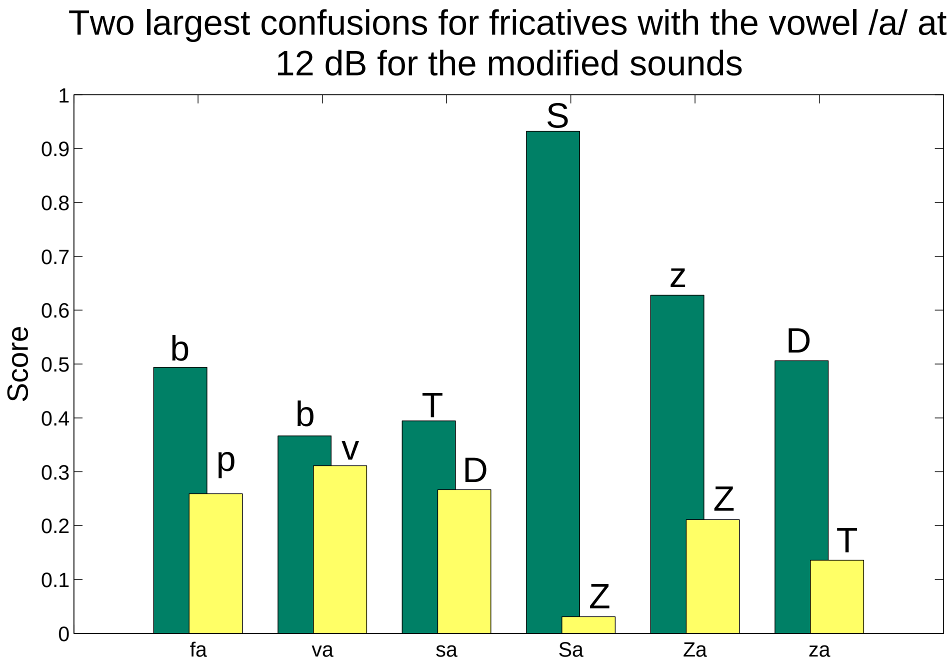


Figure 8.3: Bar plots of the two highest confusions for each fricative CV for the vowel /a/.

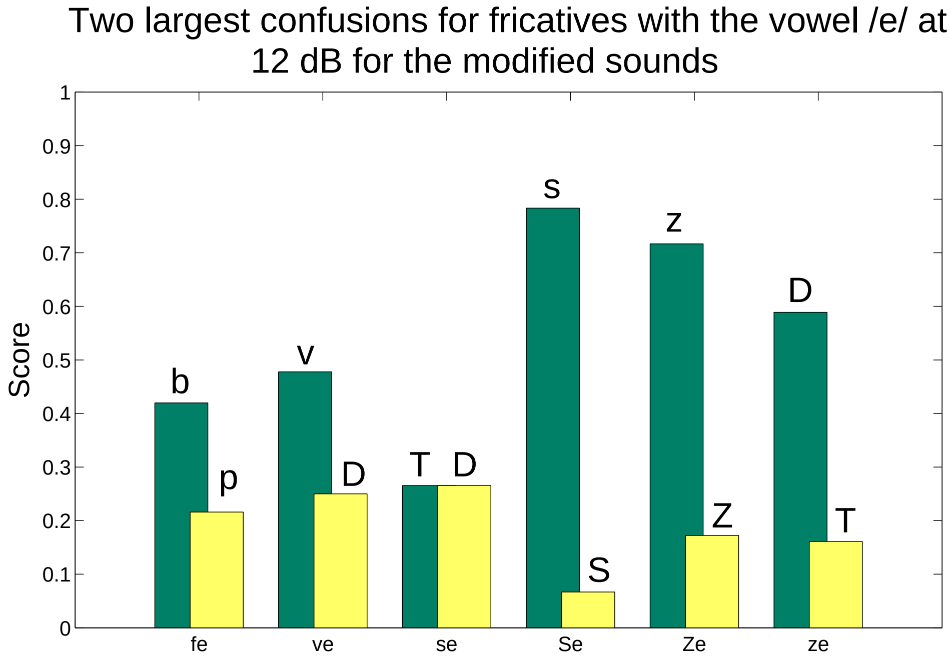


Figure 8.4: Bar plots of the two highest confusions for each fricative CV for the vowel /e/.

CHAPTER 9

CONCLUSION

The 3DDS method uses three independent methods to find the feature region that are perceptually critical for a given sound. Given these feature regions, the verification study confirms that they are critical to perception.

For the vowel /a/, /p/ and /b/ are characterized by a low frequency burst and a wide band click that contributes to the quality of the sound. Consonants /k/ and /g/ are defined by a mid frequency burst and /t/ and /d/ have a high frequency burst. The unvoiced stops preceed the vowel by $\approx$ 6-10 [cs], while the voiced stops have onsets that almost coincide with the vowel. For the fricatives, /s/ and /z/ are characterized by a frication edge at 4 [kHz] while /ʃ/ and /ʒ/ have the same at 2 [kHz]. /f/ and /v/ are relatively harder to characterize owing to the low quality of most tokens present in our database. However, both have a weak frication region that is easily masked. In cases when the frication energy is weak, the mid frequency burst also cues the sound.

The verification study shows that for most of the stops and fricatives, the feature regions are more or less invariant across the vowels /a/ and /e/. The biggest change seen is for /ke/ and /ge/ whose burst location is closely corelated to the location of F_2. Since the second formant for /e/ is higher in frequency than that for /a/, the /ke/ and /ge/ bursts have a higher burst frequency.

The future scope of this study would be to expand it to other vowels. Using VCs and CVCs would also be useful to find whether the feature regions remain the same across initial or final position in the syllable.

REFERENCES

[1] J. B. Allen. Short time spectral analysis, synthesis, and modification by discrete Fourier transform. *IEEE Trans. Acoust. Speech and Sig. Processing*, 25:235–238, 1977.

[2] F. Li. *Perceptual cues of consonant sounds and impact of sensorineural hearing loss on speech perception.* Ph.D. dissertation in Electrical Engineering and Computer Engineering, University of Illinois at Urbana-Champaign, August 2009.

[3] J. B. Allen. Harvey Fletcher's role in the creation of communication acoustics. *J. Acoust. Soc. Am.*, 99(4):1825–1839, April 1996.

[4] J. B. Allen and L. R. Rabiner. A unified approach to short-time Fourier analysis and synthesis. *Proc. IEEE*, 65(11):1558–1564, November 1977.

[5] J. B. Allen. Consonant recognition and the articulation index. *J. Acoust. Soc. Am.*, 117(4):2212–2223, April 2005.

[6] A. A.H. Alwan. *Modeling Speech Perception in Noise: The Stop Consonants as a Case Study.* Ph.D. dissertation in Electrical Engineering and Computer Science, Massachusetts Institute of Technology, February 1992.

[7] S. E. Blumstein and K. N. Stevens. Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants. *J. Acoust. Soc. Am.*, 66(4):1001–1017, October 1979.

[8] S. E. Blumstein and K. N. Stevens. Perceptual invariance and onset spectra for stop consonants in different vowel environments. *J. Acoust. Soc. Am.*, 67(2):648–266, February 1980.

[9] S. E. Blumstein, K. N. Stevens, and G. N. Nigro. Property detectors for bursts and transitions in speech perceptions. *J. Acoust. Soc. Am.*, 61(5):1301–1313, May 1977.

[10] R. A. Cole and B. Scott. Toward a theory of speech perception. *Psychological Review*, 81(4):348–374, 1974.

[11] F.S. Cooper, P.C. Delattre, A.M. Liberman, J.M. Borst, and L.J. Gerstman. Some experiments on the perception of synthetic speech sounds. *J. Acoust. Soc. Am.*, 24(6):597–606, November 1952.

[12] P.C. Delattre, A.M. Liberman, and F.S. Cooper. Acoustic loci and translational cues for consonants. *J. Acoust. Soc. Am.*, 24(4):769–773, July 1955.

[13] R. Drullman, J. M. Festen, and R. Plomp. Effect of temporal envelope smearing on speech reception. *J. Acoust. Soc. Am.*, 95(2):1053–1064, February 1994.

[14] H. Fletcher and R. Galt. The perception of speech and its relation to telephony. *J. Acoust. Soc. Am.*, 22:89–151, 1950.

[15] N. R. French and J. C. Steinberg. Factors governing the intelligibility of speech sounds. *J. Acoust. Soc. Am.*, 19:90–119, 1947.

[16] S. Furui. On the role of spectral transition for speech perception. *J. Acoust. Soc. Am.*, 80(4):1016–1025, October 1986.

[17] M. Hasegawa-Johnson. Time-frequency distribution of partial phonetic information measured using mutual information. *ICSLP-2000*, 4:133–136, October 2000.

[18] V. Hazan and A. Simpson. The eff?ect of cue-enhancement on the intelligibility of nonsense word and sentence materials presented in noise. *Speech Communication*, 24:211–226, 1998.

[19] J. M. Heinz and K. N Stevens. On the properties of voiceless fricative consonants. *J. Acoust. Soc. Am.*, 33 (5):589–596, May 1961.

[20] G. W Hughes and M. Halle. Spectral properties of fricative consonants. *J. Acoust. Soc. Am.*, 28(2):303–310, Mar. 1956.

[21] D. Klatt. Acoustical theory terminal analog of speech synthesis. *Proceedings of the 1972 International Conference on Speech Communication and Processing*, 1972.

[22] F. Li and J. B. Allen. Manipulation of consonants in natural speech. *IEEE Trans. Audio, Speech and Language Processing*, 2010, to be published.

[23] F. Li, A. Menon, and J. B Allen. A psychoacoustic methodology to study perceptual cues of stop consonants in natural speech. *J. Acoust. Soc. Am.*, 127(4):2599–2610, Apr 2010.

[24] A.M. Liberman, F.S. Cooper, D.P. Shankweiler, and M. Studdert-Kennedy. Perception of the speech code. *Psychol. Review*, 74(6):431–61, November 1967.

[25] A. Malećot. Computer-assisted phonetic analysis techniques for large recorded corpuses of natural speech. *J. Acoust. Soc. Am.*,53 (1):356–356, Jan 1973.

[26] G. A. Miller and P. E. Nicely. An analysis of perceptual confusions among some English consonants. *J. Acoust. Soc. Am.*, 27:338–352, 1955.

[27] S. Phatak and J. B Allen. Consonant and vowel confusions in speech-weighted noise. *J. Acoust. Soc. Am.*, 121(4):2312–2326, 2007.

[28] R. K. Potter, G. A. Kopp, and H. G. Kopp. *Visible Speech.* Dover Publications, Inc, New York, 1966.

[29] D. Recasens. Place cues for nasal consonants with special reference to catalan. *J. Acoust. Soc. Am.*,73 (4):1346–1353, Apr 1983.

[30] M. S. Regnier and J. B. Allen. A method to identify noise-robust perceptual features: Application for consonant /t/. *J. Acoust. Soc. Am.*, 123(5):2801–2814, May 2008.

[31] R.E. Remez, P.E. Rubin, D.B. Pisoni, and T.D. Carrell. Speech perception without traditional speech cues. *Science*, 212(4497):947–949, May 1981.

[32] K. S. Rhebergen and N. J Versfeld. A speech intelligibility index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners. *J. Acoust. Soc. Am.*, 117 (4):2181–2192, Apr 2005.

[33] C. E. Shannon. The mathematical theory of communication. *Bell System Tech. J.*, 27:379–423 (parts I, II), 623–656 (part III), 1948.

[34] R. V. Shannon, F. G. Zeng, V. Kamath, J. Wygonski, and M. Ekelid. Speech recognition with primarily temporal cues. *Science*, 270:303–304, 1995.

[35] K. N. Stevens and S. E. Blumstein. Invariant cues for place of articulation in stop consonants. *J. Acoust. Soc. Am.*, 64(5):1358–1369, November 1978.

[36] K. N. Stevens, S. E. Blumstein, L. Glicksman, M. Burton, and K. Kurowski. Acoustic and perceptual characteristics of voicing in fricatives and fricative clusters. *J. Acoust. Soc. Am.*, 91(5):2979–3000, May 1992.

[37] D. J. Van Tassel, S. D. Soli, V. M. Kirby, and G. P. Widin. Speech waveform envelope cues for consonant recognition. *J. Acoust. Soc. Am.*, 82 (4):1152–1161, Oct 1987.

[38] M. D. Wang and R. C. Bilger. Consonant confusions in noise: A study of perceptual features. *J. Acoust. Soc. Am.*, 54:1248–1266, 1973.