Aston University

Some pages of this thesis may have been removed for copyright restrictions.

If you have discovered material in AURA which is unlawful e.g. breaches copyright, (either yours or that of a third party) or any other law, including but not limited to those relating to patent, trademark, confidentiality, data protection, obscenity, defamation, libel, then please read our Takedown Policy and contact the service immediately

The Verbal Transformation Effect: An exploration of the perceptual organization of speech

Marcin Stachurski

Doctor of Philosophy

ASTON UNIVERSITY

December 2012

This copy of the thesis has been supplied on condition that anyone who consults it is understood to recognise that its copyright rests with its author and that no quotation from the thesis and no information derived from it may be published without proper acknowledgement.

Aston University

The Verbal Transformation Effect: An exploration of the perceptual organization of speech

Six experiments investigated the influence of several grouping cues within the framework of the Verbal Transformation Effect (VTE, Experiments 1 to 4) and Phonemic Transformation Effect (PTE, Experiments 5 and 6), where listening to a repeated word (VTE) or sequence of vowels (PTE) produces verbal transformations (VTs). In Experiment 1, the influence of F0 frequency and lateralization cues (ITDs) was investigated in terms of the pattern of VTs. As the lateralization difference increased between two repeating sequences, the number of forms was significantly reduced with the fewest forms reported in the dichotic condition. Experiment 2 explored whether or not propensity to report more VTs on high pitch was due to the task demands of monitoring two sequences at once. The number of VTs reported was higher when listeners were asked to attend to one sequence only, suggesting smaller attentional constraints on the task requirements. In Experiment 3, consonant-vowel transitions were edited out from two sets of six stimuli words with 'strong' and 'weak' formant transitions, respectively. Listeners reported more forms in the spliced-out than in the unedited case for the strong-transition words, but not for those with weak transitions. A similar trend was observed for the F0 contour manipulation used in Experiment 4 where listeners reported more VTs and forms for words following a discontinuous F0 contour. In Experiments 5 and 6, the role of F0 frequency and ITD cues was investigated further using a related phenomenon – the PTE. Although these manipulations had relatively little effect on the number of VTs and forms reported, they did influence the particular forms heard. In summary, the current experiments confirmed that it is possible to successfully investigate auditory grouping cues within the VTE framework and that, in agreement with recent studies, the results can be attributed to the perceptual re-grouping of speech sounds.

Marcin Stachurski

Doctor of Philosophy

December 2012

Key Phrases: ***Auditory perception, Scene analysis, Perceptual grouping, Pitch contour, Formant Transitions***

Acknowledgments

I would like to thank my supervisor Brian Roberts for his continued support, advice, encouragement, endless patience and incredible wealth of knowledge. I am also grateful to Rob Summers for his assistance with programming experiments and his continued friendship over the years.

This research was supported by Research Grant EP/F016484/1 from the Engineering and Physical Sciences Research Council, awarded to Brian Roberts and Peter Bailey.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Auditory Perception

Sounds in our environment originate from a variety of acoustic sources; these include people talking, cars passing by, music playing, or leaves rustling on a tree. They are rarely heard in isolation; situations in which only a single source of sound is active are very uncommon. When we engage in or listen to a conversation, more likely than not there will be other sound sources competing for our attention. Yet, despite this mixture, our auditory system is capable of separating out the sounds that come from different events in the environment and grouping together sound streams originating from the same source.

Bregman and Pinker (1978) defined a stream as “a psychological organisation whose function is to represent mentally the acoustic activity of a single source over time” (p. 19). We are bombarded with a constant stream of sensory information (and not just in the auditory domain) that is coming from different objects and events. In these mixtures of sensory evidence, whenever there is more than one object or event present at the same time the minimum condition for being able to identify it is to correctly detect which parts of the stimulation belong to the same object. Understanding a speaker with extraneous sound sources present (e.g. at a party) rather than one-on-one in a quiet room is made more difficult, however, in most circumstances it is still achieved with seemingly little difficulty. Although the nature and the details of the situation may vary – there can be other people talking, a plane flying by or a fire alarm going off - without the identification of the particular parts of the sensory stimulation we want to attend, the process of building a representation of it will fail.

Such failure can result in two outcomes. The first is a failure to group each subset of sounds separated in time but arising from a common source into a separate auditory stream, a failure of separation resulting in one aggregate percept which is not differentiated in any way into figure and background. The second type of failure is a misallocation of properties of streams to the wrong events, where single streams are segregated but they are inappropriately grouped. For example, you might be at a busy party and want to locate and identify your friend by their voice. If you fail to segregate all the voices from each other you will simply hear a morass of noise with all the speech components overlapping with each other. If on the other hand, you do separate the voices into their frequency components but allocate the wrong set of them to your friend’s voice, the perceived timbre of their voice might change and they will sound like a different person.

From an acoustical point of view, a single source of sound (defined as a sequence of acoustic events emanating from one place; Beauvois and Meddis, 1991) like your friend talking, usually has many frequency components. Given that a typical everyday listening situation consists of many such sources, what reaches the listener’s ear is a total sum of their spectra. The auditory system needs to partition this information and correctly allocate a given subset of these components to its respective source, e.g. the human voice. This process where our auditory sensory data are grouped and segregated into separate mental representations, called auditory streams, has been termed *auditory scene analysis* (ASA) by Albert S. Bregman (1990).

Most of the research on this process of perceptually allocating sound elements to their respective sources comes from experiments done with simple stimuli. In these studies, listeners are typically presented with repeating sequences of simple tones, often pure tones or steady-state complex tones. Relatively little has been done with complex broadband dynamic sounds such as speech and this will be addressed in the following thesis. Speech as an acoustic signal consists of elements with many different intensities, different durations, different fundamental frequencies (F0), and different spectral components. However, the relative contribution of these components to grouping is still relatively poorly understood; certainly they are not equally important for the intelligibility of speech (Darwin, 2008). Apart from the theoretical interest of the scientific community, this problem is of paramount importance to the study of computer modelling of speech recognition systems and clinical aspects of hearing loss and cochlear implant users.

The following chapter will review relevant studies that have used relatively simple auditory stimuli with respect to auditory scene analysis, making a distinction between two major types of grouping: *simultaneous* and *sequential* grouping. It will then continue by considering experiments using more complex stimuli within the ASA paradigm and review the literature on the *verbal transformation effect* and its potential for the proposed series of experiments presented in this thesis.

1.2 Auditory Scene Analysis

In the 1930s, a group of Gestalt psychologists proposed a series of principles pertaining to how our visual perception of the world is organised. Their system of rules – including the principles of *similarity*, *good continuation* and *common fate* – described how components of the sensory data may be grouped into perceptual wholes (Koffka, 1935). ASA adapts and uses these rules to explain our auditory experiences based on the idea that events in our environment tend to have some persistence and do not change abruptly (Bregman, 1990). Therefore, in any acoustical mixture, any two sounds originating from the same source are more likely to be grouped together if they strongly resemble one another – the principle of *similarity*, if they change gradually and smoothly – the principle of *good continuation*, or if they begin and end at the same time or vary together coherently – the principle of *common fate*. The aforementioned set of principles, also referred to as *primitive cues,* operate at an early stage of central auditory processing and are considered to be based on automatic, innate processes. Support for this idea comes from the phenomenon of camouflage and the demonstration of perceptual organisation in young infants (e.g. Demany, 1982). Camouflage tricks the observer into grouping parts of the object with parts of the background (inappropriate grouping of parts, as mentioned earlier in the second type of failure). For example, tigers have stripes which tend to break up their contour, and parts of their image merge with woods and grassland making them more difficult to spot. The fact that the perceptual system can be tricked into making inappropriate groupings strongly suggests that there must be a set of basic principles which are difficult to override and that are 'built-in' to our perceptual system. In another line of support, it has been possible to demonstrate perceptual organisation in young children, which at the age of around 2-3 months old, is more likely to reflect innate properties rather than learnt behaviour. Based on infants gaze, Demany (1982) used the habituation-dishabituation technique where infants would be drawn to novel sounds (dishabituation) or lose interest if they had heard them before repeatedly (habituation). When a four tone sequence (two on a high fundamental frequency - H1, H2 and two on a low fundamental frequency - L1, L2) H1-L1-H2-L2-H1-L1 etc. was played in reverse – L2-H2-L1-H1-L2-H2 etc. – it was easily discriminable from the first sequence as the order of the elements changed. If the high and the low tones were sufficiently separated in frequency, they broke into separate streams – high (H1-H2) and low (L1-L2) resulting in each of the two sequences sounding the same after reversal. However, Demany found that if the frequency separation between the low and the high notes was small, reversing the sequence order dishabituated children’s interests in the sequence making it novel again. The dependency of dishabituation, following sequence reversal, on the HL frequency separation implies that greater separations lead to stream segregation even in young infants.

We can, however, also utilise a set of perceptual mechanisms based on voluntary processing. These are thought to operate through our past knowledge and experience and Bregman (1990) described them as a set of schema driven processes. They allow us to take advantage of the properties of sounds that have a reasonably high probability of originating from a common source and may be used to aid in the interpretation of the potentially insufficient or inaccurate organisation offered by primitive processes. One example of schema based knowledge being applied to an auditory stream that might otherwise be heard as a sequence of discrete sounds is *sine-wave speech* (Bailey, Summerfield & Dorman, 1977; Remez, Rubin, Pisoni and Carrell, 1981). It is a digital synthesis technique whereby natural speech is described using a small number of time-varying sinusoids (anecdotally referred to as an “acoustic cartoon” of normal speech). Although the auditory grouping cues are minimal, listeners still report hearing it as speech. As the innate, primitive cues cannot be utilized in this instance, listeners most likely are using their prior knowledge of speech to identify the signal.

For both sets of mechanisms, the “bottom up” primitive cues and the “top down” schemas, our perceptual system is faced with the problems of ASA needing not only to find the right solution but to find it in a very short period of time (almost instantaneously). It does so with the notion of heuristics or “betting” principles, which also demonstrate how various grouping cues are in constant competition with each other. A variety of problems that we face in real world situations do not have a formal solution or a correct one. Hence, we require a set of principles which help us to make a judgment or a decision. A heuristic is a “rule of thumb” or “betting” principle that helps us to find a solution to a problem. Although the proposed outcome might not always be right, on average it is likely to provide a good solution to the problem. It is a set of principles competing with one another and whichever set of these principles dominates determines the solution that will be chosen. In most real world environments there is a plethora of cues to choose from and we usually effortlessly end up with clear and stable perceptions. In a laboratory environment these factors can be deliberately removed or one factor can be pitted against another to produce examples where perception is shifting or ambiguous. This approach of applying a whole set of principles based on the way our view of the world is structured and having those principles competing with each other, is a simple but very powerful technique for dealing with huge amounts of complex information very quickly. It is not guaranteed to give the right answer (if it does not, we may experience illusions), however it is usually very effective.

Both the primitive and schema based sets of principles contribute to the ultimate goal of ASA which is to separate out the sounds that come from different events at the same time, also referred to as *simultaneous grouping*, and to group together sound streams originating from the same source over time - *sequential grouping*. The two grouping processes are distinct but not mutually exclusive. A good illustration of how the two interact with each other was given by Bregman and Pinker (1978). In a repeated sequence of two tones (see Figure 1.1), a pure tone A is followed by a complex tone with two pure tone components, B and C. Bregman and Pinker (1978) manipulated two factors: the frequency of A and the relative timing of B and C and they showed that it is possible to hear the repeated sequence in two ways. One would be a pure tone A alternating with the complex tone BC, while the other way would be a single alternating stream of A and B tones, separate from tone C. The two options show the contrasting nature of the two grouping principles mentioned earlier - *simultaneous* grouping of B and C in the first case and *sequential* grouping of A and B in the second case.

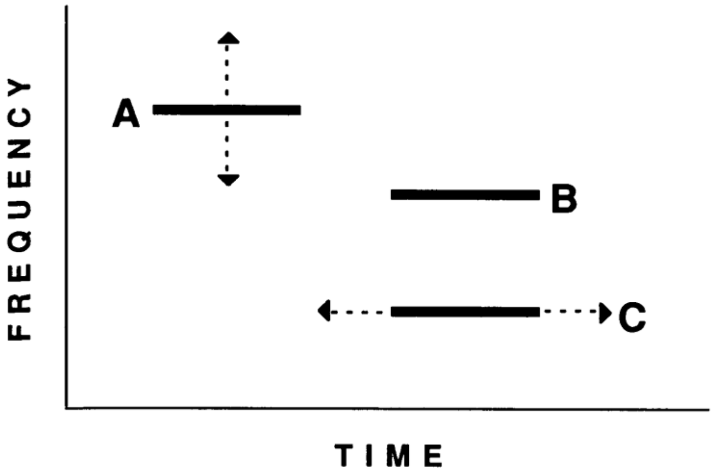


Figure 1.1 Stimulus used by Bregman and Pinker (1978)

1.3 Simultaneous and sequential auditory grouping

The primitive grouping mechanisms are claimed to operate in accordance with Gestalt perceptual mechanisms. In general, sequential grouping involves connecting spectral components that follow one another in time (i.e. tracking a source across time), whereas simultaneous grouping is used to partition concurrent sounds (i.e. overlapping in time) into different streams (Bregman, 1990). Whereas simultaneous grouping is mostly governed by the principle of common fate and by harmonic relations, sequential grouping processes adhere more to the principles of good continuation and similarity.

Sequential Grouping

The streaming phenomenon is the most common example of sequential grouping and is sometimes referred to as *fission* (van Noorden, 1975). It is thought to occur as a consequence of the ASA process and in general it can be described as follows. Frequency is one of the factors influencing our interpretation of a given auditory event. When two pure tones, A and B, of different frequency (one high and one low) and a duration of 100 ms each, are played at a slow rate in a cycle (3 tones per second, see Figure 1.2 left panel) listeners report hearing the up-and-down pitch pattern and a rhythm that contains all the tones. After speeding up the rate of repetition (12 tones per second, see Figure 1.2 right panel) the high and low tones start to separate and the sequence splits into two perceptual streams, one on the higher and one on the lower pitch. Intermediate speeds can result in an ambiguous organisation where the listener can alternate between a single ABABAB… percept or the two streams on different frequencies: AAA… and BBB… (Bregman and Ahad, 1996; van Noorden, 1975).

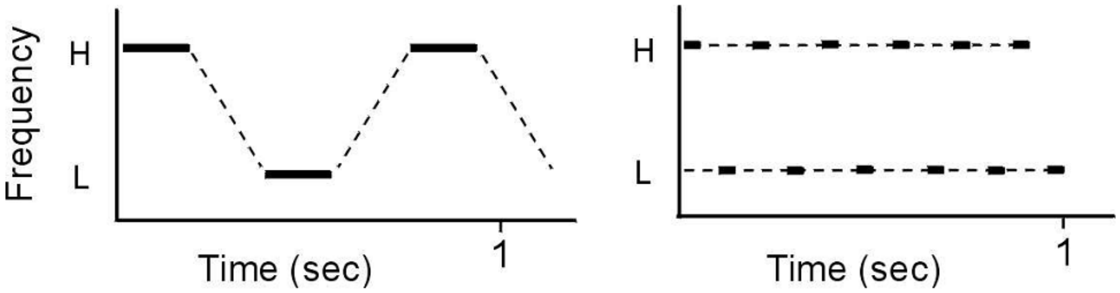


Figure 1.2 One second cycle of alternating high (H) and low frequency (L) tones. The rate of repetition is 3 tones per second on the left and 12 tones per second on the right. Dashed lines represent perceptual grouping. Adapted from Bregman (2004).

Although first described as such by Bregman and Campbell (1971), examples of the streaming phenomenon can be found much earlier in the literature (e.g., Ortmann, 1926; Miller and Heise, 1950). Using 100 ms tones, Miller and Heise (1950) found that a repeating sequence of alternating tones can be heard either as segregated or integrated depending on the frequency difference between the neighbouring tones. A frequency difference of about 15% (a whole tone, i.e. two semitones, is 12%) was sufficient for the separation into two streams to occur. Van Noorden (1975) extended the work on alternating tones by distinguishing between enforced segregation at large frequency differences and voluntary segregation at smaller frequency differences, under attentional control. Bregman and Campbell (1971) showed that as the average frequency separation is increased, a sequence of six notes breaks up into two streams. Although listeners could distinguish the temporal order of notes within a single stream, they failed to correctly judge the temporal order of notes across the two streams. In a related study, demonstrating that abrupt changes in acoustic properties can lead to segregation, Warren, Obusek, Farmer and Warren (1969) asked listeners to report the order of four sounds spliced into a repeating loop. The sounds were a hiss, a buzz, the phoneme ‘ee’, and a whistle. Regardless of the time spent listening to the sequence, participants’ performance was not different from chance. It was only improved when the sequence was slowed down to 700 ms per item (they were 200 ms each in the original sequence). Interestingly, the listeners could easily identify the sequence when the sounds described above were replaced by spoken digits (‘one’, ‘three’, ‘eight’, ‘two’, 200 ms each).

One of the first studies to look at the role of continuity (“smoothness of change”) in promoting segregation was by Bregman and Dannenbring (1973). Using alternating sequences of high and low frequency pure tones (an ABAB… sequence), they measured how large the frequency difference needed to be before it broke into the separate high and low stream in three conditions. In the *discrete* (classical) condition there was simple alternation between high and low notes (see Figure 1.3). In the *ramped* condition the silent gap between the tones was filled by introducing the frequency glide. In an intermediate condition, the *semi-ramped,* the two tones were pointing at each other without being physically connected. Bregman and Dannenbring showed that listeners tolerated the biggest frequency difference in the ramped condition, followed by semi-ramped, and finally the discrete condition, which tolerated the least separation. The authors concluded that the smoothness of change indicated by this unbroken spectral pattern, as in the ramped condition, helps to hold the sequence together. The results of the study were somewhat confounded by the fact that they could also have been explained by the frequency proximity cue. For example, the average AB frequency separation in the semi-ramped condition is effectively reduced by the tails on either end of the tone. Darwin and Bethell-Fox's (1977) study of stream segregation by abrupt changes in pitch (F0 frequency) of three vowel formants whose centre frequencies varied over time also looked at the principle of good continuation in hearing. Their study (see next section), which did not have the confounds of the Bregman and Dannenbring study, also indicated the importance of good continuation.

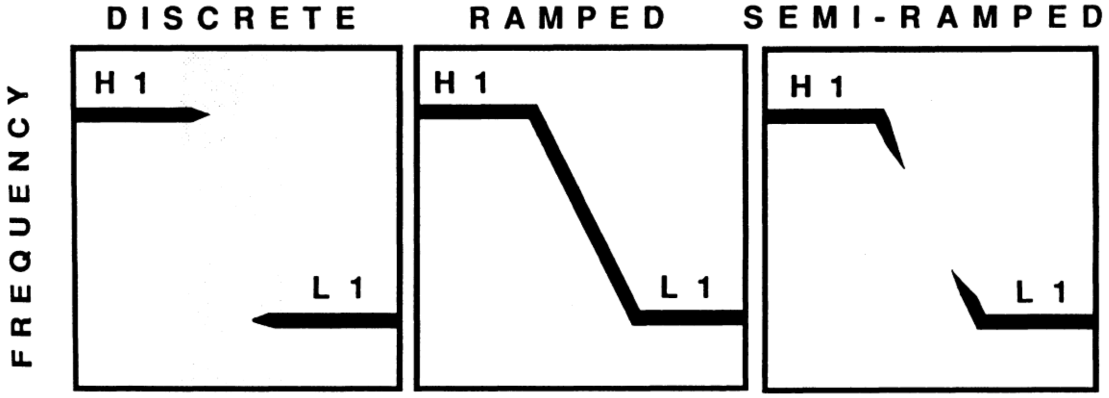


Figure 1.3 Three continuity conditions used in the study by Bregman and Dannenbring (1973). Taken from Bregman (1990)

Simply because two sounds have the same pitch does not necessarily mean that they will not segregate from one another, as they might still do so based on the differences in their timbre. Van Noorden (1975) studied this hypothesis using repeating ABA- sequences. In the Figure 1.4 below, shown on the left is an alternation between two sounds A and C which share the same underlying pitch. While A is simply a pure tone, C is a set of harmonics (the 3rd to the 10th) of the pure-tone frequency, which can be considered as the fundamental component. The two tones share the same F0, however, tone C has a missing fundamental (indicated by the dotted line). The right panel of the figure presents another variant, where both tones C and C' share the same (but missing) fundamentals. However, while tone C is defined by harmonics 3 to 5, tone C' is defined by harmonics 8 to 10. Van Noorden (1975) showed that even though these sounds all have the same underlying fundamental, if played in a sequence they readily undergo stream segregation. The reason that they do that is because, although they may share the same pitch, they have very different timbres. The ones with the low frequency harmonics sound very dull, and the ones with the high frequency harmonics sound very bright ('tinny'). Listeners exploit that difference in sound quality to segregate them from one another.

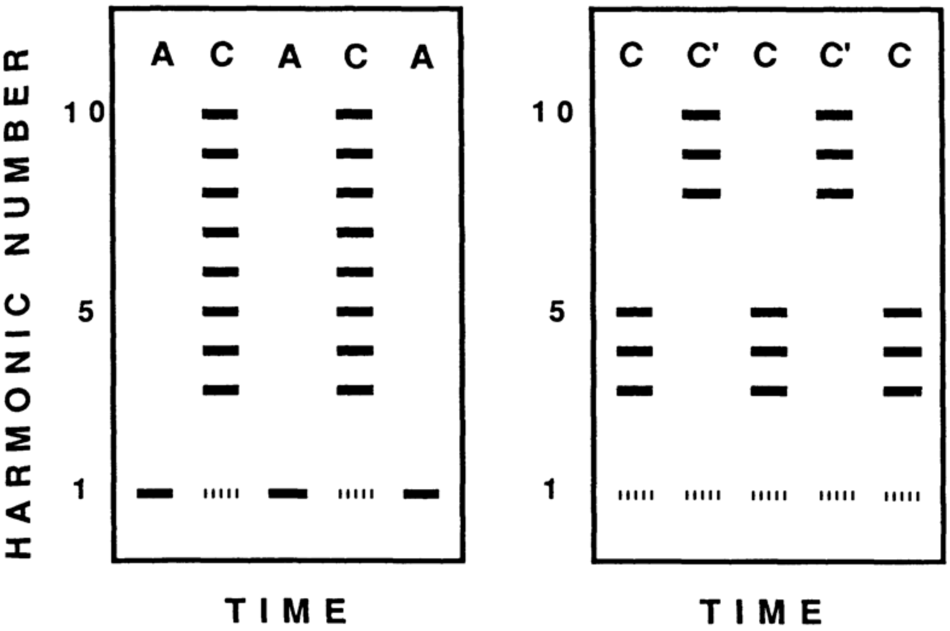


Figure 1.4 Stimuli sequences used by Van Noorden (1975).

The earlier mentioned experiment by Bregman and Pinker (1978) demonstrates the concept of competitive grouping. In their experiment, tones A and C are competing with each other to group tone B. As such, competition is a general property of ASA where different elements and principles compete with one another to control the organisation that we experience. In an arbitrary situation with streams X, Y and Z, it is possible for a cue (e.g. frequency proximity or timbre) to favour the grouping of X with Y. However, if another cue favours the grouping of Y with Z more, then that organisation will dominate our perception. In that sense, different cues compete with each other to produce the organisation that falls out of it, and the competition is an inherent part of the whole process. Hence, one situation may produce easily one type of organisation but the introduction of another cue might change this. Bregman, Liao and Levitan (1990) demonstrated this by varying how much difference there is between sounds on the different dimensions of either F0 or timbre and measuring under what circumstances listeners' responses were driven by formant frequency differences or F0 differences. Both factors influenced stream segregation, but the grouping that was heard depended on which of the two factors led to the greater perceived difference between the tones.

If streaming is viewed as a result of competitive grouping, tones tend to group with their nearest neighbour and the likelihood of them separating into two different streams depends on the acoustic dimensions of frequency and time. If the frequency separation is big enough, or the tones are repeated sufficiently rapidly, the two sounds will split into two perceptual streams (see Bregman & Dannenbring, 1973). While at the slower speed temporal separations are larger than the frequency separations, the opposite holds true for the case when the repetition rate is high. In the first situation (slower rate) tones will group with their nearest neighbour on the temporal scale and in the latter case (faster repetition rate) they will group with the nearest neighbour on the frequency scale (Bregman, 2004). Just as for sequences of pure tones which have high and low frequencies, we can segregate sounds from one another based on their fundamental frequency (F0). Hence the sounds which have a low F0 will tend to segregate from ones that have a high F0. It is not just a property of pure tones but also of periodic tones which have their own more complex pitch. This has been exploited by musicians, e.g. in effects of a difference between pitch range of the two parts in African xylophone music (Bregman & Ahad, 1996).

Exploiting differences in quality of sounds to segregate them from one another by timbre was also demonstrated by the Wessel illusion (1979), which shows how this can have complex consequences for the rhythm perceived. Wessel started with a very simple three-tone sequence of three relatively rapid sounds ascending in pitch. When the difference in F0 was modest, the sequence was heard a single stream. However, when alternate tones were played on sufficiently different timbres (every odd numbered tone had the 1st, 3rd , and 5th harmonics removed, and every even numbered tone had the 2nd, 4th, and 6th harmonic taken away) the original sequence streamed into two slower *descending* motifs.

Simultaneous Grouping

Simultaneous grouping involves the separation of a mixture of sounds occurring at the same time into separate streams. Our auditory system uses a set of cues that describe a given sound mixture, allowing the allocation of frequency components to the appropriate sound sources. For example, based on harmonicity cues, we are likely to assign sounds as if they originate from the same source if their frequency components are integer multiples of a common fundamental. Similarly, if the sound mixture contains sets of frequencies with different fundamentals, they will be treated as separate sounds. Broadbent and Ladefoged (1957) used the example of a person uttering a syllable. If the two vowel formants (resonances of the vocal tract) are given a different F0 they are assigned to different sources and two speakers are heard. If, on the other hand, the formants share a common F0, as in natural speech, the vowel sound is heard as fused. Similarly, it is much harder to separate one speaker from another if their voices are artificially modified to be on the same monotonous pitch than if their voice pitches are different. For example, a four semitone difference can increase the number of correctly identified words in a mixture of two voices from 40% to 60% (Brokx & Nooteboom, 1982). During comprehension of normal (non-synthesised) speech, listeners might also be exploiting the gaps within the speech stream associated with closures, as for plosive stops, to help separate two temporally overlapping voices. If the speech is presented without these pauses, it becomes more difficult to separate the two talkers (Bird & Darwin, 1998).

Another factor influencing simultaneous grouping is the synchrony of onsets and offsets of components, as frequency components which start and stop at different times are less likely to be grouped together and more likely to be perceptually segregated. Darwin (1984) found that the phonetic quality of a vowel can be affected if a harmonic in the F1 (first formant) region starts earlier or stops later than the other harmonics by few tens of milliseconds. In another example, mistuning a single harmonic in a sequence causes it to be heard out as separate tone. One line of evidence showing that harmonic templates can be used to pick out different fundamental frequencies comes from Brunstrom and Roberts (1998). They used a set of 14 harmonics, with three experimental conditions where certain harmonics were removed (for condition 1 it was the 6th and 7th, for condition 2 – the 6th to the 8th and for condition 3 – the 6th to the 9th). The spectral gap from the removed harmonics was replaced with a single probe. Listeners were asked to listen for a pure-tone-like sound in the complex and to adjust another pure tone to match its pitch. Brunstrom and Roberts showed that if the probe lined up with one of the missing harmonics, it tended to fuse / integrate with the other components, and hence was difficult to hear out. If, on the other hand, the probe was in a mistuned position it tended to segregate from the rest of the complex. Whenever the probe matched the position of the missing harmonics there were clear minima visible in the matching results. In their second experiment, Brunstrom and Roberts presented evidence indicating the activation of two harmonic templates at the same time. This indicated a mechanism that allows concurrent harmonic complex tones on different F0s (e.g., voiced speech on different pitches) to be segregated from one another.

An *interaural time difference* (ITD) is the difference in time it takes for a sound lateralised in the left-right plane to arrive at the two ears. By itself, an ITD is a weak cue for simultaneous grouping (Culling & Summerfield, 1995; Hukin & Darwin, 1995; Shackleton & Meddis, 1992), although it can assist other cues in segregating components (Darwin, 1997). However, the role of ITD cues in simultaneous grouping can be contrasted with the improvement in intelligibility when the on-going voices of two talkers comes from two different locations (Bronkhorst & Plomp, 1992). Darwin and Hukin (1999) showed that listeners can utilise ITD cues to track a voice across time in the presence of another sound source.

1.4 ASA and the perceptual organisation of speech

Unlike the simple stimuli on which the ASA account has been primarily based, speech has two particular features: it is acoustically diverse and it is rapidly changing. The human vocal apparatus, particularly our larynx, tongue, lips and jaw, can produce a complex signal with different sources. In a very short period of time, we can differentiate between quite disparate acoustic segments: vocal cord vibration during the production of voiced vowels, plosive bursts characterised by the stop and release of the air flow (as in 'b' for 'bat'), fricative hisses which are sounds produced by air turbulence due to constriction of our vocal tract (as in 's' for 'sit'), and formant transitions which can be defined as frequency glides between resonances when we progress from uttering one phoneme to another.

The model of speech production by Fant (1960) is known as the Source-Filter model. First proposed as a theory for vowel production it is now an accepted doctrine in speech acoustics. It describes speech production as a two stage process. Sounds are first produced at the *source* by the vibrating vocal folds (the glottal source) and then they are *filtered* by the vocal tract (whose resonances, known as formants, shape the spectrum of the vowel). Source-filter theory can be generalised to consonants, where the source may arise from frication or plosion instead of (or in addition to) voicing. Whatever the source or sources, these sounds will be modified/filtered by the shape of the vocal tract, and each one of these shapes has its own filter function or transfer function associated with it, which arise from its associated resonances. These resonances shift around as vocal-tract shape changes, providing the filter function. When the source is passed through the filter, the output emerging from the lips will be the outcome of the process. The other feature of the model is that the source and the filter are independently controllable. For example, the speaker can adjust the glottal source by changing the vibration of their vocal cords (so that they can raise or lower the pitch while keeping the filter shape exactly the same). Similarly the speaker can keep the source the same but change the filter function (hence differentiating between vowel sounds on the same pitch). For these reasons, speech is both dynamic (time-varying) and acoustically diverse. This diversity comes from the fact that there are different acoustic sources within a single speaker and that these can be switched on and off almost instantaneously. This sudden switching on and off of vocal cord vibration can trigger frication or (very short) plosive bursts. This rapidly changing distribution of energy across the frequency spectrum is accompanied by rapid switches between the buzz source of the vibrating vocal cords and a noisy source such as frication. The frequencies of the lowest three formants in particular, as well as their pattern of change over time, provide cues that help listeners determine the phonetic identities of vowels and consonants (Assmann & Summerfield, 2004).

This heterogeneous and dynamic nature of speech has major implications for grouping of its elements. Speech is usually heard as a single stream, which raises a question of how we can reconcile the rules and principles of scene analysis / perceptual grouping with these complex stimuli. Not only does speech sound coherent when heard in isolation, but we are usually able to hold speech together in a coherent stream in the presence of other speech or of non-linguistic sounds – the so called cocktail party phenomenon (Cherry, 1953). Experiments involving grouping of speech sounds or speech analogues have provided some clues as to the underlying mechanism of the cohesion of speech, but the results are still open to interpretation. Indeed, some researchers argue that the outcomes of studies involving simple stimuli cannot be applied to more complex signals such as speech (see Remez, Rubin, Pisoni and Carrell, 1981; Remez, Rubin, Berns, Pardo and Lang, 1994). There are, however, several studies that have used synthesized speech signals to investigate the relative contribution of general purpose ASA principles to understanding the perceptual organisation of speech; these are reviewed below.

Darwin and Bethell-Fox (1977) explored the sequential grouping of speech sounds using synthetic vowel-like stimuli with three formants. Figure 1.5 A shows that in the starting phase the frequency of the first formant F1 was relatively low, and the second (F2) and third formants (F3) were positioned close to each other in terms of frequency separation (on its own sounding like the vowel ‘ee’). In the next phase, the F1 frequency was higher and the F2 frequency was lower (on its own sounding like vowel ‘aa’). The two phases were linked by smooth, linear formant transitions. Darwin and Bethell-Fox showed that if the pitch of a sequence of sounds such as in Figure 1.5 A is constant, or if it changes gradually and "smoothly" over time, listeners hear a single sequence of speech and this is heard as the repeating diphthong 'yayaya…'. However, after introducing abrupt changes in F0 (the distinctive step functions in Figure 1.5 B) between the two phases (i.e., falling in the transition zones), listeners report hearing two streams – one on the low and one on the high pitch. In addition, as each voice is heard to be silent while the other is speaking, one of the voices (as shown in B) results in a formant pattern and silence that is heard as 'gagaga…'. The silence that is necessary to produce the perception of a stop consonant, such as 'g', is not physically present in the stimulus but is a result of stream segregation. In contrast, slow changes in F0 do not produce stream segregation and do not generate stop-consonant percepts. This strongly suggests that pitch contours facilitate the integration of voiced speech elements into a single coherent stream, the aforementioned principle of good continuation. The question remains about how other speech sounds, especially the voiceless plosives and fricatives, are able to cohere with the voiced speech segments and not to segregate into different perceptual stream.

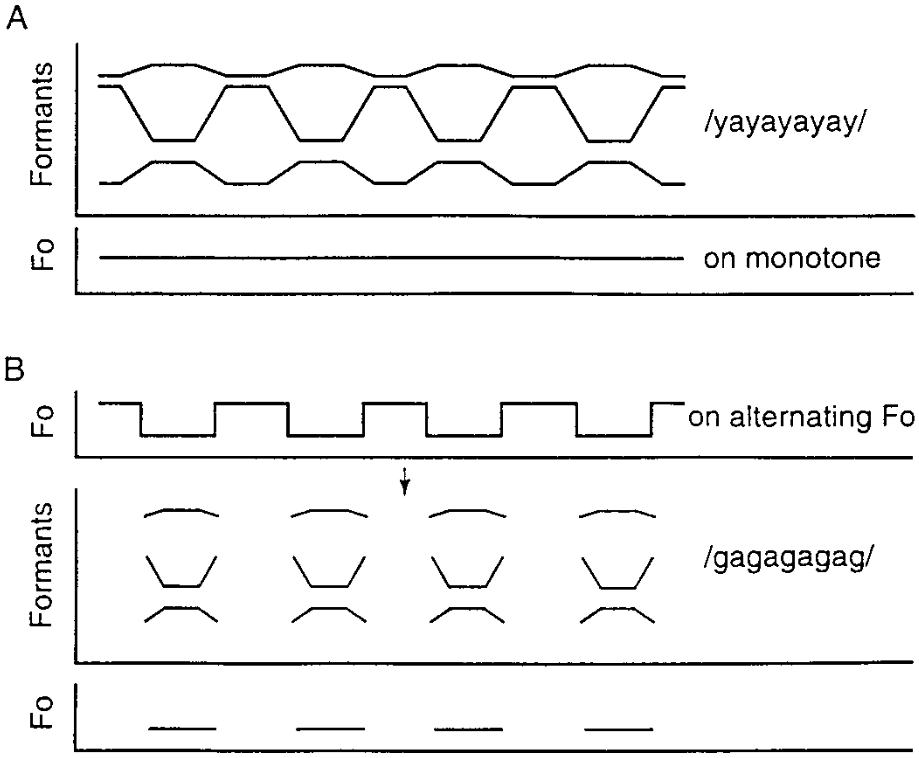


Figure 1.5 Sequences used in the Darwin & Bethell-Fox (1977) experiment

One study that has addressed this question is that by Cole and Scott (1973). They investigated the streaming of a single repeated syllable, such as ‘sa’ (presented in Figure 1.6). The left panel in the figure shows the syllable with intact natural formant transitions between the consonant and the vowel (unaltered recording) and the right panel shows the same syllable with the transitions spliced out. These formant transitions are the acoustic consequences of the changes in the configuration of the vocal tract when the articulation moves from one sound element to another – e.g. from the fricative ‘s’ to the vowel ‘a’. To investigate the importance of formant transitions in grouping, Cole and Scott used a combination of acoustically different sounds: voiceless consonants such as ‘s’, which are characterised by a noise burst, and voiced vowels which include pitch information. The high-frequency fricative burst of ‘s’ can be clearly differentiated from the first 5 formants of the vowel ‘a’ as well as from the spectral movements of the vowels into the frication – the formant transitions. During speech production the shape of our vocal tract is constantly changing and as this change, physically, cannot be totally abrupt (compare with almost instantaneous changes in the source of excitation, e.g. from frication to voicing), it manifests itself in the formant transitions.

Cole and Scott argued that formant transitions will help the voiceless fricative adhere onto the voiced vowel as that was part of their role in perceptual grouping – a process akin to other continuity cues. Using a relatively crude analogue tape-splicing procedure, they produced ‘transitionless’ versions of each tested syllable (consonants such as ‘s’, ‘sh’, and ‘v’ combined with vowel ‘a’), leaving only the consonant and the steady state formants of the vowels afterwards. Their results indicated that when played in isolation, the two versions of the syllable, with- and without- the formant transitions, were heard as essentially the same; participants had no difficulty in identifying the syllables. A difference emerged, however, when syllables were presented in a rapid sequence in a standard streaming task. Their results showed that the sequence where the transitions were preserved held together as a single stream much better than the one where the transitions had been spliced out. Syllables without the transitions broke up relatively quickly into two separate streams, one containing the voiceless fricative and the other the vowel. Syllables with the transitions would eventually stream – if the repetition rate was high enough and they were presented for a long duration – but nonetheless these stimuli were much more resistant to streaming. Cole and Scott concluded that formant transitions aid in perception of the temporal order of speech sounds.

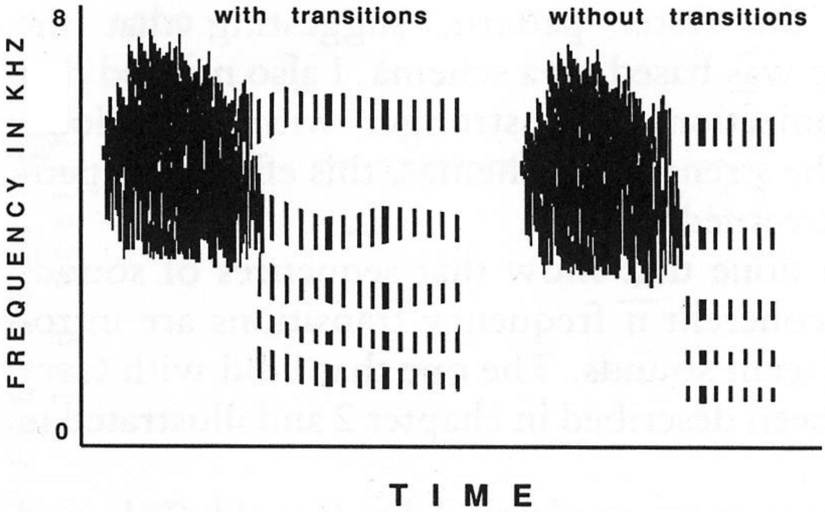


Figure 1.6 Syllable /sa/ used by Cole and Scott (1973). With formant transitions intact on the left and with formant transitions spliced out on the right.

Note, however, that formant transitions are not important only in the case of holding voiced and unvoiced speech segments together. Just like the pitch contour, formant transitions appear also to play a role in linking together voiced segments. Dorman, Cutting and Raphael (1975) presented listeners with an order judgment task. After hearing a sequence of four voiced vowels (with an F0 of 110 Hz), participants were asked to write down the order in which the vowels occurred. In all five experimental conditions (see Figure 1.7), the vowels themselves did not differ, but the context in which they were presented varied. The simplest conditions from the acoustical arrangement point of view were: the *long vowels* condition, in which the vowels occupied the whole time interval with an abrupt change from one to the next one, and the *short vowels* case, where vowels have been shortened and the space between them was filled with silences. Dorman et al. used three-formant approximations to the vowel, where the simulations were based on the lowest three formants. They found that if either of these two cases (long or short vowels) were played in a rapid sequence, vowels very quickly broke up into separate streams based on the similarity of their formant frequencies.

The *connected vowels* case was characterized by each of the vowels linked to the next one by formant transitions gliding continuously over the course of 95 msec from the steady-state formant values of one vowel into the succeeding vowel. Order accuracy for these stimuli was much higher; it was much easier to identify the order of the vowels in that sequence. This is in line with the principle of good continuation, which proposes that smooth progressive changes between the sounds will facilitate the cohesion of a stream.

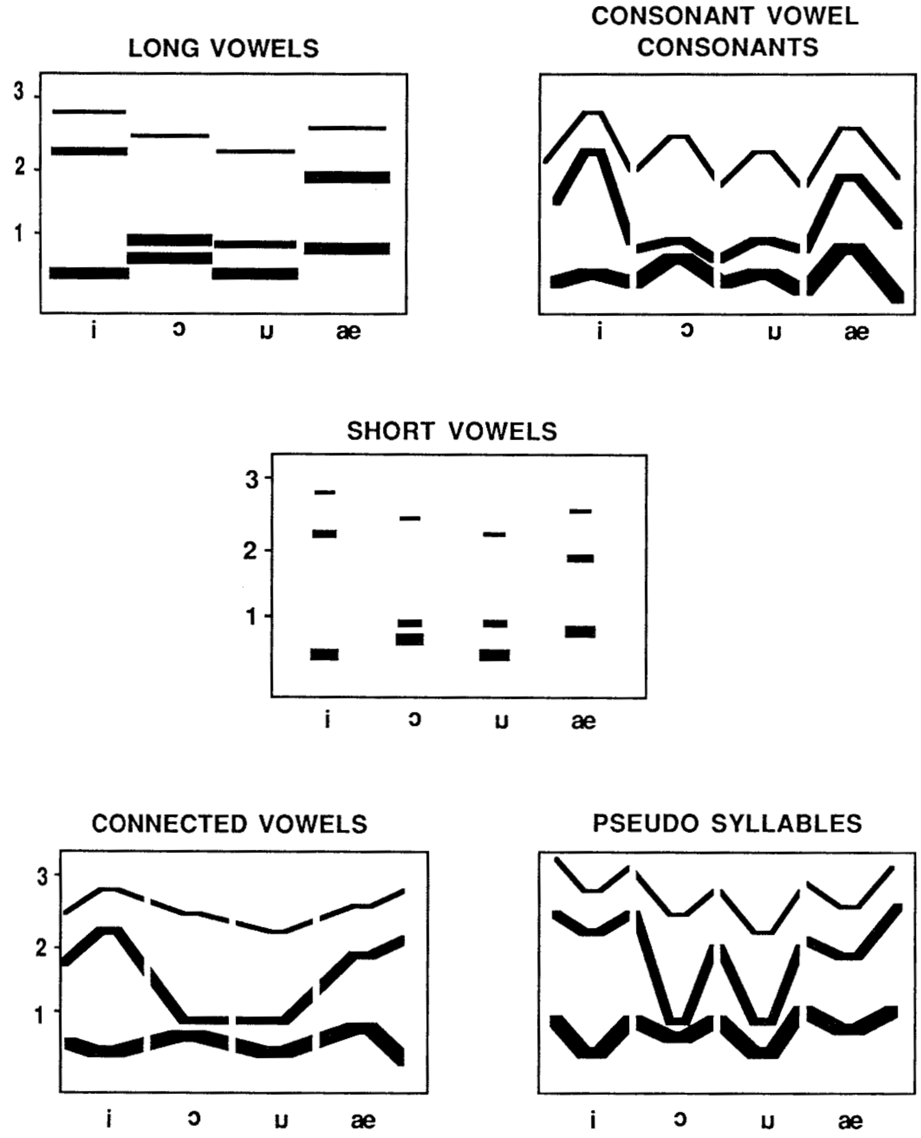


Figure 1.7 Five experimental conditions from the Dorman et al. (1975) study.

The most interesting outcome of the study was the results for the final two conditions: *consonant-vowel-consonants* (CVC) and *pseudo syllables*. The CVC case had the four vowels specified by the same formant frequencies as in previous conditions, however they included formant transitions that moved into and out of them (see top right picture in Figure 1.7). These formant transitions were designed to stimulate a stop consonant, in this case: ‘b’. Results indicated that for the CVC sequence (like for the *connected vowels* condition), it was much easier to judge the order of the vowels than for the isolated tokens (*long vowels* condition), even though the pattern of formant transitions for linking the vowels together was much more complex than in the connected vowels condition. The pseudo syllables case differed from the CVC configuration in only one respect – the formant transitions were inverted, which resulted in their frequencies falling into the vowel and rising out of it. The advantage of this design was that the formant transitions in the pseudo syllables condition produced implausible speech sounds. Even though their continuity pattern was similar to that of CVC syllables the order judgments were very poor.

The results of this study suggest that the auditory system is not exclusively using simple continuity cues for the sequential grouping of speech sounds. Rather, it appears to use our knowledge of the structure of language. If the sequence is formed of plausible formant motions, listeners are able to integrate it into a single stream and to successfully identify the order of the sounds in it. If, on the other hand, the sequence contains implausible formant transitions between the vowels, such a sequence is prone to break up into separate streams regardless of the apparently similar degree of continuity. This provides evidence that, aside from the importance of formant transitions and F0 in holding speech sounds together perceptually, there are other factors (such as the linguistic plausibility of the transitions) which suggest that the grouping of speech is governed by both primitive and higher order schema-based organisations.

In summary, sequential grouping can be affected by differences between complex tones in their spectral composition – timbre (Wessel, 1979), spatial location from the listener (Deutsch, 1979), repetition rates of their waveforms – pitch (Darwin and Bethell-Fox, 1977), or transitions between the two sounds (Dorman et al., 1975). The last point will be evident with the discussion of the studies on the importance of formant transitions in speech. In essence, if tone A changes gradually into tone B it will be heard as a single changing sound. If on the other hand, the change is abrupt, the listener will tend to perceive the second tone B as a different sound from a different source.

Clearly, sequential integration is not just involved in the grouping of a sequence of discrete tones but is also applicable to the understanding of perceptual grouping involved in more complex sounds. According to Warren (2008), successful grouping of sounds based on the above properties has several effects on the perception of more complex stimuli. (1) Judgments of the timing and order of two sounds are easier if they belong to one stream; this is especially critical to speech where one must hear the intended phonemes in the right order for speech to be comprehensible (Warren et al., 1969; Bregman & Campbell, 1971); (2) melodies and rhythms are formed within auditory streams, and (3) sudden changes in the fundamental of the voice result in the loss of speech continuity and emergence of a new stream, which sounds as if one talker has been replaced with a different one (Darwin, 1997).

1.5 The Verbal Transformation Effect

Bregman (1990) argued that scene analysis "involves putting evidence together into structure" (p. 15). Illusions demonstrate a failure to achieve that structure even though the particular elements of it have been identified. Even though the assignment process of the evidence is taking place, the resulting descriptions of our environment are not correct. Studies of auditory illusions have provided valuable information on the general mechanisms underlying auditory perception as well as their role in understanding speech (Warren, 1996).

One phenomenon believed to reflect the operation of perceptual mechanisms under difficult listening conditions is the verbal transformation effect (VTE). Upon listening to a recycled word, participants report hearing illusory changes to the initial stimulus. For example, a 3-minute presentation of a repeated word "ripe" may include the following responses: *ripe*, *right*, *white*, *white-light*, *right*, *right-light*, *ripe*, *right*, *ripe*, *bright-light*, *right*, *ripe*, *bright-light*, *right*, *bright-light* (after Warren, 1961a). The changes can be quite complex phonetically and they sometimes suggest semantic linkages. The usual procedure involves a 3 to 5 minute presentation of a repeated syllable, word or a sentence. Two measures are taken – the number of verbal transformations (any change to a previously reported utterance) and the number of forms (unique transformations). Therefore, in the above example there are a total of 14 transformations but only 6 forms (ripe, right, white, white-light, right-light, bright-light). In the early days of VTE research, listeners wrote their responses phonetically on a sheet of paper or spoke them out loud to an experimenter sitting in front of them (see Fig. 1.8). Later on (from the mid 70's onwards) listeners were seated in a sound booth and spoke their transformations into a microphone. Interestingly, although no systematic exploration of this has been reported in the context of the VTE, in the closely related *phonemic transformation effect* (PTE, see Chapter 4), Chalikia and Warren (1991) noted that there is no evidence of inconsistencies between written and verbal reports in the PTE.

Richard Warren was the first to investigate the VTE experimentally. He postulated that, upon experiencing a repeated word, the initial organisation of the speech sounds into words or phrases may not be confirmed by contextual information. Under such unusual listening conditions, verbal transformations (VTs) may be temporarily accepted in order to decipher what the speaker is saying (Warren, 1968). As such, the VTE can be seen as a product of the normal constructive nature of speech processing, guarding listeners against error under the imperfect listening conditions (Kaminska, Pool, and Mayer, 2000). Given that speech comprehension can be regarded as a highly skilled perceptual process that happens without conscious effort on our part, the underlying mechanisms remain hidden in everyday life. Warren (1996) argued that as an illusion, the VTE demonstrates breakdown in perceptual accuracy, and therefore, at least in the experimental setting can be used to study normally inaccessible processes (Warren, 1996). In general, the VTE appears to be related to the mechanisms employed normally for the prevention of errors and resolving ambiguities in speech perception. The paradigm for the VTE seems to involve two general principles – *verbal satiation* (loss of a particular verbal organization resulting from a continuous exposure to a stimulus) and consequently the emergence of a different form resulting from a *shift in perceptual criteria*. Next, due to the lack of normal context, the process is recycled and the new form undergoes satiation and replacement (Warren, 1996). In other words, the recycled word activates several candidate lexical items and the item with most activation is the one perceived. However, repeated stimulation causes fatigue of the activated items, and this is greatest for the most activated item, which in turn results in a change of currently perceived word.

Following on from observations by Warren and Gregory (1958), in the first reported systematic VTE study, Warren (1961a) asked participants to listen to a monosyllabic word, polysyllabic word, or short sentence played in a loop on a tape. Eighteen listeners were asked to call out what they heard initially and subsequently to report any changes that occurred to the stimulus. All listeners were unaware of the illusion and reported that the changes to the stimulus they experienced during presentation were real. The major findings were that all experimental stimuli evoked verbal transformations and that the rate of verbal utterances across a 3 minute presentation after the first VT occurred at an approximately constant rate for transformations but decreased for new forms.

Initially, Warren compared the illusion to visual reversible figures (Warren and Gregory, 1958), although he later pointed out that marked differences exist between the two. This comparison is considered further below. However, in broad terms, both phenomena seem to reflect the principle stated by John Locke (1690, as cited in Warren, 1996) that no particular

Figure 1.8 Experimental procedure in the early VTE study. The participant listens to the recording of repeated words over headphones and reports VTs out loud. At the same time, the experimenter notes down the responses. (Taken from Warren and Warren, 1970)

thought or perceptual organization can be maintained without change for any length of time. In summary, (i) VTs occur over a wide range of stimuli like syllables, words or phrases; (ii) they sometimes involve considerable distortions from the original percept; (iii) responses vary considerably between participants, and (iv) they generally produce more forms in 2 or 3 minutes - while with reversible figures there are typically only two forms possible, VTE can potentially elicit an indefinite number of forms.

Warren's pioneering study initiated a considerable body of research, peaking in the '60s and '70s, which concentrated on the different aspects of the illusion. The studies which followed indicated that the VTE may be a valuable tool for studying speech perception. Some of the themes investigated included phonetic analysis of VTs as seen for words and syllables (Ohde and Sharf, 1979; Lass and Golden, 1971, Naeser and Lilly, 1970), the effect of age (Warren, 1961b, Warren and Warren, 1966), the effect of listeners' phonetic training (Lass and Gasperini, 1973), or the influence of inter-trial time interval (Warren, Healy, and Chalikia, 1996). Researchers have also looked at the effect of adding continuous noise on the VTE (Warren, 1961a, Sadler, 1989 as cited in Warren, 1996) or the effect of transitions evoked by concurrent nonverbal stimuli – such as repetitions of white noise bursts (Lass, West, and Taft, 1973), tone bursts (Fenelon and Blayden, 1968; Perl, 1970; Lass, West, and Taft, 1973), and melodic phrases (Guilford and Nelson, 1936; Lass, West, and Taft, 1973).

In general, the stimuli used in these studies have ranged from steady-state vowels, through syllables and words, to short sentences. The rate of VTs, despite considerable individual differences, is around 5 – 10 changes per minute for young adults (18 – 25 yr). This rate is higher for children (8 yr) at 33.7 VTs/3min (9.7 Forms/3min) and lower for the elderly (62 – 86 yr) at 5.6 VTs/3min (2.6 Forms/3min) (Warren and Warren, 1966). Young adults tend to report neologisms and, while children often violate phonotactic rules, older participants almost always restrict their responses to lexical items. Phonetically trained listeners tend to report more forms and VTs; they also require fewer repetitions of the stimuli to report the first illusory change. Reducing the repetition rate by introducing silent gaps between cycles results in a proportional decrease in the rate of VTs (i.e., the same number of repetitions produces the same number of changes). Additionally, participants tend to report more transformations when listening to pseudowords than words (Natsoulas, 1965). The rate of VT is equivalent for monaural and diotic presentations (Warren, 1961b). Interestingly, adding continuous noise has an adverse effect on the number of VTs, in the sense that the partially masked VT stimulus elicits fewer transformations (Warren, 1961a).

The VTE phenomenon is still being used today to investigate various aspects of auditory perception. Bashford, Warren and Lenz (2006) found that 78% of the first reported VTs were lexical neighbours (using the scale of frequency-weighted neighbourhood density – FWND), differing from the original stimulus by a single phoneme. In addition, the amount of time the stimulus was heard "non-veridically" (i.e., as different from the original percept) declined during presentation and decreased with both increasing neighbourhood density and increasing neighbourhood spread (i.e. the number of stimulus phonemes that can be changed to form a lexical neighbour; Bashford, Warren & Lenz, 2009). More recently, in a study employing fMRI analysis and investigating the neuroanatomy of the VTE, Kashino and Kondo (2012) reported activity in frontal areas of the brain while listeners were asked to respond by pressing a button either to a repeated word 'banana' in a verbal transformation condition or a tone pip in a tone detection condition. While both tasks evoked activation within the primary auditory cortex, additional activation was found in anterior cingulate cortex, the prefrontal cortex and the left inferior frontal gyrus for the VT task only.

The VTE can also be viewed within a framework of the *multistability of perception* , which has seen recent advances and interest in the auditory domain (Schwartz et al, 2012). Multistability, originally studied extensively in vision, refers to perceptual organisation where a single physical stimulus can produce alternations between different subjective percepts. The classic example of an ambiguous image is the vase-faces illusion (the Rubin's vase, see Schwartz et al, 2012) where a single figure can be viewed as either an outline of a vase or as two faces. The image is perceptually segregated into two percepts with viewers able to switch between the organisations with ease. In the auditory domain, stream segregation and its build up over time has been extensively studied in the past (e.g. Bregman & Campbell, 1971; Miller & Heise, 1950; van Noorden, 1975) where two auditory streams are perceived while listening to a single sequence of sounds. In general, for a repeating sequence of high and low tones, the likelihood of segregation increases over time and this build-up of stream segregation is most noticeable in the first few seconds of a tone sequence (Bregman, 1978; Anstis & Saida, 1985). It has been recently shown, however, that after the initial period of a strong bias to a single-stream organisation, the subsequent percepts are bi-stable and continue to change between segregation and integration (Pressnitzer & Hupé, 2006; Denham & Winkler, 2006). The VTE can be viewed as a case of multistability where the initial organisation (e.g., a repeated word), can produce alternations between many different subjective percepts – verbal transformations of that word. Interestingly, the multistability aspect can also manifest itself as bi-stability, where for prolonged periods of time when the repeating sequence is presented, listeners experience switching between two dominant forms (Ditzinger, Tuller & Kelso, 1997). Based on the data from the few neuroimaging studies using the VTE, Basirat, Schwartz and Sato (2012) described the mechanisms of the VTE within the general framework of multistability and, more specifically, the ‘predictive coding’ approach where the sensory input is constantly compared with its pre-stored schema like prediction. When the two entities do not match an error signal is generated. Within that framework, a prediction – e.g. the expectation of a word embedded in some linguistic context such as a sentence – is compared against an input signal – the repeating sequence of words. An error message is then sent to the perceptual system allowing the re-evaluation of the sensory input – and the re-emergence of a different verbal transformation.

Despite the myriad of themes in the above mentioned studies, very few researchers have focussed specifically on the acoustic-phonetic factors involved in VTE. The existing studies have either looked at this problem indirectly or in insufficient detail, but they do constitute first attempts to quantify the nature of verbal transformations from the perspective of perceptual organisation and to identify the patterns involved in the grouping of speech sounds in the context of this phenomenon. Barnett (1964, as cited in Warren, 1996) was the first to attempt to analyse VTs by looking at the phonetic content of the responses to a variety of words. Both consonants and vowels were prone to change and produced illusory changes and “stability was noted for the voicing property of consonants and the type of movement characteristic of individual consonants and vowels. Intervowel glides were generally stable both in position and type of movement” (Barnett, 1964, as cited in Warren, 1996, p.453).

Naeser and Lilly (1970) looked primarily at the difference of responses between linguists and non-linguists when listening to the repeated word “cogitate”. They noted that both groups gave similar responses but more interestingly they commented on the type of phonetic changes given. Consonants generally were substituted by place of articulation but not by manner for example, plosives were usually substituted with other plosives. On the other hand, vowels were most often substituted on the basis of similarity of the position of the tongue. Clegg (1971) used 18 separate repeating syllables, each consisting of a different consonant followed by the vowel ‘ee’. He was interested in analysing the transformations reported by listeners according to several linguistic features: voicing, nasality, affrication, duration and place of articulation. Focusing on the transformations of the consonants, Clegg concluded that a consonant and its transform tended to share the features of voicing, nasality, and affrication but not of duration and place of articulation. Using a similar methodology, Evans and Wilson

(1968) looked at VTs reported to a series of syllables consisting of a range of consonants and one vowel. Their analysis of VTs for the consonant revealed a high frequency of responses involving the aspirated phoneme ‘h’.

Using six vowels and six consonants, Goldstein and Lackner (1973) constructed 30 nonsense syllables and analysed the responses in terms of VTs and forms elicited by participants. After summarising the types of changes according to distinctive features they concluded that VTs are “very systematic” (cf. Lackner, Tuller, and Goldstein, 1977). Although the phenomenon continues to draw attention, it is still believed to defy a satisfactory theoretical explanation and the dynamics of the changes in the VTE are poorly understood (Tuller, Ding, and Kelso, 1997; Kaminska & Mayer, 2002). Some recent investigations, however, have indicated that the switching between lexical interpretations shows properties in common with the perception of visual reversible figures (as first observed by Warren & Gregory, 1958), particularly rapid and long alterations between pairs of transformations (Ditzinger, Tuller & Kelso, 1997; Ditzinger, Tuller, Haken & Kelso, 1997), and that the perceptual regrouping of speech sounds plays a key role in the VTE (Pitt & Shoaf, 2002; Shoaf & Pitt, 2002).

Perceptual Re-Grouping in the VTE

Ditzinger, Tuller and Kelso (1997) showed that although listeners may experience a large number of different forms in the course of a VTE experiment, these do not occur at random; rather, they are usually organised into pairs. By slowing the rate of stimulus presentation and considerably increasing the number of stimulus presentations, Ditzinger et al. were able to quantify the characteristics of verbal transformations. The syllable /ke/ repeated 1000 times with a 500-ms silent interval between each repetition was presented to the listeners in 10 sessions. All participants experienced changes, but notably these were characterized by oscillations between two perceptual pairs, where one of the percepts was always /ke/ and the other was different for each listener. Those oscillations occurred much faster than new forms and the authors noted that for these pairs: “…perception remains tied to the acoustics because the actual syllable presented is always one of the two most often reported forms. Moreover, listeners tend to cycle between only two forms at a time, not three or more.” (p. 31). These results indicate that such pairwise oscillations may show underlying mechanisms similar to alternating between interpretations of ambiguous visual figures.

Another factor contributing to our understanding of the inner workings of the VTE was investigated by Pitt and Shoaf (2002), who studied perceptual regrouping of phonetic segments in VTs. Pitt and Shoaf (2002) demonstrated that streaming based VTs depend on the acoustic properties of the stimuli. More specifically, phonetic elements such as fricatives, affricates and plosive stops cohere less well with adjacent phonemes and therefore are more prone to streaming. Therefore, participants' responses to a repeated stimulus should reflect the properties typical of regrouping percepts, such as grouping based on frequency proximity or good continuation. Pitt and Shoaf presented listeners with CVC pseudowords with varying degrees of acoustical binding between the consonants and the vowel. The experimental conditions included the *Intact* condition where the consonants were approximants and nasals (e.g., /lom/ and /wEm/), the *Final* condition where fricatives and affricates were at the terminal position in the syllable (e.g., /lodZ/, /wEtS/), and the I+F (*Initial plus Final*) condition where fricatives, affricates and stops occupied both consonant positions (e.g., /podZ/, /pEtS/). Pitt and Shoaf argued that, based on the principles of perceptual streaming, the first condition should be the most resistant to streaming as both nasals and vowels are periodic signals occupying similar frequency regions. Listeners were instructed to report the transformations and the number of streams they heard. The presence of multiple streams was reported 60% of the time and in all of those cases the transformation consisted of a foreground stream including a consonant and a vowel and a background stream containing only a consonant. Across the three conditions, the relative cohesiveness with the vowel was representative of whether the consonants will split off. As expected, in the Final condition only the terminal consonants streamed off. In the I+F case, typically only one consonant split off (usually the final one), but there were also reports of both consonants segregating at the same time, with the vowel forming a separate stream.

In 1976, Warren and Ackroff demonstrated that it is possible to stimulate each ear with the same repeated word without hearing the word as a single fused image. Two copies of the word 'tress', with a repetition period of 492 ms, were separated from one another by an interaural delay of half the repetition rate so that temporally offset but otherwise identical stimuli were heard in each ear over headphones (see Figure 1.9). Neither ear could be considered as leading with the half-cycle delay.

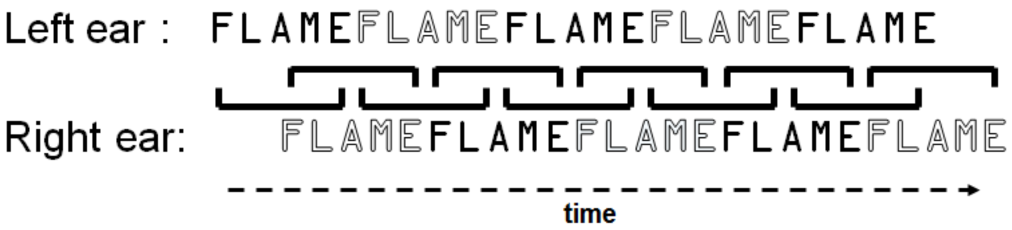


Figure 1.9 Warren and Ackroff's (1976) experimental setup, using the example word 'flame'

Warren and Ackroff were interested in whether or not the same illusory changes would be heard simultaneously on the right and left. It was found that for each of 20 subjects, the times at which changes occurred were uncorrelated at each ear. Also, the forms heard at the two sides were independent, so that while the word "dress" might be perceived at one ear, a word as far removed phonetically from the repeating stimulus "tress" as "commence" might be heard at the other. However, no description was given as to how this was measured.[1] Warren and Ackroff were interested in the so-called right ear advantage (e.g., Kimura, 1961) and investigated whether separate or identical linguistic processors are used for processing the acoustically identical verbal stimuli. However, the study is of some interest in relation to the issue of the perceptual regrouping of acoustic elements in the speech signal. Namely, as the two recycling words are in competition with each other, systematic investigation of the effect of factors such as fundamental (F0) frequency and interaural time difference could inform us about their respective role in the perceptual regroupings of the verbal transformations. This relates to the classic cocktail party situation, where more than one person is speaking at once and our auditory system needs to separate the required information from a mixture of broadband dynamic sounds.

1.6 Summary and Orientation to the Thesis

To reiterate, a difficult task for our auditory system is to separate out the sounds that come from different events in the environment and to group together sound streams originating

[1] In their discussion, Warren and Ackroff (1976) also reported an unpublished finding of independent transformations achieved with three asynchronous versions of the same word presented at the same time: two monaural inputs (on the left and right), and one diotic input, forming a centralised auditory image. This finding was replicated by Zuck (1992), however, the measure of independence used was a difference in the overall number of VTs heard for each sequence rather than the difference at any given time between linguistic forms for the left, right and central percepts.

from the same source. As outlined in the previous section, auditory scene analysis is governed by a set of general principles for grouping sound elements; however, despite a large body of research these general principles do not seem to account sufficiently for the fact that the rapidly changing and diverse acoustic elements of speech cohere to form a single perceptual event (Remez, Rubin, Berns, Pardo, and Lang, 1994). Although repeated speech seems to break into separate streams for the same reasons as tones and other nonspeech objects (Chalikia and Warren, 1994; Cole and Scott, 1973), acoustic features of speech such as alternation of aperiodic noises (e.g., fricative consonants) and periodic segments (e.g., vowels) often seem to violate the principles of perceptual organisation as specified by auditory scene analysis (Pitt and Shoaf, 2002).

Recent research has demonstrated that the VTE can be a useful tool for exploring the perceptual organisation of speech sounds. It reveals the perceptual changes to linguistic form that can occur with an unchanging pattern of acoustic stimulation. It has been argued that the VTE reflects mechanisms involved in the correction of errors under difficult listening conditions. In the situation where two repeating sequences of speech material are presented at the same time, there is a potential for simultaneous grouping factors to influence the rate and the type of VTs. However, the only study of this type reported to date (Warren and Ackroff, 1976) used two sequences presented to separate ears. This configuration largely precludes re-grouping interactions between phonetic segments across the two sequences.

Warren & Ackroff (1976) used a half-cycle offset between the two sequences of the same stimulus simply to avoid the formation of a single, centrally located, percept. Listening “set” may also influence cross-ear re-grouping. For example, distributing one’s attention across both sequences might plausibly increase cross-ear VTs and focussing one’s attention on one ear or other might plausibly decrease them. A potentially informative way of extending Warren and Ackroff’s approach would be to use a cue other than dichotic presentation to maintain the percept of two repeated sequences. Such a cue ensures that both sequences are present in both ears, increasing the possibility of across-sequence interactions. For example, one approach would be to create left- and right-lateralised sequences using ITDs. This might be expected to increase cross-sequence VTs (cf. Darwin & Hukin’s (1999) exploration of tracking by ITD vs. by F0). Another would be to introduce differences in fundamental frequency between the two sequences. Whilst this does introduce an acoustic difference between the sequences, it is specific and limited. For example, systematic pitch differences between the two sequences might be expected to reduce re-groupings involving voiced segments from both stimuli, but not, e.g., re-groupings involving a voiced segment from one stimulus and a voiceless fricative from the other.

The VTE also provides a potentially informative approach to investigating further the role of pitch contours and of formant transitions in binding together the speech stream. For example, the role of formant transitions in the VTE can be guided by the findings of Cole and Scott (1973) and Dorman et al. (1975), who concluded that formant transitions play an important role in holding the disparate speech segments into a single sequential stream. The findings of these studies might be replicated and extended using a wider range of stimuli under more controlled experimental conditions. Using careful digital editing, the VTE can be tested using words with intact or with spliced out formant transitions. It is possible that removing transitions that do not appreciably affect the intelligibility of isolated words may affect re-grouping when the word is repeated, with consequent changes in the frequency and type of VTs.

Despite researchers agreeing that speech perception is governed by both general and speech-specific auditory grouping factors, the precise nature of this influence is not known (Darwin, 2008). As speech is highly redundant, under even the most favourable listening conditions the cues available for successful perception are more than required. Therefore, it is important to identify and characterise how this information benefits the auditory system to comprehend speech when in competition with other sound sources such as noise, distortions or other speech (Darwin, 2008). The following set of experiments will address some of these issues with respect to the VTE and the grouping of speech sounds. In the process, these studies will also bear on the question of the relative contribution of grouping “primitives” (Bregman, 1990) and of speech-specific factors to the perceptual coherence of speech.

Chapter 2

Grouping and the Verbal Transformation Effect: The influence of fundamental frequency, ear of presentation, and interaural time-difference cues

2.1 Introduction

The following two experiments investigated the influence of several auditory grouping factors on the VTE when two repeating sequences of the same word were presented together. These were fundamental frequency (F0), ear of presentation, and interaural time-difference (ITD). The extent to which these cues are manipulated should affect the type and pattern of verbal transformations elicited through the general procedure in the VTE paradigm.

Warren and Ackroff (1976) investigated the effect of stimulating both ears with the same repeating stimuli while preventing the fusion of the two word tokens by offsetting their relative timing by half of their duration (see Introduction). The first experiment replicated and extended their findings by adding F0 and ITD cues to the existing conditions, as well as manipulating the ear of presentation of the two words. In Warren and Ackroff's experimental design, the two sequences of words are in competition with each other for listeners' attention; hence, by adding the F0 and ITD cues, the relative contribution of both factors can be investigated with respect to the perceptual separation of simultaneously occurring speech stimuli. The change from dichotic presentation to stimulus arrangements in which both sequences can interact within the same ear encourages competition between different perceptual organisations (VTs). Hence, this approach may offer an effective means of identifying and characterizing the grouping factors (primitive and speech-specific) of key importance to speech perception in complex listening environments. In the second experiment, the effect of pitch differences demonstrated in the first study was further explored from the perspective of the experimental task demands. While considering the role of fundamental frequency in the perceptual separation of simultaneous voices, a number of studies have demonstrated that the intelligibility of speech in the presence of interfering speech can be improved by introducing a difference in fundamental frequency (ΔF0) between the competing messages (Brokx and Nooteboom, 1982; Bird and Darwin, 1998; Assmann, 1999).

When sound travels towards us, the differences in arrival time between the two ears are referred to as interaural time differences (ITDs). It has been shown that even though ITD is a weak simultaneous grouping cue (e.g., Shackleton and Meddis, 1992), it is quite effective as a sequential grouping cue. ITD cues allow the listener to lock on to a location and to track a sound at that location over time (Darwin, 1997). Evidence of this in speech perception from a study by Darwin and Hukin (1999) indicates that listeners can use differences in ITD much more effectively than differences in F0 when tracking a speaker over time, at least for ΔF0s of a few semitones. Both F0 differences and ITD cues could potentially be used to extend the VTE paradigm. If applied to the modified condition of Warren and Ackroff's study, such that the two recycled streams are on different pitches and additionally separated spatially by ITD cues, this could improve the segregation of the two streams. This would allow exploration of the circumstances in which both sequences are present in both ears but separated using either the pitch or ITD cue. In this respect, the addition of an ITD cue could potentially improve the segregation of the two streams, if the F0 difference on its own is not sufficient. In relation to this notion, Darwin (2008) notes that: "when the listener has some independent way of grouping together the frequency components that make up different sound sources, then ITD differences between the sources give improved identification" (p. 7). Adding in an ITD cue would create the sense of lateralisation. Each ear would still be receiving both signals, with the difference that the streaming would be cued by the difference in F0, and by the ITD as well. This is important as some parts of the speech which are not voiced (e.g., fricatives) could receive an additional benefit from being differentiated by an ITD cue. As such, this would extend Warren and Ackroff's study by adding a condition where the only cues for segregation would be F0 and ITD. In addition, the ITD cues could inform us about the frequency and pattern of verbal transformations across conditions, for example whether streaming of plosive sounds is affected by the extent of ITD.

2.2 General method

The experiments presented in this thesis share some common procedures and these are outlined below. Any differences, especially the creation of stimulus sets and their manipulation are described in their respective sections.

2.2.1 Overview

All six experiments used the same behavioural measure to elicit participants' responses, which was a modified protocol of the early Verbal Transformation Effect studies (e.g., Warren, 1961a). In any given experiment, each listener was presented with a number of 3-minute presentations which consisted of continuously repeated tokens of digitally modified natural speech (either a ~0.5 s word [Experiments 1-4] or a sequence of vowels of around 330 ms [Experiments 5 and 6]). The 3-minute duration for each sequence was in accordance with Pitt and Shoaf (2002), who observed that after that time participants tend to stop reporting changes due to fatigue.

Between every 3-minute presentation there was 1-minute break during which participants were not exposed to any sound. In any one complete session there were always six 3-minute presentations; therefore, one session lasted ~30 minutes and each session was taken on a different day. All experiments employed a within-subjects design. Stimuli in all six experiments were presented at approximately 75 dB SPL.

2.2.2 Instructions

Participants were told that they would hear a series of words (Experiments 1 to 4) or speech sounds (Experiments 5 and 6) played repeatedly over headphones. At the onset of each presentation, listeners were required to speak into the microphone – positioned ~18 inches away – what they heard (whether it be a word, non-word, phrase, sentence, or syllable). Subsequently, their task was to report any changes occurring to the initial percept, this being a change to a similar word, pseudoword, nonword, syllable or to a different word altogether. Listeners were told they might also hear the current percept revert to a previous form which they would also need to report. It was emphasised that a non-response was as important as a response so that listeners did not feel under pressure to report if they did not hear a change. Listeners were assured that there was no right or wrong answer to the presented stimulus. In addition, in some experiments, listeners also used a keyboard to indicate the sequence for which the change occurred, e.g. on the high or low pitch (Experiments 1 and 2), in the right or left ear (Experiment 6), or on the higher or lower voice timbre (Experiment 5 and 6).

2.2.3 Apparatus and recording procedure

All experiments were completed in a single-walled sound-attenuating chamber (Industrial Acoustics 401A) which was housed within a quiet room. Participants' verbal responses for each 3-minute presentation were saved on a PC computer as 8-bit audio (.wav) files at a sampling rate of 11.03 kHz. The keyboard presses indicating which pitch, location or voice the change has occurred on, were stored as text (.txt) documents, where each response entry consisted of the timings of when the key button was pressed down and released and the identity of the key pressed (e.g. UP or DOWN).

On any 3-minute trial, the presentation of a stimulus over headphones, the recording of the VTs over the microphone and the recording of the key presses, were time-locked; all started simultaneously. It was therefore possible to accurately assign verbal responses to individual key presses. For example, for a text file entry of "14.028, DOWN Pressed – 14.852, DOWN Released", the experimenter would search and transcribe the respective audio file for a verbal response occurring between 14.028 s and 14.852 s (e.g., "flane"). It would then be recorded that for that particular instance, the verbal transformation "flane" occurred on the low pitch 14.028 s after the start of the trial.

Stimuli were presented using Sennheiser HD480-13II headphones at ~75 dB SPL; the headphones were calibrated using a sound-level meter (Brüel and Kjaer, type 2209) coupled to the earphones by an artificial ear (type 4153). The presentation software, custom written in VB.Net (Microsoft Visual Studio 2005), was run on a PC computer with a Turtle Beach Santa Cruz sound card. Each 3-minute presentation began with the presentation volume being ramped up from zero and at the end it was ramped down to zero. The duration of the ramps depended on the length of the stimulus itself (it used one full cycle, either the first or last to be played). This is common practice in the VTE literature; for example, Warren (1961b) increased the volume of his stimuli from 0 to full in one second.

2.2.4 Stimuli

The stimuli used in the first four experiments were monosyllabic words and for the last two experiments they were short sequences of vowels. All stimuli used in the experiments reported in this thesis were 16-bit audio files, derived from 16-bit recordings.

Monosyllabic words were used, because an increase in the number of phonetic elements in a stimulus tends to restrict the number of verbal transformations evoked (see Warren, 1961a). After initial recording, all stimuli were monotonised to the required fundamental (F0) frequency (for details, see each experiment). This technique allowed large F0 frequency separations to be introduced between the two sequences, so that auditory streaming would take place with ease (see Brokx and Nooteboom, 1982). This was particularly important in Experiments 1 and 2 where two sounds were played simultaneously. Monotonisation also precludes the possibility of pitch cross-over effects in cases where the F0 frequencies of two concurrent items were close.

In line with previous experiments, most notably those by Warren and Ackroff (1976) and Pitt and Shoaf (2002), there were no silent intervals between concurrent cycles of the stimuli. This allowed maximal re-segmentation or perceptual regroupings of phonetic elements within the presented stimuli.[2]

2.2.5 Participants

All listeners were native speakers of English and reported no hearing problems. They received either cash or course credit for participation (the vast majority were Aston University Psychology undergraduates).

[2] Interestingly, Warren, Healy and Chalikia (1996) found that, for repeating sequences of vowels, listeners reported similar (or identical) syllables either with or without silent gaps between the two iterations of the six 70-ms vowel sequence.

Around 10% of listeners across all experiments (2, 2, 1, 0, 2 and 1 for Experiments 1-6, respectively) showed little or no tendency to transform (less than 10 responses in a single session with six 3-minute presentations). On the basis that the VTE cannot be used as a tool to explore the perceptual regrouping segments in these listeners, their data were excluded and they were substituted with different listeners. It is important to note here that, although the experimenter's encouragement to report *any* perceived changes in verbal form might increase the total number of responses, it is highly unlikely that it would account for any differences observed across conditions.

2.2.6 Statistical analyses

The principal form of analysis was within-subjects ANOVA, using SPSS. All post-hoc analyses were performed using Fisher's LSD (Least Significant Difference) tests, with the restriction that the factor being explored must be associated with a significant main effect in the ANOVA (the restricted LSD test; Snedecor and Cochran, 1967; Keppel, 1991). The measure of effect size reported in the following ANOVAs was partial eta squared (η^2).

2.3 Experiment 1

Pitt and Shoaf (2002) showed that perceptual regrouping is one of the potentially many causes of the VTE where, on repetition, certain phonetic elements such as fricatives have a tendency to segregate from the others (see Introduction). Warren and Ackroff (1976) used separation of the two sequences by ear as a lateralization cue. However, it is possible to perceive clearly two sequences of words at the same time without dichotic presentation by distinguishing the two repeating tokens using two different fundamental frequencies. These words would come from the same location and they would be derived from the same original recording of speech, but they would be separated by the difference in pitch. Just as for Warren and Ackroff's procedure, in the present study the two words would be presented half a cycle out from each other in order to prevent across-ear fusion. In Experiment 1, over conditions which include differences in F0 and ITD between the two sequences, we would expect to find different frequencies and patterns of VTs. Specifically, it is hypothesised that conditions where the two sequences can interact within each ear will result in more re-grouping opportunities between the acoustic elements comprising the perceived words. Hence, listeners will report more VTs and forms in conditions with two sequences in each ear rather than just one.

2.3.1 Method

Participants

Twelve listeners (2 males, 10 females) took part in the experiment. They were all native speakers of English and reported normal hearing. At the end of the study they were either paid cash or received course credit. The mean age of the listeners was 22.2 years old (*s.d.* = 5.31).

Stimuli and Conditions

A modified version of Warren and Ackroff's (1976) experimental design was used (see Introduction). Two versions of the same word derived from the same recording were played with a half-cycle offset (half the duration of a given stimulus word), thus preventing diotic fusion of the two stimuli in instances where they are physically identical. The two versions had the same duration but differed in that they were re-synthesized on two F0 frequencies with a 10-semitone difference. In addition, three lateralization cue conditions were introduced. The no-ITD (diotic) condition resulted in the perception of both sounds coming from the central azimuthal position. For this condition, the only separation cue was the difference in pitch between the recycled words. The second lateralization condition, 680-µs ITD, resembled a maximum natural ITD difference for a typical adult male of about 680 µs. This arrangement meant that, in both the no-ITD and 680-µs ITD conditions, the two sequences were physically present in the same ear. This allowed for perceptual regroupings across- as well as within-sequence, and could potentially have an effect on the number and type of VTs heard by listeners. The final condition used was dichotic presentation. The last condition resembled that of Warren and Ackroff (1976), except for the pitch difference between the two sequences.

Six monosyllabic words were used – **face**, **right**, **sleep**, **see**, **noise**, and **flame** – all spoken by the same male voice with no obvious regional accent. The selected words come from previous VTE studies, and were chosen on the basis that they produce a variety of verbal transformations as determined by a pilot study. The duration of 550 ms for each word was also decided on that basis (resulting in 327 repetitions in 3 min). The speaker produced several examples of each utterance with the assistance of an on-screen metronome to help pace speech production. From that recording session, instances with clear articulation and which were very close to the desired duration were chosen. Using CoolEdit software, exact durations were achieved by small manual adjustments to the stimuli; for example, copying in or deleting a few ms of fricative noise or plosive silence. Next, amplitude contours of every 550-ms file were adjusted such that the start and end were ramped up and down (5-ms ramps) using CoolEdit. All stimuli were MONO, 16-bit recordings with a 22.05 kHz sampling rate and duration of 550 ms. These duration-adjusted and ramped recording were then processed using PRAAT software (Boersma and Weenink, 2009) as follows:

(i) monotonized, using PSOLA – a time domain speech manipulation algorithm which identifies glottal pulses and aligns them equally in time (Moulines and Charpentier, 1990).

(ii) LPC (Atal & Hanauer, 1971) resynthesized on two different F0 frequencies (with 10 semitones difference), at 100 Hz (low pitch, male range) and at 178 Hz (high pitch, female range). LPC – linear predictive coding, allows separation of the excitation source from the filter function and after manipulating the source (e.g. F0 frequency), the modified source can be fed back through the original filter.

Finally, using MITSYN (Henke, 1997), 680-μs ITD instances of each word were created for the 680-μs ITD lateralization-cue condition. Opposite lateralizations were used for the two repeating sequences – i.e., one sequence was perceived as coming from the left ear and the other sequence as coming from the right ear. An additional word, **train**, was transformed in the same way as the experimental stimuli described above and used in the practice trial for this experiment.

After listeners read the instructions, the experimenter answered any questions and reiterated the methodology. Participants then completed a training session which comprised a 1-minute presentation of the word **train** (processed in the same way as for the 680-μs ITD condition). The main experiment comprised six 3-minute presentations with 1-minute breaks between each presentation. Each 3-minute presentation consisted of two copies (one on the low and one of the high pitch) of a given stimulus word, played half a cycle out of phase with each other.

Participants were instructed that they would hear a word or words spoken by two voices, one on a low pitch and one on a high pitch. They were asked to monitor both voices continuously, to speak into the microphone as soon as they were able to identify what each voice appeared to be saying, and at the same time to indicate using the 'up' or 'down' arrow key on the keyboard whether what they heard was on the high or the low pitch, respectively. For example, if a listener heard the word 'book' spoken on the high pitch, they should press the 'up' arrow key (therefore displaying 'HIGH' on the screen), say the word 'book' into the microphone, and then release the button. Note that, although this procedure allows for continuous and effective monitoring of both sequences, it must be acknowledged that on any occasion when participants hear transformations almost at the same time on the two pitches they cannot in principle respond to them both simultaneously. Listeners were instructed to keep on listening to the stimuli, and to speak into the microphone each time as soon as the words seem to change, using key presses as indicated. A change was defined as either a new word or a return to a word which they had reported before. It was pointed out to participants that there were no correct or incorrect responses and that in some cases they may hear few or no changes over the course of a trial.

Each of the three sessions, each corresponding to one of the three lateralization-cue conditions, took ~30 minutes to complete and were taken on a different day. The order of the three conditions was fully counterbalanced between participants, requiring six people to complete a full set (the Experiment therefore included two sets). Within each session, trials using particular words were presented in random order.

The three within-subject conditions were lateralization cue (no-ITD, 680-µs ITD, or dichotic), pitch (high or low) and word (**noise**, **flame**, **face**, **sleep**, **see**, or **right**). For each listener, the number of verbal transformations and forms were calculated. A verbal transformation was defined as any change to the reported stimulus (this could be a change to a new form or back to one previously reported) while a new form was defined as a case where a given transformation had not occurred before on that trial. Therefore, as long as at least two forms were reported, a listener could experience an infinite number of verbal transformations. All participants' responses were included in the analysis, including non-words and pseudowords (see Appendix 1 for the types of the responses given for each word on the high and low pitches).

2.3.2 Results and discussion

The results presented below reflect four major aspects of the data analysis, (i) number of verbal transformations reported, (ii) number of forms reported, (iii) timing of the first verbal transformation, and (iv) the dependency index as a measure of the extent to which VTs observed for one sequence are related to those observed for the other (described below in a separate section).

Verbal Transformations

Table 2.1 shows the mean numbers of verbal transformations reported in 3 minutes for each listener. The grand average reported across all conditions in 3 minutes was 13.17 verbal transformations.

Table 2.1 Average no. of VTs for each lateralization cue and stimulus word across all listeners. The breakdown between high pitch responses (H) and low pitch responses (L) is given in brackets. Standard errors of the mean are in italics.

	Verbal Transformations reported (in 3 min)
no-ITD	**14.04** *±2.88* (H=7.57 *±1.51*, L=6.47 *±1.61*)
680 µs ITD	**15.24** *±3.49* (H=9.92 *±2.38*, L=5.32 *±1.20*)
Dichotic	**10.24** *±2.58* (H=6.61 *±1.66*, L=3.63 *±1.05*)
Noise	**8.86** *±2.64* (H=5.86 *±1.85*, L=3.00 *±0.95*)
Flame	**17.00** *±4.30* (H=10.81 *±3.03*, L=6.19 *±1.94*)
Face	**10.47** *±2.23* (H=5.75 *±1.33*, L=4.72 *±1.28*)
Sleep	**15.92** *±3.30* (H=9.31 *±2.04*, L=6.61 *±1.54*)
See	**17.64** *±3.80* (H=11.08 *±2.92*, L=6.56 *±1.29*)
Right	**9.14** *±2.46* (H=5.39 *±1.67*, L=3.75 *±1.17*)

A three-way repeated-measures ANOVA was performed on the data. The within subjects factors were lateralization cue (no-ITD, 680-μs ITD or dichotic), pitch (high or low), and words (‘noise’, ‘flame’, ‘face’, ‘sleep’, ‘see’, or ‘right’). The ANOVA results are summarized in Table 2.2. Significant terms are shown in bold.

Table 2.2 Summary of three-way ANOVA for verbal transformations.

Source	*df*	F	p	η^2
Lateralization Cue (L)	2,22	2.16	=.14	=.16
Pitch (P)	**1,11**	**14.23**	**<.01****	**=.56**
Word (W)	**5,55**	**4.16**	**<.01****	**=.27**
L x P	2,22	2.26	=.13	=.17
L x W	10,110	0.85	=.59	=.07
P x W	5,55	0.61	=.69	=.05
L x P x W	**10,110**	**2.21**	**<.05***	**=.17**

The main effect of pitch indicates that listeners reported verbal transformations more often on the high pitch (8.03 VT/3 min, s.e.= ±1.60) than on the low pitch (5.14 VT/3 min, s.e.= ±1.09). Fisher’s LSD post-hoc analysis for the main effect of word showed that ‘noise’ (8.86 VT/3 min) transformed significantly less than did ‘flame’ (p<.01), ‘sleep’ (p<.05), and ‘see’ (p<.01). In addition, ‘flame’ transformed more than ‘face’ (p<.05) and ‘right’ (p<.01), ‘face’ transformed less than ‘see’ (p<.05), ‘sleep’ transformed more than ‘right’ (p<.05) and ‘see’ transformed more than ‘right’ (p<.01). No other pairwise comparisons were significant. The effect of the lateralization cue is apparent only in the context of the significant three-way interaction term (though there is perhaps a suggestion of a trend for the main effect and for the L x P interaction term). Inspection of the Figure 2.1 reveals that the three-way interaction comes from significantly higher transformation rates for the words ‘flame’ and ‘see’ on the high pitch in the 680-μs ITD condition.

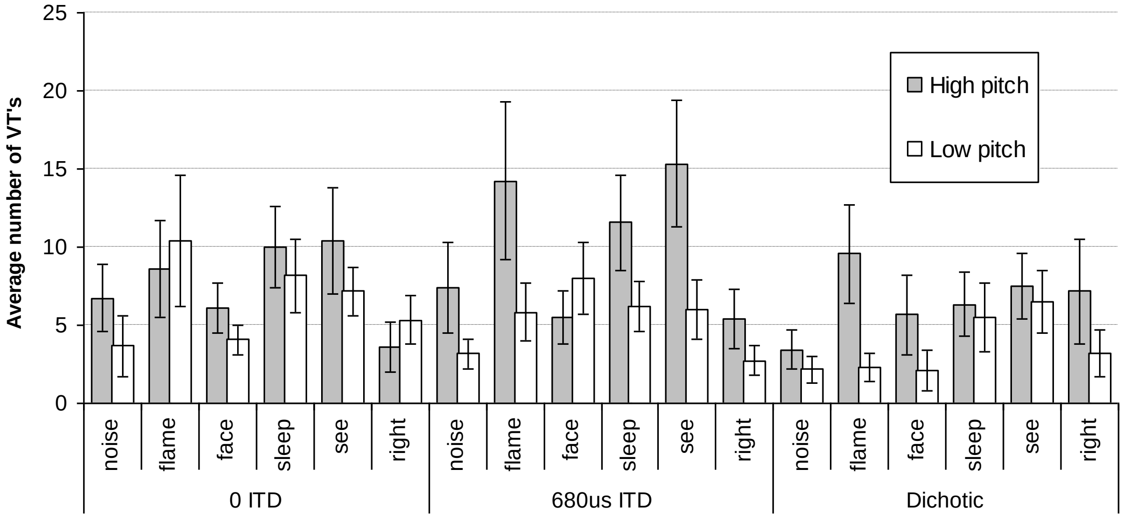


Figure 2.1 Three-way interaction for verbal transformations.

Forms

Just as for verbal transformations, a 3x2x6 ANOVA (lateralization cue, pitch, and word) was performed on the number of forms reported. Table 2.3 shows the means for the three conditions. The grand average reported across all conditions in 3 minutes was 3.52 forms.

The ANOVA results showed significant main effects for all three factors; there was also a significant two-way interaction between lateralization cue and word and a significant three-way interaction. Table 2.4 presents a summary of that analysis.

For the main effect of pitch, responses on the high pitch (1.90 Forms/3 min, s.e.= ±0.30) were more frequent than on the low pitch (1.62 Forms/3 min, s.e.= ±0.29). Fisher's LSD post-hoc analysis for the main effect of the lateralization cue revealed that significantly fewer forms were reported for the dichotic condition than for no-ITD (p<.01) and 680-µs ITD (p<.01) conditions. The no-ITD (diotic) and 680-µs ITD conditions did not differ from one another (p>.7). This pattern is consistent with the (non-significant) differences in the number of VTs reported across these conditions. For the main effect of word, 'noise' had significantly fewer forms than 'face' (p=.03), 'sleep' (p=.02), and 'see' (p=.03), and 'right' had significantly fewer forms than 'flame' (p=.04), 'face' (p<.01), 'sleep' (p<.01), and 'see' (p=.01).

Table 2.3 Average no. of Forms for each lateralization cue and stimulus word across all listeners. The breakdown between high pitch responses (H) and low pitch responses (L) is given in brackets. Standard errors of the mean are in italics.

	Different forms (in 3 min)
no-ITD	**4.04** *±0.62* (H=2.03 *±0.31*, L=2.01 *±0.35*)
680 µs ITD	**3.90** *±0.70* (H=2.15 *±0.40*, L=1.75 *±0.32*)
Dichotic	**2.63** *±0.54* (H=1.53 *±0.28*, L=1.10 *±0.29*)
Noise	**2.08** *±0.55* (H=1.19 *±0.32*, L=0.89 *±0.25*)
Flame	**3.78** *±0.80* (H=2.14 *±0.47*, L=1.63 *±0.37*)
Face	**4.33** *±0.81* (H=2.47 *±0.46*, L=1.86 *±0.43*)
Sleep	**4.50** *±0.92* (H=2.36 *±0.47*, L=2.14 *±0.47*)
See	**4.31** *±0.82* (H=2.25 *±0.44*, L=2.06 *±0.40*)
Right	**2.14** *±0.48* (H=1.00 *±0.22*, L=1.14 *±0.31*)

Table 2.4 Summary of three-way ANOVA for forms.

Source	***df***	**F**	**p**	**η^2**
Lateralization Cue (L)	**2,22**	**7.72**	**<.01****	**=.41**
Pitch (P)	**1,11**	**5.36**	**<.05***	**=.33**
Word (W)	**5,55**	**4.68**	**<.01****	**=.30**
L x P	2,22	1.39	=.27	=.11
L x W	**10,110**	**2.47**	**=.01****	**=.18**
P x W	5,55	1.20	=.32	=.10
L x P x W	**10,110**	**2.20**	**<.05***	**=.17**

The significant lateralization cue x word interaction mainly reflects the fact that, for stimulus words ‘face’ and ‘sleep’, new forms were reported significantly more often for the no-ITD and 680-µs ITD conditions than for the dichotic case. This was confirmed by LSD post hoc tests, where the number of forms for ‘face’ and ‘sleep’ did not differ for conditions no-ITD and 680-µs ITD (p>.7 for ‘face’ and p>.2 for ‘sleep’ respectively), however for the dichotic case both words had significantly fewer forms than either the no-ITD or 680-μs ITD conditions (p<.01 for all four comparisons). For an illustration of this, see Figure 2.2. This pattern suggests that the impact on forms of whether or not the two sequences can interact within the same ear depends on the acoustic properties of individual stimulus words.

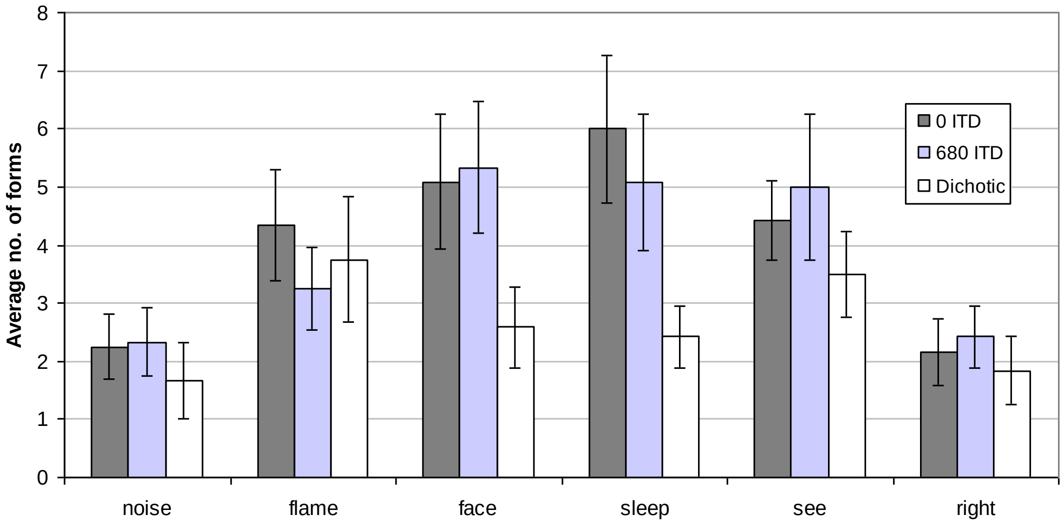


Figure 2.2 Lateralization cue x word interaction for forms.

An inspection of Figure 2.3 shows that the significant three-way interaction is mainly driven by a greater number of forms reported for the stimulus words: ‘face’ in the no-ITD condition on the high pitch, ‘sleep’ in the no-ITD condition on the low pitch, and ‘sleep’ in the 680-μs ITD condition on the high pitch.

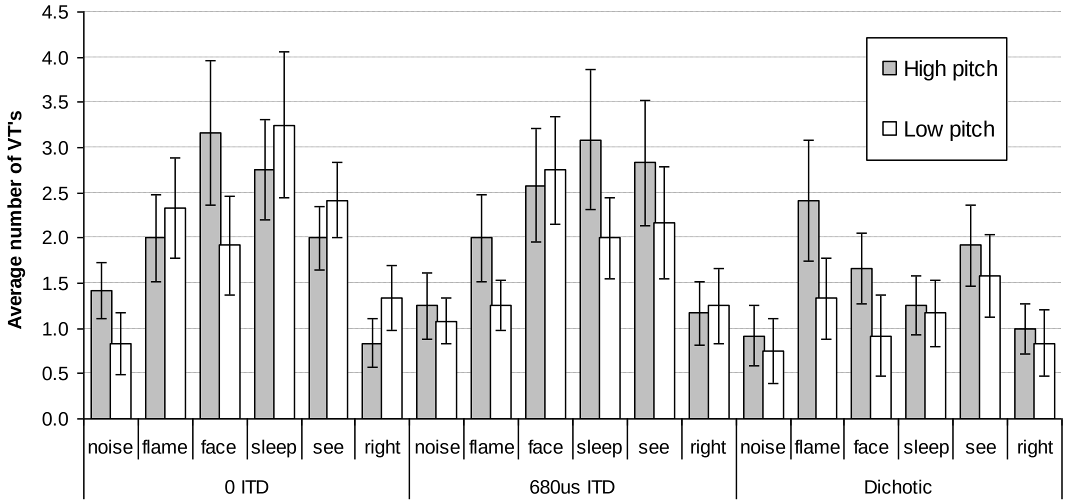


Figure 2.3 Three-way interaction for forms.

Timing of the first verbal transformation

For the following analysis, two values were extracted from any 3-minute presentation: the timing of the first VT on the high pitch and the timing of the first VT on the low pitch. Note that a nil response within any 3-minute presentation was marked as 180 s (maximum time within which a transformation could occur). Hence, the mean response time might appear spuriously late in cases where there were a substantial number of trials with nil responses (see Table 2.5). Hence, a separate analysis is also presented, for which nil responses were excluded from the data.

Table 2.5 Average times of the first VT for each lateralization cue and word on the two pitches. Standard errors of the mean are in italics.

	Time of first VT (in seconds)	
	High pitch	**Low pitch**
no-ITD	65.35 *(11.73)*	70.44 *(9.26)*
680 µs ITD	54.33 *(10.13)*	69.85 *(10.14)*
Dichotic	82.53 *(10.12)*	111.50 *(9.13)*
Noise	86.22 *(11.41)*	104.89 *(11.72)*
Flame	59.75 *(10.49)*	80.42 *(11.53)*
Face	60.94 *(10.60)*	87.11 *(12.40)*
Sleep	52.83 *(10.76)*	61.61 *(10.90)*
See	43.67 *(8.28)*	56.81 *(10.62)*
Right	101.00 *(11.56)*	112.75 *(11.79)*

The data were analysed using a three-way 3x2x6 ANOVA. The within subjects factors were lateralization cue (no-ITD, 680-µs ITD or dichotic), pitch (high or low), and words (‘noise’, ‘flame’, ‘face’, ‘sleep’, ‘see’, or ‘right’). A summary of this analysis is presented in Table 2.6.

Table 2.6 Three-way ANOVA for the timing of the first VT (pitch separated).

Source	***df***	**F**	**P**	**η²**
Lateralization Cue (L)	**2,22**	**14.19**	**<.01****	**=.56**
Pitch (P)	**1,11**	**6.45**	**=.03***	**=.37**
Word (W)	**5,55**	**4.32**	**<.01****	**=.28**
L x P	2,22	1.76	=.20	=.14
L x W	10,110	0.67	=.75	=.06
P x W	5,55	0.24	=.96	=.02
L x P x W	10,110	1.31	=.24	=.11

The average timings for the lateralization cue and words on the two pitches are shown in Table 2.5. Note that all three main factors were significant; none of the interaction terms were significant. Post hoc analyses revealed for the main effect of the lateralization cue that the first VT for the dichotic case (97.01 s) occurred significantly later than for the no-ITD (67.90 s, p<.01) and 680-µs ITD cases (62.09 s, p<.01). For the main effect of pitch, listeners tended to report their first VT on the low pitch (83.93 s) later than on the high pitch (67.40 s). Finally, for the main effect of word, the pairs showing significant differences were 'sleep' (57.22 s) vs. 'right' (106.88 s, p=.02), and 'see' (50.24 s) vs. 'right' (106.88 s, p=.01).

When nil-response cases were excluded from the analysis (i.e., treated as missing values), a two-way 3 (lateralization cue) x 2 (pitch) ANOVA still revealed a significant main effect for the lateralization cue in the same direction [F(2,22)=6.60, p<.01]. As for the previous analyses, the average first VT for the dichotic condition (54.27 s) occurred significantly later than for the no-ITD (37.80 s; p=.02) and for the 680-µs ITD conditions (37.44 s; p=.01). No other effects were significant. Hence, it can be stated with confidence that the apparent tendency for later first responses to occur for the higher-pitched sequences and for the dichotic condition is not an artefact of changes in the number of nil responses.

The average percentages of trials with a nil response in a 3-minute presentation for each lateralization cue and pitch are presented in the Table 2.7.

Table 2.7 Average % of nil responses in 3-min.

	High pitch	**Low pitch**
no-ITD	15.28 %	26.39 %
680 µs ITD	16.67 %	34.72 %
Dichotic	22.22 %	33.33 %

Pitch integrated

To pursue the above analysis further, in any 3-minute presentation, the timing of the *very first* verbal transformation reported was used, irrespective of whether it occurred for the high or the low pitch sequence.

Table 2.8 Average times of the first VT for each lateralization cue and word (pitch integrated). Standard errors of the mean are shown in italics.

	Time of first VT (in seconds)
no-ITD	45.68 *(8.90)*
680 µs ITD	43.04 *(9.83)*
Dichotic	70.15 *(10.37)*
Noise	73.64 *(16.27)*
Flame	47.28 *(10.56)*
Face	39.78 *(8.20)*
Sleep	39.86 *(11.54)*
See	31.97 *(12.87)*
Right	85.22 *(13.06)*

A two-way within-subjects 3 x 6 ANOVA (lateralization cue x word) again revealed a significant main effect of the lateralization cue [F(2,22)=7.83, p<.01]. Fisher's LSD post-hoc tests show that the average time of the first verbal transformation for the dichotic condition (70.15 s) occurred significantly later than for both the no-ITD condition (45.68 s, p=.01) and the 680-µs ITD case (43.04 s, p<.01). For the main effect of word [F(5,55)=4.97, p<.01], a significant difference was found for the following word pairs, 'noise' vs. 'see' (p<.05), 'face' vs. 'right' (p=.02), 'sleep' vs. 'right' (p=.02), and 'see' vs. 'right' (p<.01). These pairwise differences seem to reflect the effect of particular phonetic segments and the likelihood of them 'cleaving off' perceptually from the rest of the stimulus word. Specifically, the voiceless fricatives 'f' and 's' show a greater tendency for stream segregation than do the voiced approximant 'r' or the voiced nasal 'n'. For the average times of the first VT, refer to Table 2.8.

As indicated previously, the above averages are affected by the fact that a nil response within any 3-minute presentation was coded as 180 s, which can create an impression that the responses are spuriously late. However, this ensured that an average time of a first VT for a listener who gave few responses did not appear earlier than an average time for a listener who provided more (but slower) responses. When nil-response cases were excluded from the pitch-integrated analysis altogether, the one-way ANOVA for lateralization cue [F(2,22)=3.87, p<.05] revealed a similar pattern of results as for the previous test, indicating that this outcome was not due to 180 s substituting for an empty data cell. The average time of the first transformation for the dichotic case (44.63 s) was significantly later than for no-ITD condition (29.01 s; p=.02) and for 680-µs ITD (31.15 s; p<.05), with no difference between the latter two conditions. The average percentage of trials with no response in a 3 minute presentation for each lateralization cue was: 11.11% for no-ITD, 9.72% for 680-μs ITD, and 19.44% for the dichotic case. This pattern is consistent with the observed tendency for first responses to occur later in the dichotic condition.

Difference between reports for right and left ears (VTs and forms)

In their dichotic study, Warren and Ackroff (1976) reported no significant quantitative differences between responses to the left and right ears for VTs and forms. These results have been replicated in the current experiment. Such an analysis was possible in the dichotic and 680-μs ITD conditions as the two pitch percepts, high and low, corresponded to the left and the right ear (or side or space), respectively. In both the 680-μs ITD and dichotic conditions, the presentation of high and low pitch stimuli was counterbalanced such that half of the listeners had high-pitch words presented to their left ear and for the other half, high-pitch words were presented to their right ear. Note that listeners were not explicitly instructed to associate particular pitches with particular locations; rather, their task was to focus on the voice pitches.

Similar to Warren and Ackroff's (1976) finding, in the current experiment there were no significant effects of differences in the direction of lateralization (by ITD cues or by ear) for either of the conditions. For VTs in the 680-μs ITD condition, listeners reported an average of 6.64 VTs/3 min as coming from the left ear and 8.60 VTs/3 min as coming from the right ear. For the dichotic condition, these differences were as follows: 5.32 VTs/3 min in the left ear versus 4.92 VTs/3 min in the right ear. The 3x2 ANOVA (lateralization cue x direction of lateralization) revealed no significant effects. Specifically, for the main effect of lateralization F(2,22)=2.16, p>.1; for the main effect of direction of lateralization F(1,11)=.21, p>.6; and for the two-way interaction F(2,22)=1.05, p>.3.

The results for the number of VTs were mirrored by those for forms – i.e., no effect of direction of lateralization was found. For the 680-μs ITD condition, listeners reported 1.68 Forms/3 min as coming from the left ear as opposed to 2.04 Forms/3 min as heard in the right ear. For the dichotic condition, the corresponding means were: 1.39 Forms/3 min (left ear) and 1.24 Forms/3 min (right ear). The 3x2 ANOVA for forms did not yield significant effects for the direction of lateralization. Specifically, for the main effect of lateralization cue, F(2,22)=7.72, p=.003 (see earlier section on Forms describing this result), for the main effect of direction of lateralization, F(1,11)=.17, p>.6, and for the two-way interaction F(2,22)=1.10, p>.3.

Dependency index measure

Warren and Ackroff (1976) suggested that dichotic VTs occur independently of one another, and so provide evidence of separate linguistic processors for identical stimuli. Although Warren and Ackroff claimed that listeners experienced independent VTs in the two ears, they did not specify precisely how their measure of independence was obtained. The explanation given by Warren and Ackroff was as follows: "The major finding was that VTs occurred independently on each side. Each of the 20 subjects listening dichotically and monitoring the identical stimuli on both sides reported hearing phonetically different words on the two sides at the same time for some period during the test." (p.476). Although, on any 3 minute trial in the current experiment, such instances were also found by visual inspection of the data, the issue of independence was explored and quantified in a more systematic way. This is important because the near-simultaneous occurrence of different VT forms to two repeating sequences of the same stimuli does not necessarily imply their independence.

The Dependency Index measure was used to quantify the relationship between the responses for the two presented streams of repeated words, one on the high and one on the low pitch. This custom measure of relatedness had the advantage over existing correlational measures of allowing each response on a given sequence to be compared with both adjacent responses on a second sequence – the one immediately preceding it and the one immediately following it.

Two measures were used to assess the relatedness of responses to the high- and low-pitched sequences. The main measure, the *dependency index*, compared each response to one sequence with both the previous and the subsequent response to the other, flagging each decision as 1 (hit) if the preceding/following response to the other sequence was the same and otherwise flagging it as 0 (miss)[3]. The dependency index then is the total number of hits divided by the total number of responses to that sequence (see Figure 2.4). Scores ranged from 0 (independent/unrelated VTs) to 1 (fully dependent/related VTs). The second measure, the *temporal overlap index*, gave the proportion of time (for the remaining interval after the

[3] Note that the task constraints prevented a listener from indicating a simultaneous change on both sequences.

first VT occurred) for which responses to both sequences were the same. The temporal overlap index becomes important when differences between conditions are found for the dependency index. This is because, in principle, there is a circumstance in which a reduction in the value of the dependency index could occur without a change in the degree to which responses to the two sequences are related. Specifically, this is where there is an increase in the proportion of anti-correlated responses to the two sequences. Such a change would, however, be reflected in the temporal overlap index. Hence, a substantial reduction in the dependency index accompanied by a relative lack of change in the temporal overlap index would indicate that the responses are indeed more independent of one another, and not simply an artefact of a change from correlations to anticorrelations.

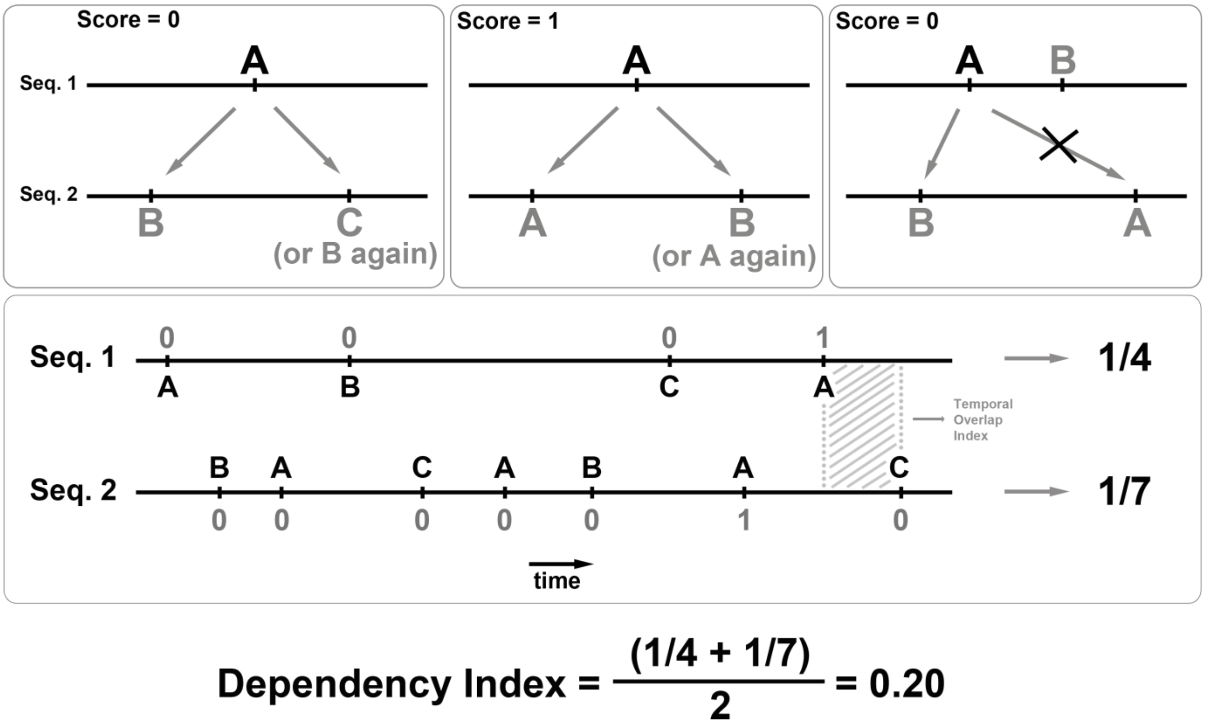


Figure 2.4 Calculation of the Dependency Index measure. Top three panels: each response (A) on one sequence is compared with both the previous and subsequent response on the other sequence. Response (A) is given a value of 0 (miss) if neither response on the other sequence matches A; note this can be true even when both responses on the other sequence are the same (top left panel). A value of 0 (miss) is also given if there is an intervening response on the same sequence that does not match (top right panel). Response (A) is given a value of 1 (hit) if either or both of the responses on the other sequence matches A (top middle panel). The bottom panel shows the scoring procedure applied to an example series of responses. For each sequence, the number of hits is divided by the total number of responses on that sequence. The dependency index measure is obtained by averaging the sum from both sequences. The temporal overlap index is represented by the shaded area and it shows the proportion of time for which the responses to both sequences were the same, in this case response A.

Table 2.9 shows the results for the Dependency and Temporal Overlap Indices across the lateralization-cue conditions and individual stimulus words. A two-way 3 x 6 ANOVA revealed a main effect of the lateralization cue [F(2,22)=8.03, p=.002, η^2=.42] showing more independence in the dichotic case (0.09) [no-ITD vs. dichotic, p=.03 and 680-µs ITD vs. dichotic, p=.01]. Neither the main effect of word [F(5,55)=.75, p>.5], nor the interaction [F(10,110)=.53, p>.8] were significant. In addition, the fact that the dependency index measure was low in all three conditions (with the largest being 0.26, for the 680-µs ITD condition) suggests that most VTs were found to be relatively independent for the high- and low-pitched sequences in all three conditions.

Table 2.9 Means across each Lateralization cue and Word. Standard errors of the mean are shown in italics.

	Dependency Index	**Temporal Overlap Index**
no-ITD	0.23 *(0.05)*	0.41 *(0.05)*
680 ITD	0.26 *(0.04)*	0.34 *(0.04)*
Dichotic	0.09 *(0.02)*	0.36 *(0.04)*
Noise	0.19 *(0.06)*	0.39 *(0.08)*
Flame	0.19 *(0.04)*	0.43 *(0.04)*
Face	0.17 *(0.05)*	0.25 *(0.04)*
Sleep	0.22 *(0.05)*	0.38 *(0.04)*
See	0.26 *(0.05)*	0.42 *(0.05)*
Right	0.15 *(0.06)*	0.34 *(0.06)*

For the temporal overlap index, the mean values obtained were fairly similar across conditions (no-ITD = 0.41, 680-µs ITD = 0.34, dichotic = 0.36). Indeed, the corresponding 3 x 6 ANOVA showed that there was no main effect of the lateralization cue [F(2,22)=.70, p>.5]; this outcome indicates that the lower dependency index for the dichotic case was not a spurious consequence arising from greater anticorrelation in the responses to the two sequences. In addition, neither the main effect of word [F(5,55)=1.77, p>.1] nor the interaction term [F(10,110)=.68, p>.7] were significant.

When collapsed across lateralization conditions, the results indicate a relatively low dependency index (0.19) of the responses when the two F0 are pooled together (see Table 2.10). The results are also shown separately here for the cases where the low-pitch or the high-pitch sequence was designated as the reference sequence to which the other was compared when computing the dependency index. Note that the effect of which one is used as the reference case is relatively small.

Table 2.10 Overall Means for the Dependency Measure. Standard errors of the mean are shown in italics.

Low Pitch Dependency Index	High Pitch Dependency Index	Overall Dependency Index	Temporal Overlap Index
0.21 *(0.02)*	**0.18** *(0.02)*	**0.19** *(0.02)*	**0.37** *(0.02)*

2.3.3 Summary and Conclusions

As the difference in apparent lateralization difference between the two sequences increased, the number of forms reported was reduced, with the fewest number of forms heard in the dichotic condition (same trend for VTs, but not significant). Consistent with this pattern, the first verbal transformation occurred significantly later for the dichotic case than for the no-ITD or 680-µs ITD conditions.

The dependency index showed relatively low dependency, suggesting that most of the responses on the two streams were fairly independent of one another. Additionally, the responses were significantly less independent when there was no separation of the two sequences by ear (i.e., where both two sequences were presented to both ears).

There was a general tendency for responses to the high-pitched sequence to be more numerous, to display more forms, and to occur earlier than responses to the low-pitched sequence. These effects of sequence pitch (high vs. low) on verbal transformations, which were evident throughout the analyses presented above, were explored further in Experiment 2.

Overall, the results are consistent with the hypothesis that verbal transformations are facilitated by the possibility of additional re-groupings offered by conditions where two sequences are present in each ear (no-ITD and 680 ITD). The two sequences can interact with each other across and within ears, creating more opportunities for VTs to occur than in the dichotic condition, where one sequence only is present in the right ear and one in the left ear.

2.4 Experiment 2

Experiment 1 revealed that the high-pitched sequence was associated with significantly more VTs and forms, and with a significantly shorter time to first VT. In addition, there was a trend towards a location cue x pitch interaction for all three measures. Although the interaction itself was not significant (see Tables 2.2, 2.4, and 2.6) there is some evidence of a tendency for the 680-µs ITD and the dichotic conditions listeners to be associated with more transformations, forms, and shorter first-response times on the high pitch rather than the low pitch (see Tables 2.1, 2.3, and 2.5). For example, in the case of number of VTs, the tendency to respond to the high pitch increased substantially for 680-µs ITD and dichotic conditions compared to the no-ITD case. While in the no-ITD condition there were about 15% fewer responses on the low pitch, these ‘losses’ grew to 46% for the 680-µs ITD condition and to 45% for the dichotic condition (see Table 2.11).

Table 2.11 Average number of VTs in 3 min across location cue and pitch for Experiment 1.

Condition	no-ITD		680 ITD		Dichotic	
Pitch	High	Low	High	Low	High	Low
Average	7.57	6.47	9.92	5.32	6.61	3.63

Two possible explanations for these differences suggested either qualitative differences between the high and low pitch stimuli or variable task demands between the conditions of Experiment 1. Although this was not evident from the participants’ comments after finishing the study, they could have experienced the high-pitch stimuli as more phonetically salient, e.g. as sounding clearer or louder (even though the experimental manipulation of all stimulus words was uniform). Alternatively, the demand characteristics for listeners attending two sequences at once could have made the task more difficult, such that they failed to report all the VTs they might have heard (especially on the low pitch). It remains unclear the extent to which attention is required for the build-up of stream segregation (Carlyon, Cusack, Foxton and Robertson, 2001). Nonetheless it is clear at least for tone sequences that switching attention between streams can reset the build-up of stream segregation (Cusack, Deeks, Aikman and Carlyon, 2004). In Experiment 1 listeners could have been switching their attention between the two sequences they were monitoring causing a reset of the build-up of stream segregation which resulted in fewer VTs reported. In other words, one way in which task demands might influence the rate of VTs is because there is too much to monitor and report. However, it is also possible that the task demands have the direct effect on perception through attentional switching (Cusack et al., 2004). On either ground, we might expect fewer responses. The task demands may have been particularly high when these sequences were heard as spatially separated (either by ITD cues or different ear). This possibility was investigated in Experiment 2; it was hypothesised that the difference between the reports of VTs on the high- and low-pitched sequences will be absent with lower task demands (i.e., when listeners are asked only to attend to one sequence at a time).

2.4.1 Method

Participants

Fifteen participants (three sets of five rotations of the conditions used) completed the study. None of them took part in Experiment 1. They were all native speakers of English and reported normal hearing. At the end of the study they were either paid cash or received course credit. The mean age of the listeners was 22.5 years old (s.d. = 3.02). There were 12 females and 3 males.

Stimuli and Conditions

The differences between conditions in the number of VTs reported in Experiment 1 were explored further in the current study. Each participant attended five sessions corresponding to the conditions below (all presented diotically). In conditions 1 and 2, each of the 3 minute presentations consisted of an on-going repetition of a single word. In conditions 3 to 5, there were two on-going repetitions (one on the high and one on the low pitch) of a word played half a cycle out of phase (i.e., half the duration of a stimulus word). The description of the conditions given here includes the number of sequences present (e.g., one sequence in Condition 1: ‘Low’ and two sequences in Condition 3: ‘High/Low’), and the instruction as to which sequence they have to attend to (represented by an underscore, e.g. in Condition 3 listeners are asked to report VTs on the low-pitched sequence only: 'High/Low'). Hence the summary of all the conditions is as follows:

Condition 1 (Low) – listeners were presented with a single sequence of low-pitch stimuli and asked to report all transformations

Condition 2 (High) – listeners were presented with a single sequence of high-pitch stimuli and asked to report all transformations

Condition 3 (High/Low) – listeners were presented with two sequences (high and low pitch) at the same time (one sequence delayed by half the duration of a stimulus word) and asked to report transformations on the low pitch only

Condition 4 (High/Low) – listeners were presented with two sequences (high and low pitch) at the same time (one sequence delayed with half a cycle offset) and asked to report transformations on the high pitch only

Condition 5 (High/Low) – listeners were presented with two sequences (high and low pitch) at the same time (one sequence delayed with half a cycle offset) and asked to report transformations on both pitches (for convenience, in the results analysis this condition has been split into two: one including only transformations reported on the low pitch – High/**LOW**, and the other on the high pitch – **HIGH**/Low)

The conditions were counterbalanced across listeners using a five-cycle rotation, which meant that the condition order for the first listeners in each set was 1-2-3-4-5, the second was 2-3-4-5-1, the third was 3-4-5-1-2, and so on. The stimuli used were the same as in Experiment 1. However, here the words were presented diotically in all conditions. The data were recorded and transcribed in the same way as in previous experiment.

2.4.2 Results and discussion

In order to analyse the extent of task demands on the number of VTs (and forms) the five conditions have been condensed into three, as demonstrated in Table 2.12, which shows the average numbers of VTs for each experimental condition. The three conditions were: M1 – single sequence presented and listeners asked to report what they hear (conditions Low +

High), M2 – two sequences presented but listeners asked to attend to only one of them, and report either the high pitched or the low pitched voice (conditions High/Low + High/Low), and M3 – two sequences presented and listeners asked to attend to both of them at the same time, reporting changes on both sequences (conditions High/**LOW** + **HIGH**/Low). If condition M3 places the greatest constraints on listeners' attention, it should result in the fewest VTs and forms being reported.

Table 2.12 Average no. of VTs for each location cue and stimulus word across all listeners. Standard errors are shown in brackets.

		Verbal Transformations reported (in 3min)	**Different forms (in 3min)**
M 1	**Low**	9.90 *(3.40)*	3.58 *(1.11)*
	High	10.13 *(3.21)*	3.70 *(1.02)*
M 2	**High/Low**	13.22 *(2.73)*	5.37 *(1.53)*
	High/Low	13.42 *(4.25)*	5.18 *(1.45)*
M 3	**High/LOW**	7.22 *(1.70)*	3.38 *(0.69)*
	HIGH/Low	10.40 *(2.58)*	4.17 *(0.80)*
Noise		7.55 *(2.57)*	3.38 *(1.58)*
Flame		12.75 *(5.38)*	4.38 *(1.13)*
Face		10.28 *(2.50)*	4.80 *(1.56)*
Sleep		11.03 *(3.06)*	4.63 *(1.15)*
See		12.58 *(4.46)*	4.57 *(1.17)*
Right		10.08 *(3.40)*	3.62 *(1.23)*

Similar to Experiment 1, the number of VTs and forms were the main focus of the current study. In addition, the effect of condition on time to the first VT was explored. In those conditions where listeners were expected to monitor and respond to both sequences at once, the dependency and temporal overlap indices were completed as for Experiment 1.

Verbal Transformations

The three-way ANOVA – 3 (condition: M1, M2, M3) x 2 (pitch) x 6 (word) – did not reveal any significant effects, although the main effect of condition nearly reached significance [F(2,18)=3.48, p=.053, η^2=.23]. The means for the three conditions were: M1 – 10.02 VTs/3 min, M2 – 13.32 VTs/3 min, and M3 – 8.81 VTs/3 min. This suggests a trend for the fewest number of VTs in the M3 condition, where listeners had to monitor both sequences at the same time. The fact that in condition M2, where listeners were presented with the same stimulus arrangement (i.e., two sequences present), there was an increase in the VTs reported suggests that whether listeners had to respond to only one or to both sequences at the same time matters. Warren and Ackroff (1976) reported a similar observation in the comparison of their two dichotic conditions. When participants were required to report VTs from both sequences, they produced fewer responses than when they had to monitor either the sequence played to the left ear or the right. In terms of the proportional change in the means of the two dichotic conditions in Warren and Ackroff's study, they reported a loss of around 50% of VTs when participants had to report from both sequences (mean of 10.85 VTs/5 min) compared with the case when they only had to report from one sequence (mean of 24.5 VTs/5 min). The equivalent conditions in the present study, M2 and M3 show a decrease or a loss of around 33% of VTs in favour of condition M2 (13.32 VTs/3 min for M2 vs. 8.81 VTs/3 min for M3). The difference in the relative proportion of the loss of the number of VTs might be attributed to the fact that in Warren and Ackroff's study listeners had to shift their attention between two spatially separate positions, whereas they did not in the current experiment. It is possible that shifting attention between left and right sides of space is more demanding than switching between sequences only distinguished by a difference in F0, with no spatial cues present. In addition, differences in the overall rate of responses per unit time might explain the proportional change of VTs between the two studies. While listeners in Warren and Ackroff's study reported more VTs on average than in the present experiment, their sequences lasted for 5 minutes compared to the 3 minutes used here. Hence, in the present study, the cost of monitoring both sequences might not have had relatively as great an impact on the attentional load for the listeners compared to Warren and Ackroff's study.

Note that the number of instances where listeners reported VTs on the low pitch while monitoring and reporting on both voices – condition High/**LOW** (7.22 VTs/3 min) elicited significantly fewer responses than the condition where participants heard both sequences but were required to report only the low pitch voice – condition High/Low (13.22 VTs/3 min) (p=.003). The equivalent condition Low did not differ from either High/Low or High/**LOW**. Taken together, these outcomes suggest that: (i) the presence of the two sequences boosts the number of VTs and (ii) the task demand of monitoring both sequences reduces the number of VTs.

The ANOVA summary table for the above description is presented in Table 2.13.

Table 2.13 Three-way ANOVA for VTs in Experiment 2.

Source	***df***	**F**	**P**	**η^2**
Condition (C)	2,18	3.48	=.053	=.28
Pitch (P)	1,9	0.93	=.36	=.09
Word (W)	5,45	1.64	=.17	=.15
C x P	2,18	1.42	=.27	=.14
C x W	10,90	1.27	=.26	=.12
P x W	5,45	0.67	=.65	=.07
C x P x W	10,90	1.04	=.42	=.10

Forms

For the number of forms, the ANOVA revealed a significant main effect of condition [F(2,18)=3.92, p=.04, η^2=.30], where condition M1 (3.64 Forms/3 min, s.e.=1.06) elicited significantly fewer new forms than condition M2 (5.28 Forms/3 min, s.e.=1.46) (p=.03). This indicates that the presence of the other sequence increases the number of forms perceived, even though listeners are (presumably) not monitoring it. This result is consistent with the notion that there is a significant opportunity for across-sequence re-groupings when both sequences can interact within the same ear of presentation.

There was also a significant main effect of word [F(5,45)=3.50, p=.01, η^2=.28], where LSD pairwise comparisons showed that ‘noise’ (3.38 Forms/3 min) differed significantly from ‘face’ (p<.01), ‘sleep’ (p=.04) and ‘see’ (p=.03), while stimulus ‘right’ (3.62 Forms/3 min) elicited significantly fewer forms than ‘flame’ (p=.01), ‘sleep’ (p=.03) and ‘see’ (p=.01). There was also a significant interaction between condition and word factors [F(10,90)=1.99, p=.04, η^2=.18] which was driven by the fact that only for ‘face’ and ‘sleep’ there were fewer forms reported in condition M1 compared to condition M2. Refer to Table 2.14 for the summary of the results of the analysis for Forms.

Table 2.14 Three-way ANOVA for Forms in Experiment 2.

Source	*df*	F	P	η²
Condition (C)	**2,18**	**3.92**	**=.04***	**=.30**
Pitch (P)	1,9	0.86	=.38	=.09
Word (W)	**5,45**	**3.50**	**=.01***	**=.28**
C x P	2,18	2.22	=.14	=.20
C x W	**10,90**	**1.99**	**=.04***	**=.18**
P x W	5,45	1.64	=.17	=.15
C x P x W	10,90	0.91	=.53	=.09

Timing of the first Verbal Transformation

Table 2.15 Average time of the first VT across participants (nil-responses marked as 180 s). Standard errors are shown in brackets.

	Average first VT (sec)
Low	69.65 *(19.26)*
High	66.63 *(18.76)*
High/Low	28.55 *(5.39)*
High/Low	26.77 *(6.66)*
High/LOW	46.53 *(9.08)*
HIGH/Low	29.42 *(10.71)*
Noise	70.62 *(20.95)*
Flame	41.58 *(8.58)*
Face	41.08 *(9.55)*
Sleep	36.13 *(6.96)*
See	32.43 *(7.30)*
Right	45.70 *(15.27)*

The results of a three-way, 3 (condition: M1, M2, M3) x 2 (pitch) x 6 (word) ANOVA for the timing of the first verbal transformation were as follows. There was a significant main effect of condition [F(2,18)=6.28, p=.01, η^2=.41] where supporting the results for Forms, listeners reported hearing first VT in condition M1 significantly later than in condition M2 (p=.01). For the significant main effect of word [F(5,45)=2.77, p=.03, η^2=.24], ‘noise’ differed from ‘face’ (p=.04) and ‘see’ (p=.04). No other effects were significant.

Dependency measure for conditions High/LOW & HIGH/Low

Table 2.16 Overall Means for the Dependency and Temporal Overlap indices for Experiment 2 (conditions High/**LOW** and **HIGH**/Low). Standard errors are shown in brackets.

Low Pitch Dependency Index (High/LOW)	High Pitch Dependency Index (HIGH/Low)	Overall Dependency Index	Temporal Overlap Index
0.26 *(0.04)*	**0.19** *(0.03)*	**0.23** *(0.04)*	**0.25** *(0.03)*

Note that the overall value for the dependency index is very similar to that observed for the corresponding case in Experiment 1, and that the temporal overlap index is somewhat lower (see Table 2.16). Overall, this outcome suggests that responses to both sequences are relatively unrelated to one another.

2.4.3 Summary and Conclusions

Whilst Warren and Ackroff (1976) used physical separation of the two sequences (i.e., dichotic presentation), in Experiment 2 the only cue for the segregation of the two sequences was the difference in F0. Overall, the results suggest a tendency for responses (VTs and forms) to increase and for the time to the first response to fall when the second sequence is present. These changes are offset, in part or in whole, when listeners are asked to monitor both sequences at once.

The fact that the number of VTs and forms declined in conditions High/**LOW** & **HIGH**/Low compared to High/Low & High/Low suggests a constraint arising from listeners trying to monitor both streams at the same time. In essence, the difference between conditions Low/High and High/Low & High/Low seems to be primarily driven by the stimulus difference, whilst the difference between conditions High/Low & High/Low and between conditions High/**LOW** & **HIGH**/Low is driven by the limitations of the response strategy. This further indicates that the particular combination of stimuli and task used in High/Low and High/Low is the most effective in terms of eliciting a greater number of reported VTs and forms.

Additionally, stimulus context seems to be affecting the outcomes of the study. Comparing conditions Low & High with High/Low & High/Low, even though listeners are only reporting one of the pitches, the addition of another pitch in conditions High/Low & High/Low resulted in an increase of the number of VTs and forms reported. This suggests a different type of regrouping of the speech sounds between the two pairs of conditions, and is likely to be influenced by the nature of the two sequences, where both were present in both ears at the same time (unlike for a dichotic condition).

It can be concluded that the effect of sequence pitch observed in Experiment 1 was not attributable to the resynthesis of the stimulus words per se, but rather to the demand characteristics of the task itself.

Chapter 3

Grouping and the Verbal Transformation Effect: The influence of formant transitions and pitch contour

3.1 Introduction

Pitt and Shoaf (2002) presented repeating sequences of standardised CVC syllables as stimuli to their listeners. They observed that some phonetic segments (e.g., voiceless fricatives, plosives) segregate into a separate stream much more easily than others (e.g., nasals, approximants). They did not, however, examine the effects on streaming of manipulating formant transitions. One of the arguments put forward about the cohesion of speech is that the formant transitions help to prevent auditory stream segregation. Studies by Cole and Scott (1973) and Dorman et al. (1975) have suggested that formant tracks "aid to preserve the temporal order of acoustic segments in on-going speech" and this notion can guide the kinds of manipulations of formant transitions to be tested.

Cole and Scott's (1973) study compared the tendency for a repeating cycle of CV syllables to undergo stream segregation in two conditions – unedited CV syllables vs. CV syllables edited (by analogue tape splicing) to remove the formant transitions between the consonant and vowel segments. Although the transitionless CV syllables sounded indistinguishable from the unedited versions when heard in isolation, when repeated rapidly the edited consonant and vowel segments segregated after only 2 or 3 repetitions into two different streams. For example, Cole and Scott found that rapidly repeated sequences of the CV syllable /sa/ tended to lead to the perceptual segregation of the unvoiced fricative from the vowel. It remains unclear, however, whether the resulting effect was due specifically to the removal of the formant transitions or was instead an artefact of the tape-splicing process.

3.2 Experiment 3 – Formant Transitions

The current experiment aims to develop Cole and Scott's study in the context of the VTE, but using precisely controlled digital editing to manipulate the transitions between the first two segments of monosyllabic words. This enables us to look more systematically at increases in the number and type of verbal transformations occurring when the critical transitions have been removed. For these purposes, formant transitions involving more substantial frequency changes from the initial consonant to the vowel (i.e., magnitude of the change in the second formant, F2) will be referred to as *strong transitions* whereas those involving smaller excursions will be referred to as *weak transitions*. If formant transitions help to prevent auditory stream segregation in the context of VTE, then removing them will result in more VTs and forms being reported. It is expected that taking out formant transitions from words with strong formant transitions will increase the number and type of transformations heard, showing evidence of perceptual regrouping. It is predicted that removing transitions that do not appreciably affect the intelligibility of isolated words may affect regrouping when the word is repeated, with consequent changes in the frequency and type of VTs. In contrast, taking out formant transitions from words with weak formant transitions should not increase the number and type of transformations heard. In the event that removing the formant transitions has a similar impact irrespective of whether they are weak or strong, this would suggest that the findings of Cole and Scott were simply an artefact of analogue splicing. As in Experiments 3 and 4, the manipulation of stimuli adhered primarily to the principle of good continuation, where the sequential (as opposed to simultaneous) grouping of speech elements is explored, single-sequence presentations were used (rather than two concurrent sequences played at the same time like in Experiments 1 and 2).

3.2.1 Method

Participants

Twelve participants (3 males, 9 females), all of whom reported normal hearing, completed the experiment (3 sets of 4 listeners). They were all native speakers of English and at the end of the study were either paid cash or received course credit. Listeners' mean age was 26.4 years old (s.d. = 4.83).

Stimuli and Conditions

Stimuli were chosen from a set of CVC monosyllabic words with non-centralised tense (long) vowels, as the corresponding longer durations of the steady-state portions made it easier to manipulate the formant transitions between the initial consonant and the vowel. The words included voiceless fricatives (*f*, *th*, *s*, *sh*), voiceless plosives (*p*, *t*, *k*) or the voiceless affricate (*ch*). The words, spoken and recorded by the author, were slightly hyperarticulated to obtain clearer transitions. Before the experiment proper, it was established in a pilot study that the chosen set generated a reasonable number of VTs and forms. The stimulus set included 12 monosyllabic words: 6 with strong formant transitions between the initial consonant and the vowel: **short**, **chart**, **sharp**, **seek**, **thought** and **torch** (*Strong* set) which were paired with 6 words producing weak transitions: **fort**, **park**, **sheep**, **peak**, **caught** and **porch** (*Weak* set). Words were paired such that the first pair was 'short-fort', the second was 'chart-park' and so on.

All test words were recorded (natural utterances) and then processed to create the reference stimuli using PRAAT & Adobe Audition. The reference stimuli were each set to be 500 ms long, which gave 360 repetitions in 3 min. Words were monotonized, and resynthesized at an F0 of 130 Hz (similar to the mean pitch of the speaker). In addition, an edited version of each word was created from the reference set of 12 stimuli. For each word pair (e.g. 'short' – 'fort'), the amount of editing (in ms) applied to the strong transition word (e.g. 'short'), as determined by the duration of these transitions, equalled the amount of editing applied to the corresponding weak transition word (e.g. 'fort'). The edited durations for each word pair are presented in the Table 3.1.

Table 3.1 Edited durations of each word pair used in Experiment 3. 'C' refers to the amount of editing applied to the initial consonant while 'V' represents the amount of editing applied to the vowel.

Word Pair			C (ms)	V (ms)
short	–	fort	21	69
chart	–	park	17	69
sharp	–	sheep	11	61
seek	–	peak	26	30
thought	–	caught	14	69
torch	–	porch	9	46

To create the edited version of each word, formant transitions consisting of the last 9-26 ms of the consonant and the first 30-69 ms (4-9 glottal pulses) of the vowel were removed. To replace the cropped-out segment of the vowel, a single glottal pulse from the steady-state portion was iterated several times in its place. For the removed portion of the consonant, a corresponding middle segment of the same consonant was copied and spliced in. Figure 3.1 shows spectrograms for the first stimulus pair: ‘short-fort’. The red regions correspond to the manipulated areas where the transitions have been edited. Note the clear formant transitions visible in the top-left spectrogram. For the complete set of spectrograms of the stimuli words refer to the Appendix 2 where, as in the below example, the regions highlighted in red indicate those parts of the stimuli that were subject to digital editing.

After the editing procedure (see below) the amplitude envelope (extracted from the original – monotonised – recording) was applied using a PRAAT script to the ‘no transition’ version of the stimuli.

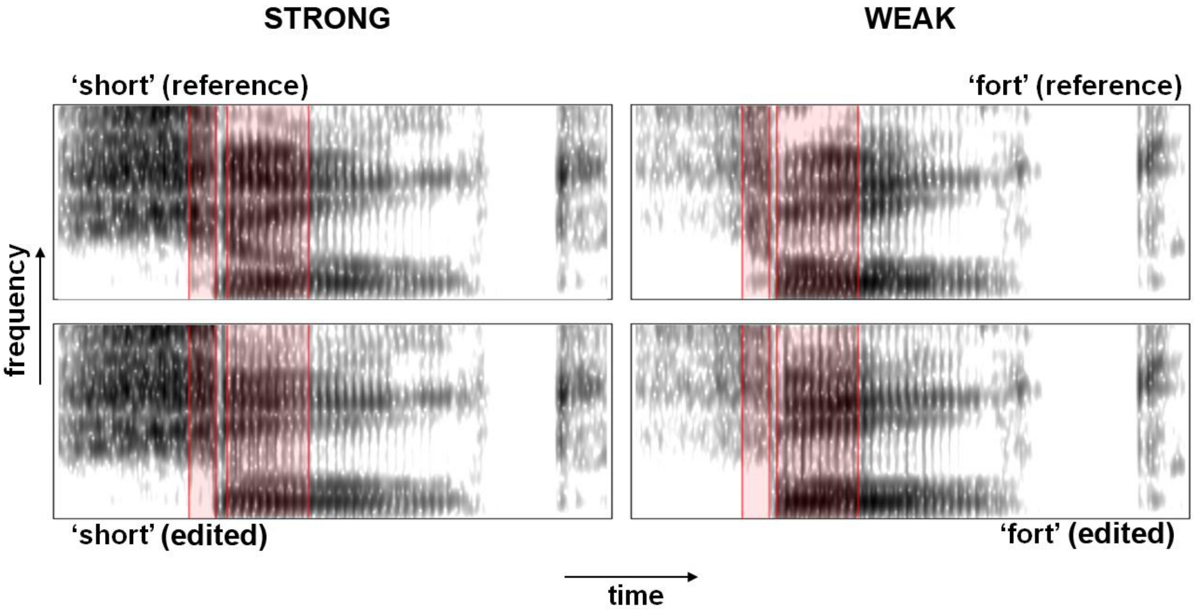


Figure 3.1 Spectrograms for the first word pair ‘short’ – ‘fort’.

Each listener attended 4 sessions (each consisting of six 3-minute presentations) and the counterbalancing procedure is presented in Table 3.2. Words were arbitrarily divided into ‘First 6’ – (short, fort, chart, park, sharp and sheep) and ‘Second 6’ (seek, peak, thought, caught, torch, porch) so that both groups included equal number of words with strong and weak transitions. As the difference between the ‘First 6’ and ‘Second 6’ was not of primary interest to the study, the order of the four sessions for each participant was ‘First 6’-‘Second

6’-‘First 6’-‘Second 6’. However, the factor of editing - whether the word was in its *reference* form or had its formant transitions *edited* - was counterbalanced such that the session sequence for the odd numbered listeners was reference-edited-edited-reference, and for the even numbered listeners it was edited-reference-reference-edited.

Table 3.2 Session counterbalancing in Experiment 3.

		Odd numbered listeners	**Even numbered listeners**
Session number	**1**	First 6 (reference)	First 6 (edited)
	2	Second 6 (edited)	Second 6 (reference)
	3	First 6 (edited)	First 6 (reference)
	4	Second 6 (reference)	Second 6 (edited)

As in previous experiments, listeners were asked to report every change in word identity that they heard. Within each session, every experimental stimulus (strong transitions, henceforth referred to as Strong) that was presented was always accompanied by its own control (weak transitions, henceforth referred to as Weak). Measures taken included the number of verbal transformations (VTs, any change to the reported stimulus), the number of Forms (any transformation that has not occurred before), and the timing of the first verbal transformation.

The data were analysed in terms of the difference in transformations reported between the reference and edited versions of the stimulus words for each set and the hypothesis was that this difference would be significant for the Strong transitions set but not for the Weak transitions set. As the stimuli from the Weak set showed little movement of their formant frequencies (especially the second and third formant) during the initial CV segment of the word, the editing procedure should not have an effect on the rate and type of VTs, whereas the opposite should be true for the Strong set. The three within-subjects factors were: **Transitions** (Strong, Weak), **Editing** (Reference, Edited) and **Word pair** (short-fort, chart-park, sharp-sheep, seek-peak, thought-caught, torch-porch). Therefore, in terms of statistical outcomes from the resulting three-way repeated-measures ANOVA, the interaction between the first two factors, i.e. Transitions x Editing was of most interest. However, all other effects are presented and described below for each of the three measures: VTs, forms and the timing of the first VT.

3.2.2 Results

The mean values for the four conditions from the crucial Transitions x Editing interaction are presented in Table 3.3. These are collated for all three measures taken in the current experiment.

Table 3.3 Mean values for the four conditions from the Transitions x Editing interaction for all three measures taken. Inter-subject standard errors of the mean are reported in brackets.

	Average no. of VTs reported in 3 min *(±SE)*	**Average no. of new Forms reported in 3 min** *(±SE)*	**Average time of first VT (in seconds)** *(±SE)*
Strong Reference	16.26 *(2.33)*	6.38 *(0.82)*	17.64 *(2.15)*
Strong Edited	17.17 *(2.36)*	7.40 *(1.11)*	17.71 *(2.04)*
Weak Reference	13.79 *(2.15)*	6.50 *(0.88)*	20.97 *(3.07)*
Weak Edited	15.81 *(2.80)*	6.65 *(1.03)*	20.93 *(3.52)*

In terms of the statistical outcomes/effects from the 3-way ANOVA, a few considerations have to be taken into account. As some of those effects are of greater importance/relevance to the study than others, the nature of each term will be described below:

- ❖ Main effect of Transitions

  This effect describes the difference between the Strong (transitions) set of words and the Weak (transitions) set, irrespective of whether the transitions have been spliced out or not (Reference vs. Edited). As the words in the two groups are different, the significance or otherwise of this main effect is of little interest.

- ❖ Main effect of Editing

  This effect compares all the Reference words (where the transition has been left intact) with their edited versions, irrespective of whether they were from the Strong or the

Weak group. In other words, it looks at whether the editing procedure had an effect on its own – whether, for example, it will tend to increase overall the number of VTs or new Forms.

One of the motivations for the current experiment was to test the possibility that the results from Cole and Scott's (1973) study were due to the editing procedure itself. The manual process of removing the transitions from analogue tape recordings and replacing the excised segment with an alternative (most probably using splicing tape) could have introduced artefacts (clicks, noise) that may have influenced the results. Of itself, a significant main effect would be of limited interest, as it cannot distinguish between relevant and artefactual consequences of editing. However, taken together with the interaction terms involving Editing, the main effect of Editing can guide a possible discussion of how any observed effects of taking out the formant transitions might have arisen.

- Main effect of Word pair

Given the somewhat arbitrary nature of the word pairs chosen, on its own this effect does not contribute to understanding the processes investigated in this study. Each word pair consists of both a Strong and a Weak stimulus and it is balanced across Reference and Edited word tokens. Hence, a significant main effect would simply indicate that some word pairs produce more responses than others.

The reason for the inclusion of this factor is to investigate whether the crucial two-way Transitions x Editing interaction (described below) can be potentially influenced by the stimulus words themselves. If that was the case, it would be indicated by a significant three-way interaction.

A separate, descriptive analysis of the results for all stimulus words used in the study, for each of the three experimental measures, is presented in a later section.

- Transitions x Editing interaction

This is the crucial interaction for the experimental hypothesis of the study, which states that the procedure of taking out and replacing the formant transitions will affect the Strong set of words but not the Weak set. Such an interaction might potentially be observed in either the reported number of VTs, new Forms or the time of the first VT. It is worth noting that, in practice, it is not possible to completely eliminate the effect of the editing process itself as the word tokens are nevertheless altered through digital manipulation. However, it is reasonable to assume that although editing will have some effect in general, it will have a considerably larger effect on the Strong set compared to the Weak one. In particular, for the Strong set it was hypothesised that listeners will report more VTs and forms for words with the formant transitions removed and replaced compared to the words with the formant transitions unaltered. This difference will not be observed for the Weak set, where removing and replacing the formant transitions should have little or no effect on the number of VTs and forms reported.

- Transitions x Word pair interaction

Given that there is no distinction involving editing here, this interaction merely informs us about which word pairs are associated with larger differences between the Strong and Weak conditions. It is, therefore, of little relevance to the study.

- Editing x Word pair interaction

Similar to the above interaction, this one is of little importance. It looks at which word pairs were more affected by the editing procedure regardless of the Transitions factor. Since the experimental hypothesis is based on the factors of Transitions and Editing, the Word pair factor only becomes relevant in the context of the three-way interaction, described below.

- Transitions x Editing x Word pair interaction

The relevance of the three-way interaction is based on the possibility that some word pairs may show a greater Transitions x Editing interaction than others. If statistically significant, it can help to identify which word pair(s) might be driving the interaction between Transitions and Editing, in which case a more detailed analysis of the phonetic structure between the words, e.g. of the duration or velocity (rate of change) of the formant transitions, might be required.

In view of the above considerations, the following results will concentrate on the critical interaction between Transitions and Editing and will comment on the remaining effects only if they are relevant to the interpretation of the study. However, full ANOVA tables with all the statistical results have been included at the end of each section.

Verbal Transformations

The Transitions x Editing interaction did not reveal a significant effect for VTs [F(1,11)=0.58, p>.4]. Removal and replacement of the formant transitions did not have a differential effect on the number of VTs reported for the Strong transition words and Weak transitions words. There was an overall trend towards more VTs reported for edited stimuli (15.03 VTs/3 min for Reference vs. 16.49 VTs/3 min for Edited), however, the corresponding main effect of Editing did not quite reach significance [F(1,11)=3.84, p=.08]. The ANOVA summary Table 3.4 for VTs is presented below.

Table 3.4 Summary of three-way ANOVA for verbal transformations.

Source	*df*	F	p	η^2
Transitions (T)	**1,11**	**13.00**	**<.01****	**=.54**
Editing (E)	1,11	3.84	=.08	=.26
Word pair (W)	**5,55**	**2.56**	**=.04***	**=.19**
T x E	1,11	0.58	=.46	=.05
T x W	**5,55**	**5.49**	**<.01****	**=.33**
E x W	5,55	2.27	=.06	=.17
T x E x W	5,55	1.49	=.21	=.12

Forms

As for the number of VTs, there was some evidence of a general trend towards more forms being reported when the stimuli had their transitions removed and replaced (6.44 Forms/3 min for the Reference stimuli vs. 7.03 Forms/3 min for their Edited equivalents). Nonetheless, the main effect of Editing did not quite reach significance [F(1,11)=3.50, p=.09]. Crucially, however, the two way interaction between Transitions and Editing was highly significant [F(1,11)=11.96, p<.01, η^2 =.52]. Post hoc analysis using the restricted LSD test revealed statistically significant differences between the set of words with *strong* transitions that have not been altered – Strong Reference stimuli, and words with *strong* transitions that have been spliced out and replaced with steady-state segments – Strong Edited stimuli (p<.01). On average, participants reported 6.38 Forms/3 min when listening to the Strong Reference stimuli, which was significantly less than 7.40 Forms/3 min when they were presented with the same words but with no formant transitions – the Strong Edited set. The difference between the Reference and Edited stimuli from the Weak set (words with *weak* transitions) was not significant [6.50 Forms/3 min for Weak Reference vs. 6.65 Forms/3 min for Weak Edited]. The three-way interaction was not significant indicating that the interaction of interest – Transitions x Editing was not driven more by some word pairs than others. Table 3.5 includes the ANOVA summary for Forms.

Table 3.5 Summary of three-way ANOVA for forms.

Source	***df***	**F**	**p**	**η²**
Transitions (T)	1,11	1.05	=.33	=.09
Editing (E)	1,11	3.50	=.09	=.24
Word pair (W)	**5,55**	**3.80**	**<.01****	**=.26**
T x E	**1,11**	**11.96**	**<.01****	**=.52**
T x W	5,55	1.65	=.16	=.13
E x W	5,55	2.31	=.06	=.17
T x E x W	5,55	1.57	=.19	=.13

Timing of the first verbal transformation

The three-way ANOVA for the average time of the first VT did not reveal anything of interest – see Table 3.6.[4] The general trend to report more VTs and Forms for the edited stimuli was not upheld for this measure. Listeners were equally quick to provide their first response to the Reference words (with an average time of 19.31 s) as for the Weak ones (with an average time of 19.32 s); the main effect of Editing was not significant (p=0.99). Contrary to the prediction, for the Strong set participants were marginally quicker to report the first VT in the Reference condition – 17.64 s than in the Edited condition – 17.71 s. However, this simply reflects chance variability, as the critical Transitions x Editing interaction was not significant (p=0.98).

[4] Note that the issue of timeouts seen in Experiment 1 and 2 was not a significant contributor to the main outcome of the present study. There were only two trials where a listener did not report any VTs (see section commenting on individual words). For the purposes of the analysis these cases were marked as 180 s.

Table 3.6 Summary of three-way ANOVA for time of first VT.

Source	***df***	**F**	**p**	**η²**
Transitions (T)	1,11	2.71	=.13	=.20
Editing (E)	1,11	<.001	=.99	<.01
Word pair (W)	**5,55**	**3.80**	**<.01****	**=.26**
T x E	1,11	=.001	=.98	<.01
T x W	5,55	2.14	=.07	=.16
E x W	5,55	1.49	=.21	=.12
T x E x W	5,55	0.62	=.69	=.05

Comments on the individual stimulus words

Table 3.7 includes the average number of VTs, Forms, and the time of the first VT for each stimulus word collapsed across all conditions used in the study. Visual inspection allows us to comment on the particular measures as well as inspect the scores for individual words. It is evident that the time of the first VT can be heavily influenced by a ‘no response’ from a listener. One participant in the study did not experience any VTs for the single word ‘park’, which resulted in the average time to the first VT for this word to be coded spuriously as late (this instance was marked as 180 s for the first VT). Also, there does not seem to be any pattern with regards to Reference vs. Edited stimuli. For six out of 12 words, responses were quicker for Reference words and five produced the opposite result (one was equal in both conditions).

The lack of VTs for one word by a single participant also seemed to have influenced the average number of VTs. VTs for the word ‘park’ were reported on significantly fewer occasions (9.75) in the Reference conditions compared to other words (see Table 3.7). On the other hand, the Forms measure seems to be the most resilient to such instances – ‘park’ has a fairly typical – middle of the range – value (5.38) for the stimuli in the reference condition. Two words ‘torch’ (Strong condition) and ‘caught’ (Weak condition) show a reverse trend for all three measures compared to the study predictions. There were fewer VTs and Forms reported as well as a longer average time to the first VTs in the edited condition. While ‘caught’ was included in the Weak set (hence the editing procedure should not be the cause of this trend), the results for ‘torch’ could highlight its phonetic differences compared to the other stimuli. It is the only word from the Strong set that has an affricate ‘ch’ at the terminal position (other words end with plosives). The only word from the Strong set which includes ‘ch’ in initial position - ‘chart’ - shows a similar trend with fewer VTs and a later time to the first VT in the edited condition, however, it produced more Forms in the edited condition. For the full list of all the forms reported by participants, refer to Appendix 3 where it is also worth noting that the comparison of the specific forms reported for reference vs. weak versions of each word, did not reveal any obvious patterns.

Table 3.7 Mean values for all stimulus words in the three measures taken in Experiment 3. Standard errors of the mean are reported in brackets.

		Average no. of VTs reported in 3min *(±SE)*		Average no. of new Forms reported in 3min *(±SE)*		Average time of first VT (in seconds) *(±SE)*	
		Reference	**Edited**	**Reference**	**Edited**	**Reference**	**Edited**
Strong	**Short**	15.00 *(1.88)*	21.92 *(3.80)*	6.83 *(1.21)*	9.33 *(1.81)*	16.25 *(2.35)*	11.92 *(1.74)*
	Chart	15.33 *(2.29)*	13.58 *(2.24)*	6.83 *(0.91)*	7.33 *(1.28)*	18.50 *(3.57)*	24.25 *(6.43)*
	Sharp	13.17 *(2.58)*	16.75 *(2.97)*	5.08 *(0.70)*	7.08 *(1.26)*	25.00 *(3.52)*	19.92 *(4.47)*
	Seek	16.42 *(3.85)*	17.75 *(3.33)*	5.00 *(0.84)*	5.67 *(0.81)*	17.75 *(4.92)*	22.08 *(6.07)*
	Thought	20.67 *(3.43)*	20.33 *(2.72)*	6.92 *(0.84)*	8.50 *(1.01)*	14.25 *(2.88)*	9.17 *(0.94)*
	Torch	17.00 *(3.21)*	12.67 *(1.68)*	7.58 *(1.49)*	6.50 *(0.99)*	14.08 *(2.86)*	18.92 *(4.11)*
Weak	**Fort**	16.00 *(2.20)*	18.67 *(3.70)*	6.92 *(0.47)*	8.50 *(1.41)*	13.92 *(5.32)*	13.67 *(1.77)*
	Park	9.75 *(2.28)*	13.08 *(3.31)*	5.83 *(1.19)*	6.08 *(1.72)*	37.25 *(13.99)*	44.08 *(14.45)*
	Sheep	12.42 *(1.97)*	19.25 *(4.17)*	5.25 *(0.57)*	6.33 *(0.88)*	25.67 *(4.42)*	19.67 *(3.49)*
	Peak	15.42 *(3.55)*	16.83 *(2.64)*	6.42 *(1.05)*	6.67 *(0.97)*	22.17 *(3.64)*	18.00 *(3.77)*
	Caught	14.17 *(2.41)*	12.17 *(2.34)*	7.58 *(1.47)*	5.83 *(0.66)*	12.67 *(2.35)*	16.00 *(2.95)*
	Porch	15.00 *(2.81)*	14.83 *(2.80)*	7.00 *(1.23)*	6.50 *(1.33)*	14.17 *(1.74)*	14.17 *(2.19)*

3.2.3 Summary

Listeners reported significantly more new Forms for the edited words in the Strong condition, whereas editing had little or no effect in the Weak condition. This suggests that: (1) the effect of editing as reported by Cole and Scott was not simply an artefact of manipulating the stimuli and (2) it supports the perceptual reorganization hypothesis, whereby continuity of formant tracks facilitates the integration of rapidly cycled speech segments into a single perceptual stream, which helps to maintain the perceived temporal order of the phonetic segments

With regards to the overall effect of editing, there is some evidence of a general trend for VTs to be reported more often where the formant transitions have been removed and replaced. Even digital editing (as opposed to the analogue tape splicing used in the Cole & Scott study) may cause a propensity to hear an increased number of VTs, thus implying a greater tendency for streaming. Note that this cannot account for the significant interaction found for Forms.

The present study, and the previous two experiments on the VTE presented in this thesis, seems to indicate that Forms are likely to be more clearly influenced by changes related to grouping, and hence are better suited than the number of VTs or the time of the first VT as an experimental measure. The Forms measure appears to be more stable, with a smaller variance compared to the number of VTs. At least in part, this may be affected by listeners not always reporting every change in the stimulus when it has been heard before (e.g. during rapid oscillation between two forms, see Ditzinger, Tuller & Kelso, 1997).

There is evidence that perceptual re-grouping of repeated segments of speech is an important contributor to VTs and therefore the changes in the VTE seen here for Forms most likely reflect changes in stream segregation. Overall, the present study demonstrates a clear effect of formant transitions on the VTE, and supports the findings from related studies such as Cole and Scott (1973), who concluded that formant transitions play an important role in binding disparate speech segments together into a single auditory stream.

3.3 Experiment 4 – Pitch Contour

The results of Experiment 3 showed how smooth gliding formant transitions help to group the often disparate elements of speech together. The following experiment explored the influence of pitch contour on VTs, as a natural continuation of and extension to the formant transitions study. It was intended to test further the hypothesis that the Gestalt principle of good continuation (smoothness of change) plays an important role in holding speech segments together. Darwin and Bethell-Fox (1977) have provided evidence for the role of the pitch contour in holding the speech stream together, but to our knowledge its role has not been investigated to date in the context of the VTE. Darwin and Bethell-Fox showed how speech can break up into two different voices when artificially abrupt alterations between high and low pitches are introduced in the speech signal. Additionally, Bregman and Dannenbring (1973) illustrated how smoothness of change indicated by the unbroken spectral pattern of a sequence of high and low frequency pure tones joined by frequency glides helps to hold the sequence together.

As large jumps in F0 between adjacent words in are rare in conversational speech, it should be possible to demonstrate with the VTE that such jumps, as represented by the pitch contour of the consecutive word tokens, will result in perceptual streaming occurring more readily – i.e., listeners reporting more VTs and forms. On the other hand, if the consecutive instances of presented words follow a smooth pitch contour, listeners will report fewer VTs and forms.

3.3.1 Method

Participants

Twenty four participants (6 males, 18 females) took part (4 sets of 6, with conditions counterbalanced across participants) and as before they were asked to report any changes to the stimulus. They were all native speakers of English and reported normal hearing. At the end of the study they were either paid cash or received course credit. Listeners' mean age was 24.1 years old (s.d. = 5.56).

Stimuli and Conditions

Single-sequence recordings were played to listeners and each participant attended 3 separate sessions corresponding to the 3 possible arrangements of the direction of the pitch contour. These were: (1) *all falling* (FF) where each repetition of the word token in a 3-minute sequence followed a pitch contour from high to low, (2) *all rising* (RR) where each token in a 3-minute sequence followed a pitch contour from low to high, and (3) *alternating* (RF) where the pitch contours of successive tokens in a 3-minute sequence alternated between rising and falling.[5] The pitch contour applied to the stimulus words was a half-sine trajectory on a linear scale and was within the range of variation for a normal human voice - an octave: 100 Hz (low pitch) – 200 Hz (high pitch). The first 2 conditions with half-sine cycle waves had abrupt pitch discontinuities at the word boundaries; it was hypothesised that this will affect regroupings such that they should be more prone to streaming than the alternating RF condition. Figure 3.2 demonstrates the differences between the three conditions. Notice the

[5] The alternating falling-rising (FR) condition was not ran as well as an RF condition, as the only difference between the two would be the direction of the pitch contour for the first and last cycles. Given that the first VT occurs long after the first cycle (on average after 20 sec), and that no responses can be made after the last cycle, these differences were considered trivial.

abrupt changes in F0 frequency at each word boundary in *all rising* and *all falling* conditions, while the half-sine shape leads to a smooth F0 frequency contour in the *alternating* case.

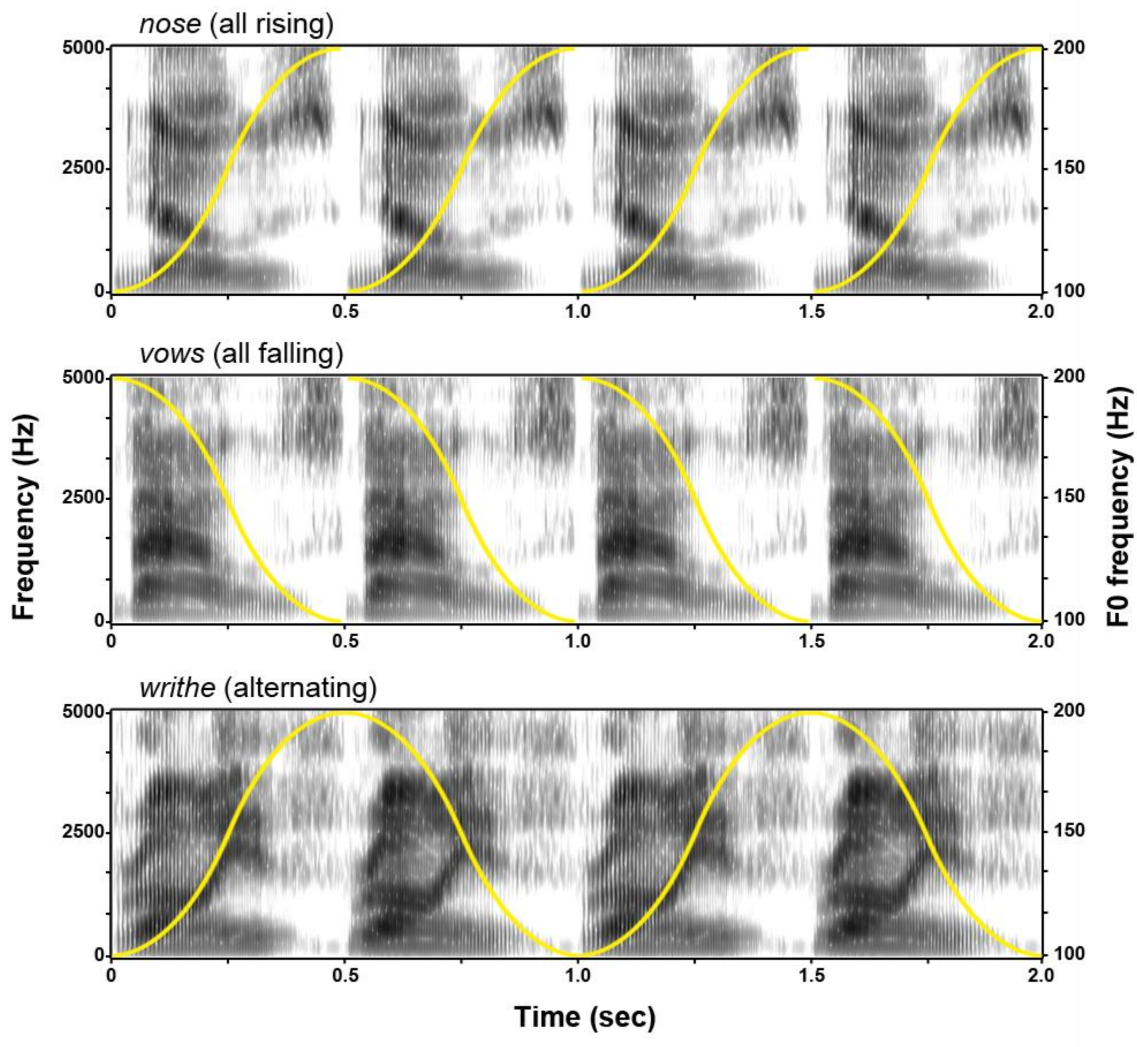


Figure 3.2 Spectrograms showing examples of three words used in Experiment 4. Each word is shown repeated four times. The yellow line represents the F0 frequency contour applied in each condition – on the top *all rising*, in the middle *all falling*, and at the bottom *alternating*.

The stimulus set consisted entirely of continuously voiced words: **vows**, **wave**, **maze**, **nose**, **lathe**, **writhe**; the training word, **rose,** was also continuously voiced. Several examples of each were recorded by the same speaker as in Experiment 1. The best tokens of these words were chosen from the recordings by looking at 3 major factors: what is their VT potential (how quickly do they transform and how many transformations do they evoke?), how good they sound when time trimmed in CoolEdit (this was the experimenter's subjective opinion), and how much of the voicing could be identified automatically in PRAAT (the algorithm was not always successful at extracting the pitch contour in full). After the final set had been identified, words (1) were time warped using CoolEdit software to 500 ms each, using the same technique as in Experiment 1; (2) had amplitude contours of 5 ms imposed on the start and end of each file using CoolEdit, and (3) had PRAAT scripts applied to create the pitch contours needed for each condition (PSOLA alogirthm).

3.3.2 Results

Two-way 3 (Condition) x 6 (Word) within-subjects ANOVA was performed separately for all three measures (VTs, Forms, and time of first VT). All three analyses yielded the same result, where the only statistically significant effect was the main effect of Word (see Table 3.8).

Table 3.8 Results for all three ANOVAs in Experiment 4 where the three levels of Condition were the pitch contours: *all falling*, *all rising*, and *alternating*.

	Source	***df***	**F**	**p**	**η^2**
VTs	Condition (C)	2,46	1.71	=.19	=.07
	Word (W)	**5,115**	**4.85**	**<.01****	**=.17**
	C x W	10,230	1.36	=.20	=.06
Forms	Condition (C)	2,46	3.10	=.06	=.12
	Word (W)	**5,115**	**15.73**	**<.01****	**=.41**
	C x W	10,230	1.65	=.09	=.07
First VTs	Condition (C)	2,46	1.34	=.27	=.07
	Word (W)	**5,115**	**6.60**	**<.01****	**=.22**
	C x W	10,230	0.26	=.99	=.01

Since the factor of Word simply compares the differences between the word tokens regardless of the experimental condition, it is of little importance to the study. For the main effect of Condition, there is evidence of a trend in the direction of the experimental hypothesis for forms (p=.06), but not for the other two measures. Furthermore, for average number of VTs and Forms, the results for condition RR appear to be intermediate between those for conditions FF and RF, rather than more similar to those for condition FF. Table 3.9 shows that in the RR condition listeners reported on average 18.64 VTs and 7.17 Forms while the equivalent averages where lower in the RF condition (16.83 VTs and 6.63 Forms) and higher in the FF condition (19.22 VTs and 7.60 Forms).

Table 3.9 Mean values for all experimental conditions, and individually for each stimulus word, for the three measures taken. Standard errors of the mean are reported in brackets.

	Average no. of VTs reported in 3 min *(±SE)*	**Average no. of new Forms reported in 3 min** *(±SE)*	**Average time of first VT (in seconds)** *(±SE)*
Condition FF	19.22 *(2.73)*	7.60 *(0.80)*	20.03 *(2.13)*
Condition RR	18.64 *(2.65)*	7.17 *(0.74)*	19.79 *(2.11)*
Condition RF	16.83 *(2.63)*	6.63 *(0.76)*	22.74 *(2.22)*
Vows	21.46 *(3.50)*	8.40 *(0.78)*	14.60 *(1.75)*
Wave	15.39 *(2.75)*	4.94 *(0.62)*	35.44 *(5.76)*
Maze	18.24 *(2.63)*	7.03 *(0.75)*	16.65 *(1.69)*
Nose	15.85 *(2.39)*	5.96 *(0.71)*	22.24 *(3.52)*
Lathe	18.78 *(2.37)*	8.60 *(0.90)*	18.43 *(2.76)*
Writhe	19.65 *(2.55)*	7.88 *(1.01)*	17.79 *(2.08)*

A speculation on the reason for this observed difference in performance between *all falling* (FF) and *all rising* (RR) contours, associated with the direction of the pitch contour is considered in the summary section below. Given this directional effect and the trend towards a main effect of condition for number of forms, an additional analysis was designed to compare the mean performance of the two discontinuous contours (all rising RR and all falling FF, collapsed together) against the *alternating* contour RF. The rationale for this was to ensure that any effects apparent in the analysis are driven by differences between conditions in pitch-contour continuity between successive word tokens, rather than by other kinds of difference between the all-rising and all-falling contour cases. Therefore, the two experimental conditions used in the additional analysis were *Continuous contour* (RF – alternating contour) and *Discontinuous contour* (mean average of all rising RR and all falling FF). Note that the results for the continuous and discontinuous conditions both involve an equal (50:50) contribution of the rising and falling pitch contours. The mean values for all conditions and words are presented in Table 3.10.

Table 3.10 Mean values for all experimental conditions, and individually for each stimulus word, for the three measures taken. Standard errors of the mean are reported in brackets.

	Average no. of VTs reported in 3 min *(±SE)*	**Average no. of new Forms reported in 3 min** *(±SE)*	**Average time of first VT (in seconds)** *(±SE)*
Continuous	16.83 *(2.63)*	6.63*(0.76)*	22.74 *(2.58)*
Discontinuous	18.93 *(2.59)*	7.39 *(0.74)*	19.92*(1.78)*
All falling only	*19.22(2.73)*	*7.60(0.80)*	*20.04(2.13)*
All rising only	*18.64(2.65)*	*7.17(0.74)*	*19.79(2.11)*
Vows	21.64*(3.50)*	8.40*(0.78)*	14.60*(1.75)*
Wave	15.39 *(2.75)*	4.94*(0.62)*	35.44*(5.76)*
Maze	18.24*(2.63)*	7.03 *(0.75)*	16.65*(1.69)*
Nose	15.85*(2.39)*	5.96*(0.71)*	22.24*(3.52)*
Lathe	18.78*(2.37)*	8.60 *(0.90)*	18.43 *(2.76)*
Writhe	19.65*(2.55)*	7.88*(1.01)*	17.79*(2.08)*

Two-way 2 (Condition) x 6 (Word) within-subjects ANOVA was performed for each of the three measures (see Table 3.11). All three analyses yielded a similar outcome to the original analysis, with a significant main effect of Word. However, there was also a significant main effect of Condition for Forms (p=.02) in the predicted direction – listeners reported more forms in the discontinuous condition compared the continuous one. There was also a trend in the same direction for the other two measures – i.e., towards more VTs and shorter times to the first VT for the discontinuous condition (p=.06 in both cases). None of the measures showed any evidence of a condition x word interaction.

Table 3.11 Results for all three ANOVAs in Experiment 4 with *continuous* and *discontinuous* pitch contour as two levels of Condition.

	Source	***df***	**F**	**p**	**η²**
VTs	Condition (C)	1,23	3.97	=.06	=.15
	Word (W)	**5,115**	**4.73**	**<.01****	**=.17**
	C x W	5,115	0.76	=.58	=.03
Forms	Condition (C)	**1,23**	**6.26**	**=.02***	**=.21**
	Word (W)	**5,115**	**16.47**	**<.01****	**=.42**
	C x W	5,115	0.14	=.98	=.01
First VTs	Condition (C)	1,23	3.93	=.06	=.15
	Word (W)	**5,115**	**5.58**	**<.01****	**=.20**
	C x W	5,115	0.35	=.88	=.02

3.3.3 Summary

The results of the additional analysis provide support for a contribution of continuity of the pitch contour to the perceptual cohesion of speech. The raw responses given by participants are included in Appendix 4. However, inspection of these data did not reveal any obvious underlying patterns. No one condition evoked a substantial number of unique Forms that were not seen in the other conditions. Anecdotally, the *all falling* FF Condition, apart from 'writhe', never had fewer forms than the other two conditions. The RF Condition, on the other hand, apart from 'vows', 'lathe', and 'writhe', never had more forms than the other two. The most frequent responses for each stimulus word are fairly similar across conditions, both in terms of number of responses for those forms and their phonetic properties.

With regards to the observed difference between the *all rising* and *all falling* configurations, it is worth noting that the role of pitch-contour direction has not been investigated before in the context of the VTE. Word tokens used in previous studies usually retained their natural pitch contours. Especially in the 'classic' studies reported in the 60's and 70's, before digital manipulation of pitch was made possible (like PSOLA), the stimulus set was obtained by a researcher attempting to speak on the monotone. As such, there is no benchmark with which to compare the current results. Although it is not obvious why the observed difference between all rising and all falling contours should occur, from the linguistic point of view the most obvious difference between them is the fact that the falling intonation contour is more common and the rising one is usually used in questions.

3.4 General Discussion

The pair of experiments in this chapter explored the effects of two continuity cues applied to the VTE. In summary, there is clear evidence that manipulation of strong formant transitions and smoothness of change in the pitch contour influence the number of forms heard. Hence, the results are consistent with the hypothesis that formant transitions between phonetic segments and the continuity of the pitch contour both influence the regrouping of phonetic segments. In the case of the formant transitions, the separation of unvoiced fricatives from the vowel was much more easily obtained, further showing that the formant transitions are important for maintaining speech cohesion. As hypothesised, results show a significant interaction between editing and the type of formant transitions involved for Forms – there was a greater increase in Forms when the Strong transitions stimuli were edited compared with the Weak transitions stimuli. In contrast, there was no interaction for the number of VTs or the time to first VT. In the case of the pitch contour manipulations, all stimuli were continuously voiced and so may have been more resistant to perceptual re-grouping than the more heterogeneous stimuli used in formant transitions experiment. Nonetheless, the one-octave change at the word boundaries in the discontinuous conditions was clearly sufficient to increase the number of Forms reported.

This outcome supports Cole and Scott’s (1973) speculation that formant transitions play an important role in holding together disparate speech segments into a single sequential stream. From their study, however, it was unclear whether the results they obtained were due to the editing procedure itself, rather than specifically to the removal of the formant transitions. Using a more sophisticated process of digital editing, the current study has shown that even digital editing produces a greater general propensity to increase the number of Forms; however, the critical difference is brought about by the interaction of the Editing and Transitions factors.

In Experiment 4, on only two trials (by single participants) there were no VTs reported in a 3 minute presentation, and there were no such cases in Experiment 3. This is in contrast with Experiments 1 and 2, where there were no VTs reported on 15% to 35% of trials for each participant. This difference is presumably related to the fact that all the conditions in Experiments 3 and 4 involved a single sequence and the average duration of the word was shorter (550 ms in Experiments 1 and 2 compared with 500 ms in Experiments 3 and 4), so there were more repetitions a 3 minute period.

Chapter 4

Grouping and the Phonemic Transformation Effect: The influence of fundamental frequency and interaural time-difference cues.

4.1 Introduction

Another factor important for maintaining the perceptual integrity of a sound source over time is timbre – which in general depends primarily on the spectral content of a sound, and ranges from "dull" (most prominent spectral components in the lower frequency regions) to "bright" (most prominent spectral components in the higher frequency regions). Timbre distinguishes two sounds on the same F0 by the way the energy is distributed across the frequency spectrum. This implies that, if played in a sequence, two sounds on the same pitch but with differences in bandwidth, spectral centroid (centre of gravity), or spectral shape will undergo segregation based on differences in timbre (van Noorden, 1975). Similarly, manipulating the timbres of alternating complex tones or steady-state vowels, as in the Wessel illusion, has been shown to separate an ascending (in F0) sequence of three tones or vowels into two separate percepts, for which the segregation changes a single, rapid, rising motif into two slowly descending motifs (Wessel, 1979, see Introduction).

Related to the phenomenon of streaming by timbre is the *Phonemic Transformation Effect* (PTE; which itself is closely related to the Verbal Transformation Effect). While originally interested in measuring listeners' abilities to discriminate between different arrangements of repeated vowels, Warren, Bashford and Gardner (1990) reported an interesting perceptual effect. For repeated sequences of steady-state vowels of 30 to 100 ms in duration, listeners experienced *phonemic transformations* into syllables, words and pseudowords. These included illusory consonants, which were not present in the signal itself and which were not heard at slow sequence rates. Additionally, different verbal organisations were heard for different permutations of the same vowels, which allowed participants to discriminate between the different orders. Warren, Bashford and Gardner (1990) argued that such transformations were possible as they occurred for rates at which the particular length of the repeated sequences matched the speech templates involved in recognition of verbal organisations such as syllables or words. This process was facilitated by the extraction of appropriate spectral components for a given syllable or word where – according to Warren, Bashford and Gardner (1990) – participants perceptually match the repeating vowel sequence to a particular verbal form. Upon repetition, the signal undergoes perceptual separation into two fractions: one is matched to the template corresponding to a syllable or word that is reported. The other corresponds to the residue or the components left over after the match has been made and this manifests itself as a non-linguistic 'noise' or a second, less salient, voice occurring at the same time as the first one. Hence, repeating a single sequence of vowels usually results in listeners hearing two different voices.

Chalikia and Warren (1991) looked more closely at the two separate verbal organisations or voices that seem to be reported during the perceptual regrouping of a given vowel sequence. They confirmed that one of those always included a verbal form while the other was either nonverbal "noise" or a secondary (less salient) verbal form. Chalikia and Warren (1991) asked participants to listen to vowel sequences (each included eight 80 ms long vowels) until they could identify two verbal organisations and to report which one was more salient. They found that all listeners could perform the task and that forms reported were syllables, words or pseudowords that followed the phonotactic rules of English. In addition, forms reported which were more salient were usually longer and they differed from the second voice in timbre, loudness and speed of enunciation. As a result, Chalikia and Warren (1991) suggested that participants must be using different spectral regions of a recycled sequence to produce the two reported forms. On occasions when two simultaneous organisations were heard, listeners could differentiate between each speaker's voice and between the phonetic content of the two percepts. Chalikia and Warren suggested two possible causes for the streaming of the sequences into two percepts. They argued that the original sequences lacked the cohesive force ("perceptual glue") provided by formant transitions, which normally prevent streaming of vowels from occurring (cf. Cole and Scott, 1973; Dorman et al., 1975). Also, they pointed out that, in general, it is repetition which drives the tendency for phonetic elements to segregate. As mentioned before, with regards to the difference between the two organisations, the authors argued that the less salient form is the result of a residue, or "leftover" spectral components from the dominant one. These would either match to another syllabic template (heard as a secondary form) or be reported as non-linguistic "noise".

Given that one voice was often reported as ‘lower’ and the other as ‘higher’ (implied pitches, based on differences in vowel brightness), Chalikia and Warren (1991) speculated that these voices can be mapped onto the properties of the formants included in these organisations. For the lower voice, this would include adjacent lower formants and vice versa for the higher voice. This spectral separation between the voices was further emphasized by the finding that the phonemes contained in the primary and secondary responses did not overlap. The vowels used in the Chalikia and Warren study varied from high front [iː] (as in ‘heat’) to low back [ɒ] (as in ‘heart’). In summary, F1 frequency is inversely proportional to vowel height, while F2 frequency is proportional to vowel frontedness (see Methodology for further details). Some vowels were more likely to be reported in the more dominant stream/voice, others in the less salient one. That was presumably influenced by the distribution of the energy differences between the vowels, most notably in formant frequencies. As the formants of consecutive vowels are not physically connected by formant transitions, and have different F1 and F2 frequencies, they group mostly based on timbre differences if synthesized on the same F0 frequency.

In a further investigation of the basis for this grouping, Chalikia and Warren (1994) demonstrated explicitly that the two organisations or voices can be separated into two separate spectral regions. Listeners were exposed to a number of repeating sequences made up of ten 60-ms vowels. Their first task was to identify what the voice, or voices were saying. If participants experienced hearing two forms at the same time they indicated which one was more salient by reporting it first. In the following session, they were asked to isolate the two voices by using the low-pass and high-pass frequency filter (controlled by a frequency knob). Chalikia and Warren (1994) found that the spectral bands used by listeners fell roughly (there were individual differences) into two regions, one for components below 1500 Hz and one for components above 1500 Hz. The authors argued that these two regions have been shown in previous research to divide speech into high-pass and low-pass ranges of equal intelligibility. Additionally, Warren, Healy and Chalikia (1996) showed that different listeners typically experience the same or very similar initial percepts (in phonetic structure) and that these verbal organisations are stable over time.

The occurrence of illusory consonants upon repetition of sounds has been demonstrated in previous research. In Darwin and Bethell-Fox’s (1977) study, a repeating diphthong broke into two different voices when it alternated abruptly between two different F0 frequencies, and this segregation also produced an illusory consonant ‘g’ being reported on one of these voices (see Introduction, p. 23). As those illusory percepts were not present in the original stimuli their occurrence at least in part can be explained in terms of perceptual regrouping. Listeners interpret the rapid changes in the stimuli as having consonants in them which shows evidence of perceptual streaming as described by Bregman and Campbell (1971). It would therefore be reasonable to assume that the VTE and PTE demonstrate similar mechanisms when it comes to the perceptual regrouping of sounds. While the original studies on PTE concentrated on the very first percept and the distinction between the two voices, those studies have not looked at the subsequent changes in the stimulus – verbal transformations which occur if the sequences are repeated for a sufficiently long time. This will be addressed in the following two experiments. Similar to earlier studies in this thesis, other constraints on the organisation of the repeated sounds will be investigated in terms of the PTE. These are F0 and ITD manipulations, neither of which has yet been explored in the context of the PTE. These cues will be introduced with the aim of opposing the "default" grouping which arises from timbre (spectral) differences when a sequence of vowels on the same F0 is presented to listeners. Timbre is a multidimensional property and differences in timbre can be brought about in many ways. For our purposes, the timbral difference between two vowels will be identified by the differences between the positions of the formant frequencies of the vowels as they are described by tongue position on the two standard dimensions (high/low and front/back). For example, the high front vowel [iː] (as in 'heat') will have a similar timbre to the low front vowel [ae] (as in 'hat') with which it shares the property of being a front vowel, but it will have different timbre to the low back vowel [ɒ] (as in 'heart') with which it shares neither height nor frontedness.

4.2 Experiment 5 – PTE and F0 cues

In this experiment, repeating sequences of four vowels were used. In a given sequence, if one pair of vowels is presented on a sufficiently different F0 from the others, it is expected that this pair will tend to separate from the others to form a separate stream. This in turn will change the type of verbal organisations that are likely to occur, relative to the case where all four vowels share a common F0. The experimental sequence will be presented diotically but one pair of vowels will be synthesised on a different F0.

The process of producing the stimuli for the current experiment is described in relation to the manipulations performed for Experiment 3 – Formant Transitions. A single glottal pulse, excised from a recording of natural speech, was iterated several times to produce a steady-state vowel on a given F0, and a sequence of four such different vowels was played in a repeating cycle to the listeners. The types of manipulations described above are in principle similar to the work of Bregman et al. (1990). They investigated judgments of the ability to pick out temporal-order patterns from a repeating cycle of four complex tones with different F0 frequencies and timbres (frequencies of single spectral peaks). By varying how much physical difference there was on the different dimensions (either in F0 or in timbre), and measuring the circumstances under which listeners' responses were driven primarily by the formant-frequency differences or by the F0 differences, they showed that one type of organisation might dominate another. Bregman et al. showed that both factors influence stream segregation and the grouping that is heard depends on which of the two factors leads to the greater perceived dissimilarity between the tones. Even when the spectra covered the same frequency range, ΔF0 was an important grouping factor; however, formant-frequency separation became more dominant with increasing sharpness of the formant peak (amplitude of the peak relative to a spectral pedestal).

In the present study, sequences of four vowels - [iː], [uː], [ae], and [ɒ] as in the words: 'heat', 'hoot', 'hat' and 'heart' - were chosen such that the default grouping (according to spectral similarity) would be the [iː] and [ae] vowels in one stream and the [uː] and [ɒ] in the other. In the absence of other differences, transformations in this case are expected to group according to timbre as defined by the differences in the tongue positions (see Figure 4.1). Hence it should be possible to distinguish percepts which share phonetic characteristics more similar to the high front vowel [iː] (as in 'heat') or the low back [ɒ] (as in 'heart'). In other words, timbre-based separation should be based on the frequencies of the second formant (F2) of the vowels, such that the pair of vowels [iː] and [ae] (with relatively high F2 frequency) will be different to the pair of vowels [uː] and [ɒ] (with relatively low F2 frequency), as shown in Figure 4.3.

F0 cues were used either to 'support' the timbral cue based on frontedness of the vowel (F2 similarity) with [iː] and [ae] synthesized on one pitch, e.g. high, and [uː] and [ɒ] on another, e.g. low, or to favour an alternative grouping by common vowel height (F1 similarity) with [iː] & [uː] presented on one pitch e.g. high, and the [ae] & [ɒ] on another pitch, e.g. low. For the alternative arrangement, if grouping occurs on the basis of pitch, listeners will report different transformations to the ones heard in a 'supported' sequence (additionally, illusory consonants would be prediceted to group with vowels which resemble their spectral properties; cf. Warren, Healy and Chalikia, 1996). Furthermore, as the two cues (timbre and pitch) will be in competition with each other, hence offering more possibilities for perceptual re-grouping, it was hypothesised that listeners would be predicted to report more VTs and forms in the alternative arrangement rather than when the F0 cues supported the timbral cues.

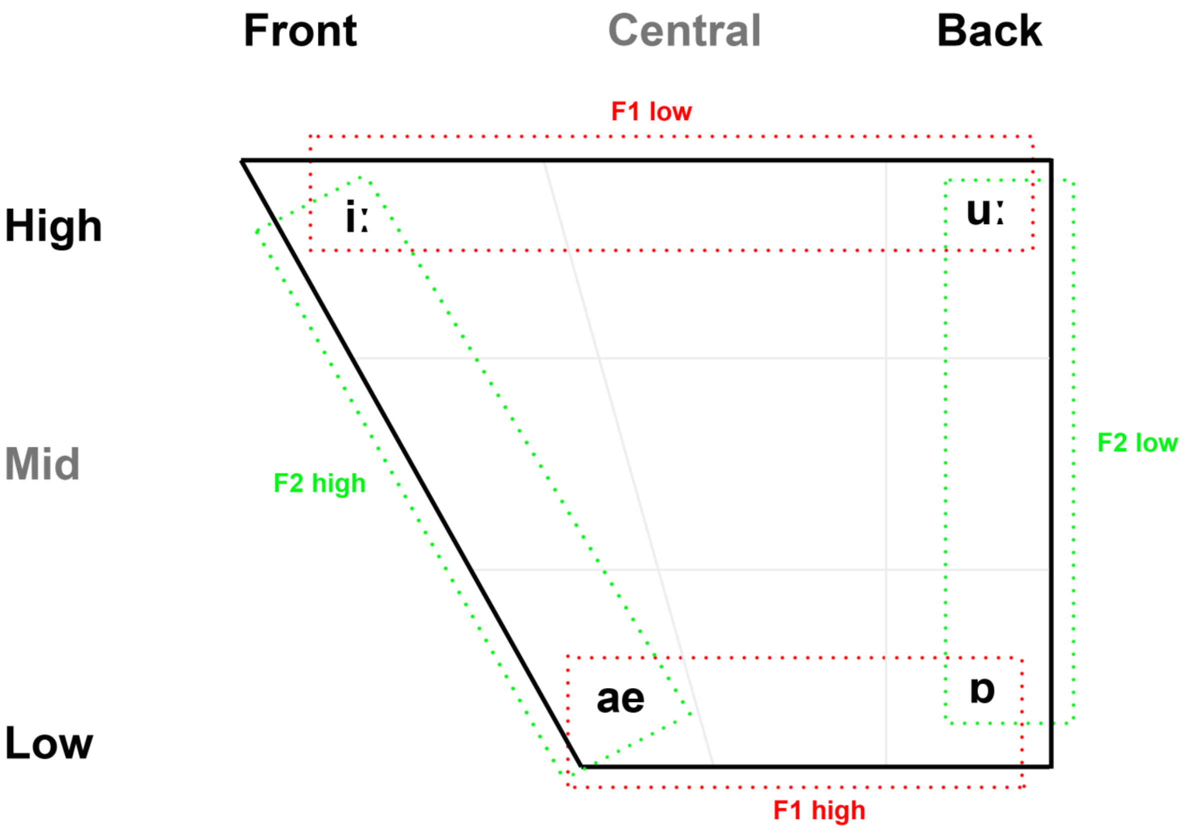


Figure 4.1 Vowel quadrilateral showing the vowels used in the study and their respective tongue positions: high front [iː], low front [ae], high back [uː], and low back [ɒ]. Also shown is the relationship between the relative frequencies of the first two formants (F1 and F2) and the four vowels: [iː] has low F1 and high F2, [ae] has high F1 and high F2, [uː] has low F1 and low F2 and [ɒ] has high F1 and low F2.

4.2.1 Method

Participants

Twelve listeners (6 males, 6 females) took part in the experiment. They were all native speakers of English and reported normal hearing. At the end of the study, they were either paid cash or received course credit. The listeners' mean age was 20.1 years old (s.d. = 3.12).

Stimuli and Conditions

For stimulus creation, individual glottal pulses, each starting and ending at zero-crossings, were selected such that the pulse peak followed immediately after the zero-crossing (see Figure 4.2). The pulses were extracted from recordings of four vowels, which were [iː], [uː], [ae], and [ɒ] as in the words: 'heat', 'hoot', 'hat' and 'heart'. For a given pitch, the pulses were excised individually from high quality recordings of BKB sentences (Bench, Kowal and Bamford, 1979), which were monotonised first to the required F0 frequency. Therefore, the selected glottal pulses were precisely the length of the period corresponding to the required F0 frequency. For example, when a word was monotonised at 120 Hz, the excised pulse was 8.35 ms long; for F0=170 Hz, it was 5.89 ms.

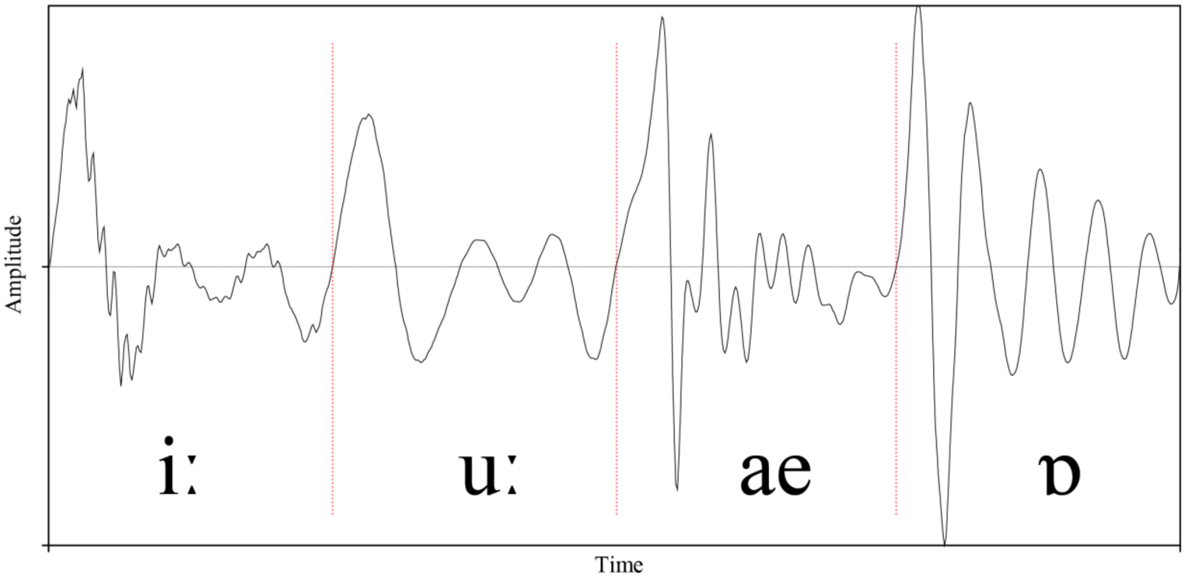


Figure 4.2 Four excised glottal pulses used to create the stimuli. Each one is shown separated by dotted red lines at the zero-crossings. Note the pulse peaks following shortly after the zero-crossings.

The excised vowels represent the four corners of the vowel quadrilateral: high front [iː], high back [uː], low front [ae], and low back [ɒ]. The first formant (F1) is inversely proportional to height, meaning that high vowels have a low F1 frequency, whereas low vowels (low tongue position) have a high F1. The second formant (F2), on the other hand, is proportionally related to frontedness - front vowels have high F2s and back vowels have low F2s. In terms of the overall spectral similarity, the natural pairing of vowels is by front vs. back because spectrally the front vowels have relatively large separations between F1 and F2 whereas for the back vowels F1 and F2 are relatively close together. That implies that for the "natural" pairing of front vowels [iː] and [ae] F1 and F2 are relatively far apart, whereas for the "natural" pairing of back vowels [uː] and [ɒ] F1 and F2 are relatively close to one another (see Figure 4.3). Please refer to Table 4.1 for the first three formant-frequencies for each of the four vowels.

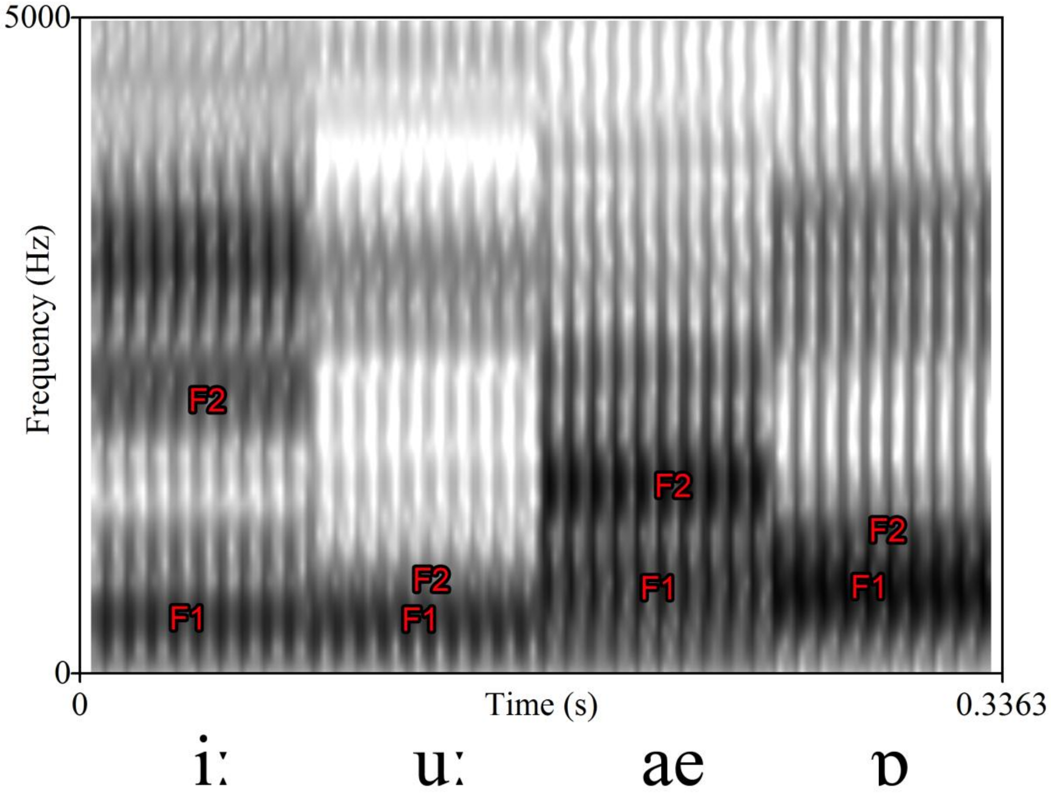


Figure 4.3 Spectrograms, highlighting the relative distance between F1 and F2 for the 4 vowels used in the study.

The experiment used sequences of four vowels. Two sequence permutations of the four vowels were used. If [iː] is 1, [uː] is 2, [ae] is 3 and [ɒ] is 4, **sequence 1** was: 1-2-3-4 and

sequence 2 was 2-4-1-3. Baseline pairings (based on the frontedness of the vowels) were of non-consecutive vowels in sequence 1, but of consecutive vowels in sequence 2. No acoustical mixing or transitional stages (e.g., amplitude ramps) from one vowel to the next were used. By building the stimuli from glottal pulses spliced out at zero crossings, transient-associated clicks were avoided.

Table 4.1 Formant frequency values (in Hz) for the four vowels used in Experiments 5 and 6.

Vowel	F1	F2	F3
[iː] as in 'heat'	386	2137	3041
[uː] as in 'hoot'	344	660	2720
[ae] as in 'hat'	685	1463	2280
[ɒ] as in 'heart'	601	962	2728

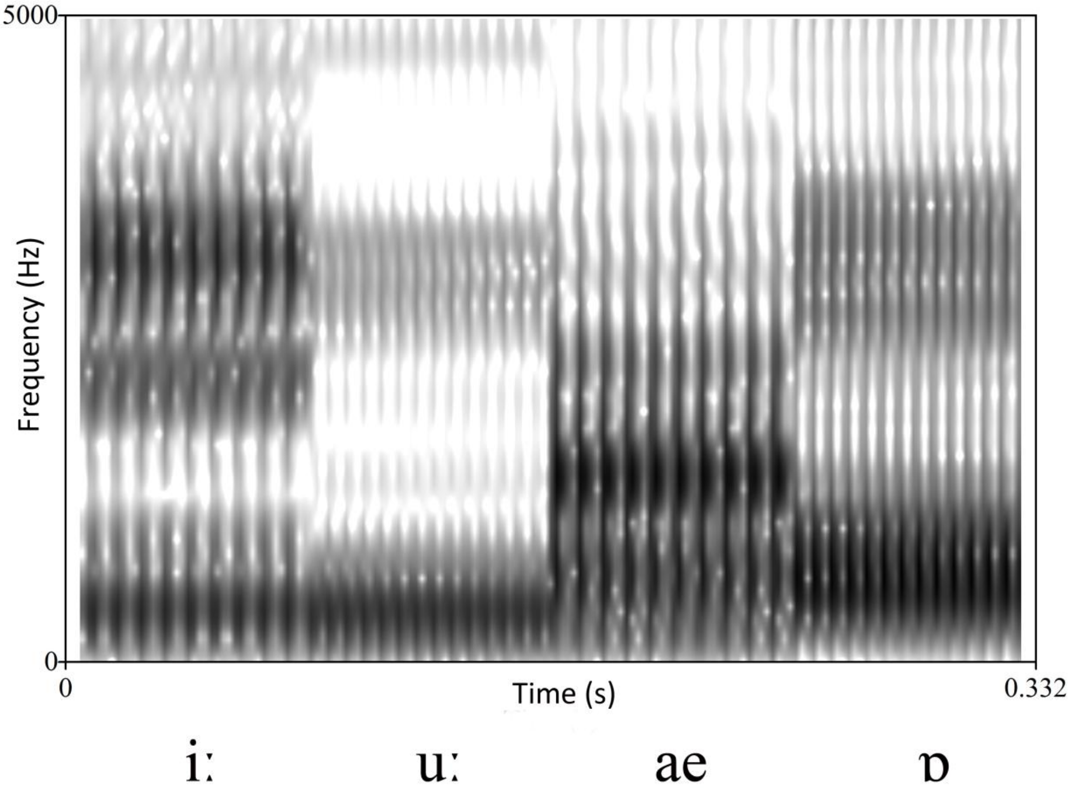


Figure 4.4 Example of a 4-vowel sequence (E1) used in Experiment 5. Vowels 1 and 3 are on F0=120 Hz and vowels 2 and 4 on F0=170 Hz. Notice the use of 10 iterations of the glottal pulse for the lower F0 and 14 iterations for the higher one, ensuring roughly equal durations for the individual vowels.

The vowels were presented either on a low (120 Hz) or on a high (170 Hz, 6 semitones higher) F0 frequency. As noted above, original recordings of BKB words including the four vowels [iː], [uː], [ae], and [ɒ] were first monotonised to the required F0 frequency before the glottal pulses were spliced out. As a result, the duration of each glottal pulse corresponded to the required F0 frequency. For F0=120 Hz, each pulse was iterated 10 times (10 x 8.35 ms); hence the four-vowel sequence was 334 ms long. For F0=170 Hz, each pulse was iterated 14 times (14 x 5.89 ms), resulting in a sequence duration of 329.84 ms. See Figure 4.4 for a wideband spectrogram of an example of one of the four-vowel sequences used; the different F0 frequencies can be seen (visible change in pulse duration). Note that the same excised glottal pulses that were used for the low-F0 vowels in this study were also used to make all of the stimuli in Experiment 6.

There were six experimental conditions in the study; all were presented diotically. For the labelling of the conditions, *TP* stands for *tongue position* and *F0* for the *pitch* hence an example combination *High(TP)Low(F0)/Low(TP)High(F0)* can be read as: vowels characterised by a high tongue position were synthesized on the low pitch whereas vowels characterised by a low tongue position were synthesized on the high pitch.

The first two conditions included all vowels synthesized on the lower F0 – *All Low(F0)*, or all vowels synthesized on the higher F0 – *All High(F0)*. As a pair, these were referred to as *baseline* conditions. Given that there were no differences in F0 to distinguish subsets of vowels in these two sequences, any perceptual regroupings arising from this arrangement will be attributed to the differences in timbre – i.e., when asked to report the two percepts while listening to the repeating sequence of vowels, it is expected that participants will associate transformations on the higher voice (sounding more 'bright') with front vowels [iː] & [ae] and transformations on the lower voice (sounding more 'dull') with the back vowels [uː] & [ɒ].

The next two conditions, described as *Front(TP)Low(F0)/Back(TP)High(F0)* and *Front(TP)High(F0)/Back(TP)Low(F0)*, are referred to as *congruent* conditions. In these cases, the F0 cues 'supported' the timbral cue – in other words, what listeners experienced as the higher and lower voices (timbre difference) was matched by an F0 cue which should act in concert to separate those two voices on different pitches. Arguably, this arrangement should result in a similar pattern of responses to the case where there is no F0 difference, i.e. to conditions *All Low(F0)* and *All High(F0)*. In other words, it is expected that the average number of VTs and forms for *baseline* and *congruent* conditions will not be significantly different.

In condition *High(TP)Low(F0)/Low(TP)High(F0)* and *High(TP)High(F0)/Low(TP)Low(F0)*, an alternative pairing of items (instead of back-back, front-front) was introduced using F0 cues. While vowels [iː] & [ae], and [uː] & [ɒ] were paired in terms of timbre, in condition *High(TP)Low(F0)/Low(TP)High(F0)*, pair [iː] & [uː] was synthesised on the lower F0 and pair [ae] & [ɒ] was synthesised on the higher F0. For condition *High(TP)High(F0)/Low(TP)Low(F0)*, this was reversed - i.e., [iː] & [uː] were synthesised on the higher F0 and [ae] & [ɒ] on the lower F0. The above two conditions can be considered to be *opposing*, because the addition of the F0 cue opposes the timbral cues for streaming. Compared to the first two *baseline* conditions, participants can now group and report the two percepts – higher and lower voice – based on the F0 frequency of the vowels, where [iː] was on the same F0 as [uː] and [ae] on the same F0 as [ɒ] (rather than based on the vowel frontedness). Hence, an effect of the F0 cue on perceptual re-grouping in the four-vowel sequence should lead to changes in the verbal forms reported. The change should be observed in more VTs and forms being reported in *opposing* conditions compared to *baseline* (with more opportunities of re-grouping in the opposing case) as well as qualitative difference in the phonetic structure of the VTs and forms between the two condition groups. For the summary of all conditions, see Table 4.2.

Table 4.2 The conditions and associated vowels arrangements in Experiment 5. For the F0 factor, 'low' refers to F0=120 Hz and 'high' to F0=170 Hz.

Cue type	Condition	Vowel arrangement
Baseline (natural timbre only)	***All Low(F0)*** ***All High(F0)***	All 4 heard on low F0 All 4 heard on high F0

Congruent	***Front(TP)Low(F0)/Back(TP)High(F0)***
	Front(TP)High(F0)/Back(TP)Low(F0)
		[iː] & [ae] on low F0, [uː] & [ɒ] on high F0
		[iː] & [ae] on high F0, [uː] & [ɒ] on low F0

Opposing	***High(TP)Low(F0)/Low(TP)High(F0)***
	High(TP)High(F0)/Low(TP)Low(F0)
		[iː] & [uː] on low F0, [ae] & [ɒ] on high F0
		[iː] & [uː] on high F0, [ae] & [ɒ] on low F0

Participants attended two sessions (on separate days) and were exposed to either 4 or 8 trials, each of which included a sequence of short vowels recycled for 3 minutes. For each condition, there were two 3-minute presentations (*All Low(F0) seq*1 and *All Low(F0) seq*2, *All High(F0) seq*1 and *All High(F0) seq*2, and so on); one for each of the two sequence permutations. Within each session, the order of 3-minute sequences was randomised.

Participants were assigned to either the ‘Odds’ or ‘Evens’ group and attended two sessions - either with conditions *All Low(F0)* and *All High(F0)* in the first session and the other four conditions in the second, or vice versa. During both sessions, while reporting the transformations, listeners had to indicate (using key presses) whether a given response belonged to the higher or the lower voice.

4.2.1 Results

The focus of the analysis for the following two experiments will be on the distinction between the two sequence permutations of the vowels used and condition type, which can be summarised as per the description above: baseline, opposite and congruent. Given that there are a greater number of prominent spectral discontinuities between neighbouring vowel segments in sequence 1, owing to the pattern of formant frequencies (sequence = front-back-front-back), it is plausible that this sequence will produce more VTs and more new forms. This will be compared with sequence 2 (front-front-back-back), where more prominent timbral differences occur between adjacent *pairs* of vowels rather than between adjacent vowels. More critically, however, the relationship between the opposite and congruent conditions can inform us about the relative contribution of the F0 manipulations (and ITD cues in the next experiment). Given that there is evidence from previous experiments in this thesis that no additional information is conveyed by the time to first response, and that the PTE illusory consonants are known to appear almost immediately (e.g., Chalikia & Warren, 1991), data on time to first response are only considered in terms of a descriptive analysis (the same is true for Experiment 6).

Verbal Transformations

The three factors manipulated in the following analyses were *condition*, *sequence permutation* (the term *permutation* will be used from now on), and *voice*. Factor condition includes stimulus manipulations from *All Low(F0)* to *Front(TP)High(F0)/Back(TP)Low(F0)*, as specified in the methods section. They can be broadly categorized into three groups: baseline (*All Low(F0)* and *All High(F0)*), congruent (*Front(TP)Low(F0)/Back(TP)High(F0)* and

Front(TP)High(F0)/Back(TP)Low(F0)), and opposing (*High(TP)Low(F0)/Low(TP)High(F0)* and *High(TP)High(F0)/Low(TP)Low(F0)*). Each condition will be presented using two sequence permutations: 1 and 2. The last factor – voice, relates to listeners' responses about what they heard in the different conditions. Warren, Healy, and Chalikia (1996) refer to the separation of the two simultaneous voices based on the spectral ranges of the vowels used as the *high* and *low* voices. In the present experiment, however, the distinction needs to be made between the baseline and the remaining conditions. As there was no manipulation of F0 for the baseline conditions (*All Low(F0)* and *All High(F0)*), participants responded to either the *bright*-timbre voice or *dull*-timbre voice. For the opposing and congruent conditions, this was referred to as the *high-F0* or *low-F0* voice, as the sequences included F0 manipulations. Note that, although the assumption here is that participants will direct their attention to voices on a particular F0, it is nonetheless possible that their responses will also be influenced by the timbral differences. Table 4.3 shows mean values for VTs heard in 3 minutes for each condition, sequence permutation, and voice.

Unless stated otherwise, all ANOVA summary tables are collated in the Appendix 5 (for the current and the next study) in the order in which the various results are described.

One purpose of the baseline conditions, *All Low(F0)* and *All High(F0)*, was to test whether participants would hear more VTs on a particular voice, as described by the bright or dull timbre. A three-way ANOVA with permutation (sequence 1, sequence 2), condition (*All Low(F0)*, *All High(F0)*) and voice (bright, dull) indicated that this was not the case. The three main effects and all the interaction terms were found not to be significant. Approaching significance (p=.06) was the permutation x condition interaction, which was driven by the fact that for condition *All Low(F0)* there were more VTs reported for sequence permutation 2 (means: 7.00 vs 7.71) while the opposite was true for condition *All High(F0)* (means: 7.50 vs 7.33). However, this was of little consequence, as in general the F0 of a sequence did not have any effect on the number of VTs reported. It would, therefore, be reasonable to assume that any effects falling out of the subsequent analyses were driven by differences in the conditions manipulated and not by the absolute value of F0.

Table 4.3 Average number of VTs reported in Experiment 5 across all conditions. For the baseline conditions, listeners classified each VT as spoken either by the bright voice or the dull voice. For the opposing and congruent conditions, each VT was classified as either on the high pitch or the low pitch.

	Condition	**Average no. of VTs reported in 3min** *(±SE)*		
		Cumulated	**Bright Voice**	**Dull Voice**
Baseline	***All Low(F0)*** **seq1**	14.00 *(3.19)*	6.67 *(1.56)*	7.33 *(2.13)*
	All Low(F0) **seq2**	15.42 *(3.26)*	9.58 *(2.34)*	5.83 *(1.47)*
	***All High(F0) seq*1**	15.00 *(3.43)*	7.83 *(1.24)*	7.17 *(2.50)*
	***All High(F0) seq*2**	14.67 *(3.61)*	8.08 *(2.12)*	6.58 *(1.71)*
			High F0	**Low F0**
Congruent	***Front(TP)Low(F0)/Back(TP)High(F0) seq*1**	15.83 *(2.87)*	8.75 *(1.90)*	7.08 *(1.33)*
	***Front(TP)Low(F0)/Back(TP)High(F0) seq*2**	16.00 *(3.40)*	8.50 *(2.42)*	7.50 *(1.43)*
	Front(TP)High(F0)/Back(TP)Low(F0) **seq1**	16.08 *(2.42)*	7.83 *(1.66)*	8.25 *(1.36)*
	Front(TP)High(F0)/Back(TP)Low(F0) **seq2**	17.42 *(3.43)*	9.67 *(2.97)*	7.75 *(1.38)*
Opposing	***High(TP)Low(F0)/Low(TP)High(F0) seq*1**	15.92 *(3.03)*	6.50 *(1.37)*	9.42 *(2.03)*
	***High(TP)Low(F0)/Low(TP)High(F0) seq*2**	15.42 *(3.23)*	8.25 *(2.47)*	7.17 *(1.17)*
	***High(TP)High(F0)/Low(TP)Low(F0) seq*1**	16.17 *(3.64)*	9.33 *(2.28)*	6.83 *(1.55)*
	***High(TP)High(F0)/Low(TP)Low(F0) seq*2**	15.00 *(2.83)*	7.75 *(2.43)*	7.25 *(1.44)*

To investigate the pattern within each condition type (baseline, opposing, and congruent), the same three-way ANOVAs were performed for conditions *High(TP)Low(F0)/Low(TP) High(F0)* & *High(TP)High(F0)/Low(TP)Low(F0)*, and for conditions *Front(TP)Low(F0)/ Back(TP)High(F0)* & *Front(TP)High(F0)/Back(TP)Low(F0)*. This approach could potentially highlight the effects of voice and the way in which it might interact with sequence. Both analyses, however, did not yield any significant results. It remains to be seen whether this outcome was a general lack of effect of the F0 manipulation between the sequences or whether, as has been shown in previous experiments, that Forms is a much more sensitive measure and can reveal effects that might not be apparent from the number of VTs.

To evaluate whether there were differences between the opposing and congruent conditions, a superordinate two-way ANOVA with condition (opposing [*High(TP)Low(F0)/Low(TP) High(F0)* + *High(TP)High(F0)/Low(TP)Low(F0)* collapsed], and congruent [*Front(TP) Low(F0)/Back(TP)High(F0)* + *Front(TP)High(F0)/Back(TP)Low(F0)* collapsed]) and permutation (sequence 1 and sequence 2) was performed. Neither the crucial interaction nor the main effects were significant. As none of the conditions produced a meaningful change in the number of VTs, it was concluded that neither the introduction of an F0 difference between pairs of vowels nor the congruence of this manipulation with the baseline timbre differences affected the number of VTs reported.

Forms

As for VTs, the same set of analyses was performed for the number of forms. Again, for the baseline and congruent conditions, no significant effects were observed. There was, however, a significant main effect of permutation [F(1,11)=6.87, p=.02, η^2=0.38] for the opposing conditions (*High(TP)Low(F0)/Low(TP)High(F0)* & *High(TP)High(F0)/Low(TP)Low(F0)*). According to this, listeners tended to report more Forms for sequence 1 (11.50 Forms per 3 min) compared to sequence 2 (9.29). This result was broadly consistent with what was a non-significant trend for fewer VTs to be heard for sequence 2 in the opposing conditions (see Table 4.4). Presumably, the emergence of a significant main effect of permutation for conditions *High(TP)Low(F0)/Low(TP)High(F0)* and *High(TP)High(F0)/Low(TP)Low(F0)* is because the F0 difference opposes the pairing that would otherwise be dictated by the voice timbre.

The means for the superordinate two-way ANOVA for Forms are presented for all conditions in Table 4.5. This analysis revealed only a significant mean effect of sequence permutation. Although this is driven primarily by the opposing condition, a similar trend is apparent for the baseline and congruent conditions. The overall tendency to hear more forms for sequence 1 is consistent with original prediction about spectral discontinuities between vowel tokens.

Table 4.4 Average number of Forms reported in Experiment 5 across all conditions. For the baseline conditions, listeners classified each VT as spoken either by a bright voice or a dull voice. For the opposing and congruent conditions, each VT was classified as either on the high pitch or the low pitch.

	Condition	**Average no. of Forms reported in 3min** *(±SE)*		
		Cumulated	**Bright Voice**	**Dull Voice**
Baseline	***All Low(F0)*** **seq1**	10.08 *(1.93)*	5.08 *(1.21)*	5.00 *(1.09)*
	All Low(F0) **seq2**	9.50 *(1.26)*	5.00 *(0.83)*	4.50 *(0.78)*
	***All High(F0) seq*1**	10.08 *(1.15)*	5.67 *(0.67)*	4.42 *(0.84)*
	***All High(F0) seq*2**	9.83 *(1.77)*	4.92 *(0.95)*	4.92 *(1.03)*
			High F0	**Low F0**
Congruent	***Front(TP)Low(F0)/Back(TP)High(F0) seq*1**	10.92 *(1.46)*	6.00 *(1.02)*	4.92 *(0.71)*
	***Front(TP)Low(F0)/Back(TP)High(F0) seq*2**	10.83 *(1.53)*	5.17 *(0.93)*	5.67 *(0.80)*
	Front(TP)High(F0)/Back(TP)Low(F0) **seq1**	11.75 *(1.74)*	5.58 *(1.18)*	6.17 *(0.90)*
	Front(TP)High(F0)/Back(TP)Low(F0) **seq2**	9.92 *(1.51)*	5.25 *(1.03)*	4.67 *(0.62)*
Opposing	***High(TP)Low(F0)/Low(TP)High(F0)*** **seq1**	11.50 *(1.80)*	5.17 *(1.04)*	6.33 *(0.97)*
	High(TP)Low(F0)/Low(TP)High(F0) **seq2**	9.67 *(1.32)*	4.83 *(0.83)*	4.83 *(0.75)*
	***High(TP)High(F0)/Low(TP)Low(F0) seq*1**	11.50 *(2.02)*	6.08 *(1.08)*	5.42 *(1.15)*
	***High(TP)High(F0)/Low(TP)Low(F0) seq*2**	8.92 *(1.23)*	4.25 *(1.21)*	4.67 *(0.61)*

Table 4.5 Mean relation for the two factors from condition x permutation for Forms in Experiment 5. The values given on the right and at the bottom are collapsed across permutation and condition type, respectively.

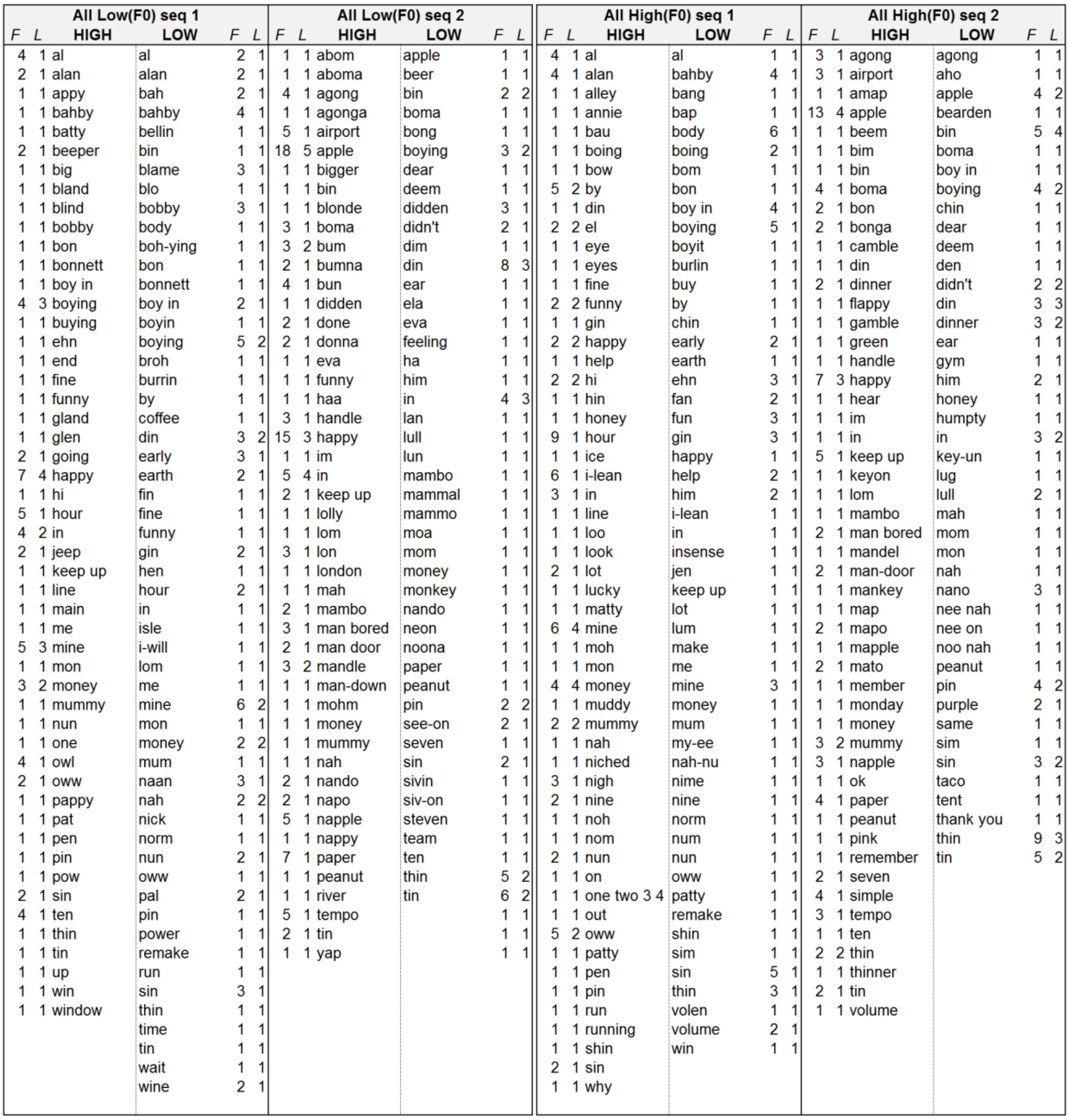
		Permutation		
		Seq. 1	Seq. 2	
Condition	Baseline	10.08 *(1.48)*	9.67 *(1.50)*	9.88 *(1.41)*
	Opposing	11.50 *(1.85)*	9.29 *(1.18)*	10.40 *(1.50)*
	Congruent	11.33 *(1.52)*	10.38 *(1.44)*	10.85 *(1.44)*
		10.97 *(1.53)*	9.78 *(1.31)*	

4.2.2 Additional analyses

As the quantitative analyses did not indicate any striking effects on the number of VTs or forms of the difference between opposing and congruent conditions, this relationship was explored further with a more descriptive approach. To facilitate this investigation, all instances of a single response made by only one participant were removed for the analysis. Hence, the criterion for including an entry was that it was heard more than once by at least one listener or once by at least two listeners. This was done to reduce noise in the data (more than 60% of data points were removed this way) and obtain a clearer picture of the underlying patterns, more specifically the extent to which the opposing and congruent cues influenced the forms heard for the two sequence permutations used. This approach also allows an exploration of the regions of overlap between the different groups of conditions in which there were common responses, either between a particular pair (e.g. opposing vs. congruent) or for all three manipulations. This idea was explored through various adaptations of Venn diagrams.

Comparison of responses to sequences 1 and 2

Considered first are the responses in the two baseline conditions (*All Low(F0)* and *All High(F0)*), for which only timbral grouping cues were present (see Figure 4.5; for the list of all responses in each condition in Experiment 5 refer to Appendices 6.1 – 6.3). For both sequences in both conditions, the responses included words and pseudowords, all of which adhered to the rules of English grammar. For the vast majority of verbal forms heard, the illusory phonetic segments were interpreted as nasals, stops or plosives, with noticeably fewer fricative sounds. This was true for the first responses as well as for the subsequent VTs. A similar pattern has been reported in previous research, notably by Chalikia and Warren (1991). Responses in conditions *All Low(F0)* and *All High(F0)* were not greatly affected by the difference in F0 between them. Within each condition, the forms reported for both voices were phonetically similar, yet there appeared to be a distinction between the two sequences in terms of the volume of different forms reported.



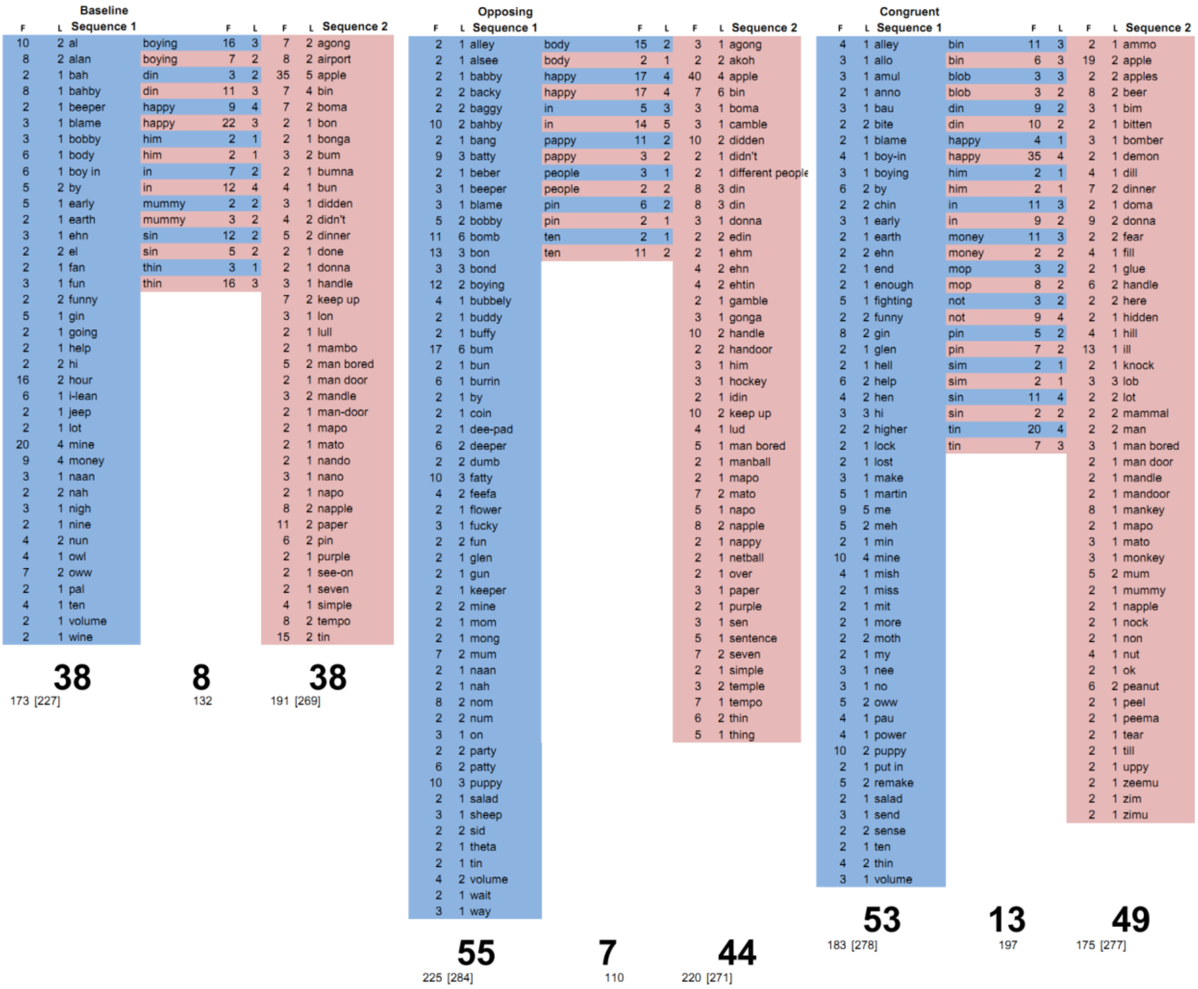
All Low(F0) seq 1					
F	*L*	**HIGH**	**LOW**	*F*	*L*
4	1	al	al	2	1
2	1	alan	alan	2	1
1	1	appy	bah	2	1
1	1	bahby	bahby	4	1
1	1	batty	bellin	1	1
2	1	beeper	bin	1	1
1	1	big	blame	3	1
1	1	bland	blo	1	1
1	1	blind	bobby	3	1
1	1	bobby	body	1	1
1	1	bon	boh-ying	1	1
1	1	bonnett	bon	1	1
1	1	boy in	bonnett	1	1
4	3	boying	boy in	2	1
1	1	buying	boyin	1	1
1	1	ehn	boying	5	2
1	1	end	broh	1	1
1	1	fine	burrin	1	1
1	1	funny	by	1	1
1	1	gland	coffee	1	1
1	1	glen	din	3	2
2	1	going	early	3	1
7	4	happy	earth	2	1
1	1	hi	fin	1	1
5	1	hour	fine	1	1
4	2	in	funny	1	1
2	1	jeep	gin	2	1
1	1	keep up	hen	1	1
1	1	line	hour	2	1
1	1	main	in	1	1
1	1	me	isle	1	1
5	3	mine	i-will	1	1
1	1	mon	lom	1	1
3	2	money	me	1	1
1	1	mummy	mine	6	2
1	1	nun	mon	1	1
1	1	one	money	2	2
4	1	owl	mum	1	1
2	1	oww	naan	3	1
1	1	pappy	nah	2	2
1	1	pat	nick	1	1
1	1	pen	norm	1	1
1	1	pin	nun	2	1
1	1	pow	oww	1	1
2	1	sin	pal	2	1
4	1	ten	pin	1	1
1	1	thin	power	1	1
1	1	tin	remake	1	1
1	1	up	run	1	1
1	1	win	sin	3	1
1	1	window	thin	1	1
			time	1	1
			tin	1	1
			wait	1	1
			wine	2	1

All Low(F0) seq 2					
F	*L*	**HIGH**	**LOW**	*F*	*L*
1	1	abom	apple	1	1
1	1	aboma	beer	1	1
4	1	agong	bin	2	2
1	1	agonga	boma	1	1
5	1	airport	bong	1	1
18	5	apple	boying	3	2
1	1	bigger	dear	1	1
1	1	bin	deem	1	1
1	1	blonde	didden	3	1
3	1	boma	didn't	2	1
3	2	bum	dim	1	1
2	1	bumna	din	8	3
4	1	bun	ear	1	1
1	1	didden	ela	1	1
2	1	done	eva	1	1
2	1	donna	feeling	1	1
1	1	eva	ha	1	1
1	1	funny	him	1	1
1	1	haa	in	4	3
3	1	handle	lan	1	1
15	3	happy	lull	1	1
1	1	im	lun	1	1
5	4	in	mambo	1	1
2	1	keep up	mammal	1	1
1	1	lolly	mammo	1	1
1	1	lom	moa	1	1
3	1	lon	mom	1	1
1	1	london	money	1	1
1	1	mah	monkey	1	1
2	1	mambo	nando	1	1
3	1	man bored	neon	1	1
2	1	man door	noona	1	1
3	2	mandle	paper	1	1
1	1	man-down	peanut	1	1
1	1	mohm	pin	2	2
1	1	money	see-on	2	1
1	1	mummy	seven	1	1
1	1	nah	sin	2	1
2	1	nando	sivin	1	1
2	1	napo	siv-on	1	1
5	1	napple	steven	1	1
1	1	nappy	team	1	1
7	1	paper	ten	1	1
1	1	peanut	thin	5	2
1	1	river	tin	6	2
5	1	tempo		1	1
2	1	tin		1	1
1	1	yap		1	1

All High(F0) seq 1					
F	*L*	**HIGH**	**LOW**	*F*	*L*
4	1	al	al	1	1
4	1	alan	bahby	4	1
1	1	alley	bang	1	1
1	1	annie	bap	1	1
1	1	bau	body	6	1
1	1	boing	boing	2	1
1	1	bow	bom	1	1
5	2	by	bon	1	1
1	1	din	boy in	4	1
2	2	el	boying	5	1
1	1	eye	boyit	1	1
1	1	eyes	burlin	1	1
1	1	fine	buy	1	1
2	2	funny	by	1	1
1	1	gin	chin	1	1
2	2	happy	early	2	1
1	1	help	earth	1	1
2	2	hi	ehn	3	1
1	1	hin	fan	2	1
1	1	honey	fun	3	1
9	1	hour	gin	3	1
1	1	ice	happy	1	1
6	1	i-lean	help	2	1
3	1	in	him	2	1
1	1	line	i-lean	1	1
1	1	loo	in	1	1
1	1	look	insense	1	1
2	1	lot	jen	1	1
1	1	lucky	keep up	1	1
1	1	matty	lot	1	1
6	4	mine	lum	1	1
1	1	moh	make	1	1
1	1	mon	me	1	1
4	4	money	mine	3	1
1	1	muddy	money	1	1
2	2	mummy	mum	1	1
1	1	nah	my-ee	1	1
1	1	niched	nah-nu	1	1
3	1	nigh	nime	1	1
2	1	nine	nine	1	1
1	1	noh	norm	1	1
1	1	nom	num	1	1
2	1	nun	nun	1	1
1	1	on	oww	1	1
1	1	one two 3 4	patty	1	1
1	1	out	remake	1	1
5	2	oww	shin	1	1
1	1	patty	sim	1	1
1	1	pen	sin	5	1
1	1	pin	thin	3	1
1	1	run	volen	1	1
1	1	running	volume	2	1
1	1	shin	win	1	1
2	1	sin			
1	1	why			

All High(F0) seq 2					
F	*L*	**HIGH**	**LOW**	*F*	*L*
3	1	agong	agong	1	1
3	1	airport	aho	1	1
1	1	amap	apple	4	2
13	4	apple	bearden	1	1
1	1	beem	bin	5	4
1	1	bim	boma	1	1
1	1	bin	boy in	1	1
4	1	boma	boying	4	2
2	1	bon	chin	1	1
2	1	bonga	dear	1	1
1	1	camble	deem	1	1
1	1	din	den	1	1
2	1	dinner	didn't	2	2
1	1	flappy	din	3	3
1	1	gamble	dinner	3	2
1	1	green	ear	1	1
1	1	handle	gym	1	1
7	3	happy	him	2	1
1	1	hear	honey	1	1
1	1	im	humpty	1	1
1	1	in	in	3	2
5	1	keep up	key-un	1	1
1	1	keyon	lug	1	1
1	1	lom	lull	2	1
1	1	mambo	mah	1	1
2	1	man bored	mom	1	1
1	1	mandel	mon	1	1
2	1	man-door	nah	1	1
1	1	mankey	nano	3	1
1	1	map	nee nah	1	1
2	1	mapo	nee on	1	1
1	1	mapple	noo nah	1	1
2	1	mato	peanut	1	1
1	1	member	pin	4	2
1	1	monday	purple	2	1
1	1	money	same	1	1
3	2	mummy	sim	1	1
3	1	napple	sin	3	2
1	1	ok	taco	1	1
4	1	paper	tent	1	1
1	1	peanut	thank you	1	1
1	1	pink	thin	9	3
1	1	remember	tin	5	2
2	1	seven			
4	1	simple			
3	1	tempo			
1	1	ten			
2	2	thin			
1	1	thinner			
2	1	tin			
1	1	volume			

Figure 4.5 Forms reported in Experiment 5 for conditions *All Low(F0)* and *All High(F0)* B, shown separately for each sequence (1 or 2) and for each voice. ‘F’ is a total number of responses for a given form and ‘L’ is a total number of listeners who reported that particular form.



Baseline

F	L	Sequence 1		F	L	F	L	Sequence 2
10	2	al	boying	16	3	7	2	agong
8	2	alan	boying	7	2	8	2	airport
2	1	bah	din	3	2	35	5	apple
8	1	bahby	din	11	3	7	4	bin
2	1	beeper	happy	9	4	7	2	boma
3	1	blame	happy	22	3	2	1	bon
3	1	bobby	him	2	1	2	1	bonga
6	1	body	him	2	1	3	2	bum
6	1	boy in	in	7	2	2	1	bumna
5	2	by	in	12	4	4	1	bun
5	1	early	mummy	2	2	3	1	didden
2	1	earth	mummy	3	2	4	2	didn't
3	1	ehn	sin	12	2	5	2	dinner
2	2	el	sin	5	2	2	1	done
2	1	fan	thin	3	1	2	1	donna
3	1	fun	thin	16	3	3	1	handle
2	2	funny				7	2	keep up
5	1	gin				3	1	lon
2	1	going				2	1	lull
2	1	help				2	1	mambo
2	2	hi				5	2	man bored
16	2	hour				2	1	man door
6	1	i-lean				3	2	mandle
2	1	jeep				2	1	man-door
2	1	lot				2	1	mapo
20	4	mine				2	1	mato
9	4	money				2	1	nando
3	1	naan				3	1	nano
2	2	nah				2	1	napo
3	1	nigh				8	2	napple
2	1	nine				11	2	paper
4	2	nun				6	2	pin
4	1	owl				2	1	purple
7	2	oww				2	1	see-on
2	1	pal				2	1	seven
4	1	ten				4	1	simple
2	1	volume				8	2	tempo
2	1	wine				15	2	tin
		38 173 [227]	**8** 132					**38** 191 [269]

Opposing

F	L	Sequence 1		F	L	F	L	Sequence 2
2	1	alley	body	15	2	3	1	agong
2	1	alsee	body	2	1	2	2	akoh
2	1	babby	happy	17	4	40	4	apple
2	2	backy	happy	17	4	7	6	bin
2	2	baggy	in	5	3	3	1	boma
10	2	bahby	in	14	5	3	1	camble
2	1	bang	pappy	11	2	10	2	didden
9	3	batty	pappy	3	2	2	1	didn't
2	1	beber	people	3	1	2	1	different people
3	1	beeper	people	2	2	8	3	din
3	1	blame	pin	6	2	8	3	din
5	2	bobby	pin	2	1	3	1	donna
11	6	bomb	ten	2	1	2	2	edin
13	3	bon	ten	11	2	2	1	ehm
3	3	bond				4	2	ehn
12	2	boying				4	2	ehtin
4	1	bubbely				2	1	gamble
2	1	buddy				3	1	gonga
2	1	buffy				10	2	handle
17	6	bum				2	2	handoor
2	1	bun				3	1	him
6	1	burrin				3	1	hockey
2	1	by				2	1	idin
2	1	coin				10	2	keep up
2	1	dee-pad				4	1	lud
6	2	deeper				5	1	man bored
2	2	dumb				2	1	manball
10	3	fatty				2	1	mapo
4	2	feefa				7	2	mato
2	1	flower				5	1	napo
3	1	fucky				8	2	napple
2	2	fun				2	1	nappy
2	1	glen				2	1	netball
2	1	gun				2	1	over
2	1	keeper				3	1	paper
2	2	mine				2	1	purple
2	1	mom				3	1	sen
2	1	mong				5	1	sentence
7	2	mum				7	2	seven
2	1	naan				2	1	simple
2	1	nah				3	2	temple
8	2	nom				7	1	tempo
2	2	num				6	2	thin
3	1	on				5	1	thing
2	2	party						
6	2	patty						
10	3	puppy						
2	1	salad						
3	1	sheep						
2	2	sid						
2	1	theta						
2	1	tin						
4	2	volume						
2	1	wait						
3	1	way						
		55 225 [284]	**7** 110					**44** 220 [271]

Congruent

F	L	Sequence 1		F	L	F	L	Sequence 2
4	1	alley	bin	11	3	2	1	ammo
3	1	allo	bin	6	3	19	2	apple
3	1	amul	blob	3	3	2	2	apples
2	1	anno	blob	3	2	8	2	beer
3	1	bau	din	9	2	3	1	bim
2	2	bite	din	10	2	2	1	bitten
2	1	blame	happy	4	1	3	1	bomber
4	1	boy-in	happy	35	4	2	1	demon
3	1	boying	him	2	1	4	1	dill
6	2	by	him	2	1	7	2	dinner
2	2	chin	in	11	3	2	1	doma
3	1	early	in	9	2	9	2	donna
2	1	earth	money	11	3	2	2	fear
2	2	ehn	money	2	2	4	1	fill
2	1	end	mop	3	2	2	1	glue
2	1	enough	mop	8	2	6	2	handle
5	1	fighting	not	3	2	2	2	here
2	2	funny	not	9	4	2	1	hidden
8	2	gin	pin	5	2	4	1	hill
2	1	glen	pin	7	2	13	1	ill
2	1	hell	sim	2	1	2	1	knock
6	2	help	sim	2	1	3	3	lob
4	2	hen	sin	11	4	2	2	lot
3	3	hi	sin	2	2	2	2	mammal
2	2	higher	tin	20	4	2	2	man
2	1	lock	tin	7	3	3	1	man bored
2	1	lost				2	1	man door
3	1	make				2	1	mandle
5	1	martin				2	1	mandoor
9	5	me				8	1	mankey
5	2	meh				2	1	mapo
2	1	min				3	1	mato
10	4	mine				3	1	monkey
4	1	mish				5	2	mum
2	1	miss				2	1	mummy
2	1	mit				2	1	napple
2	1	more				2	1	nock
2	2	moth				2	1	non
2	1	my				4	1	nut
3	1	nee				2	1	ok
3	1	no				6	2	peanut
5	2	oww				2	1	peel
4	1	pau				2	1	peema
4	1	power				2	1	tear
10	2	puppy				2	1	till
2	1	put in				2	1	uppy
5	2	remake				2	1	zeemu
2	1	salad				2	1	zim
3	1	send				2	1	zimu
2	2	sense						
2	1	ten						
4	2	thin						
3	1	volume						
		53 183 [278]	**13** 197					**49** 175 [277]

Figure 4.6 Overlap for the three condition groups between responses to the two sequence permutations used in Experiment 5. Values at the bottom of each diagram represent the number of unique forms heard (values in bold and large font) and the total number of VTs that occur (values in grey and small font). Values in brackets represent total number of VTs for a given sequence, unique plus shared. 'F' is a total number of responses for a given form and 'L' is a total number of listeners who reported that particular form.

Figure 4.6 shows the overlap for the three condition types (baseline, opposing, congruent) between the two sequence permutations used. The data have been collapsed across conditions (e.g. *All Low(F0)* and *All High(F0)* in the leftmost part of the diagram), and after removing isolated responses (a single form reported only once by only one listener) it was possible to see the extent to which responses to the two sequence permutations overlap. It is evident from these diagrams that – despite the fact that there are no differences between the two sequences in the likelihood of producing VTs or different forms in the baseline and congruent cases – there was a substantial difference in the specific verbal forms produced. This further implies that, although there is some overlap between responses to these sequences for all three grouping cues, the two permutations produced idiosyncratic instances of forms. Additionally, for all such comparisons (in the next experiment as well) there were more forms that were unique to one sequence or the other than there were forms in common to both. This is perhaps unsurprising, given the change in vowel order between the two sequences.

Nonetheless, the regions of overlap for all comparisons between sequences 1 and 2 always included those forms which were reported most often. Consider the ratio of the total number of responses (small font number) to unique responses (figure in bold). It is evident that although most of the forms are not common between the two sequences, the small number that are in common account for a much higher proportion of the total number of responses than would normally be expected by chance.

Comparison of shared and unique forms across condition types

The following Venn diagrams illustrate the relationship between responses in the baseline, opposing, and congruent conditions, shown separately for the two sequence permutations (Figures 4.7 and 4.8). The different coloured areas represent the unique forms for a given condition while the grey areas of overlap show forms which are either common to two conditions (e.g. opposing and congruent) – the peripheral areas – or forms which are common to all three conditions – the area in the middle. Note that the smaller numbers in grey show the total number of transformations reported rather than unique forms. For both sequences, it can be seen that there are more forms reported in the opposing and the congruent cases than there are in the baseline condition. This trend is reflected in the mean numbers for the three groups (9.88 Forms per 3 minutes for baseline, 10.40 for opposing, and 10.85 for congruent) although the main effect of condition (in the superordinate ANOVA) was not significant. This suggests that, irrespective of whether the additional cue was opposing or congruent, adding an F0 difference between pairs of vowels in the sequence generated more forms.

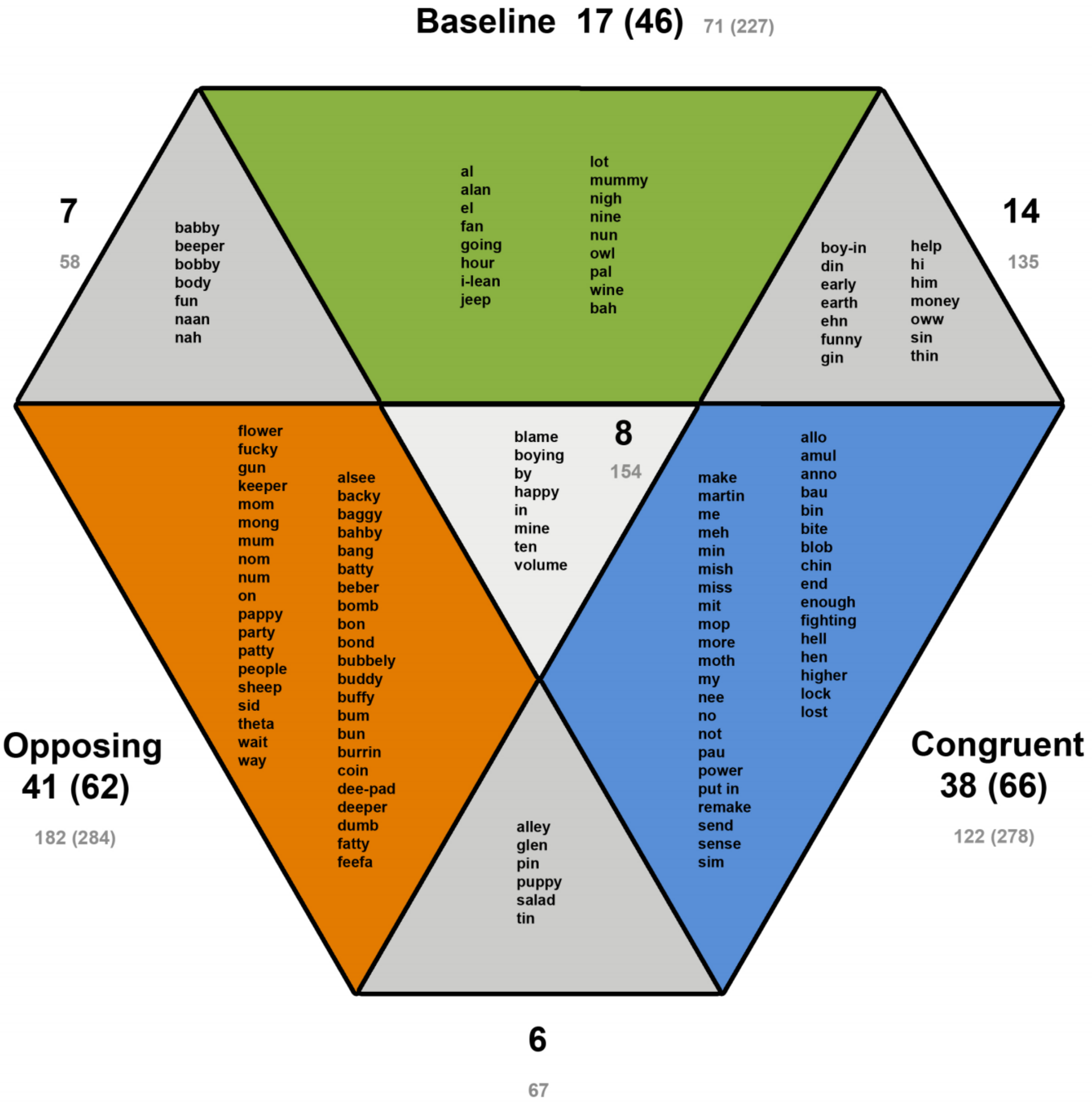


Figure 4.7 Relationship between the baseline, opposing and congruent grouping cues for Sequence 1 in Experiment 5. Values in bold represent number of unique forms heard and values in grey show the total number of VTs that occurred. Values in brackets represent total number of VTs for a given condition, unique plus shared.

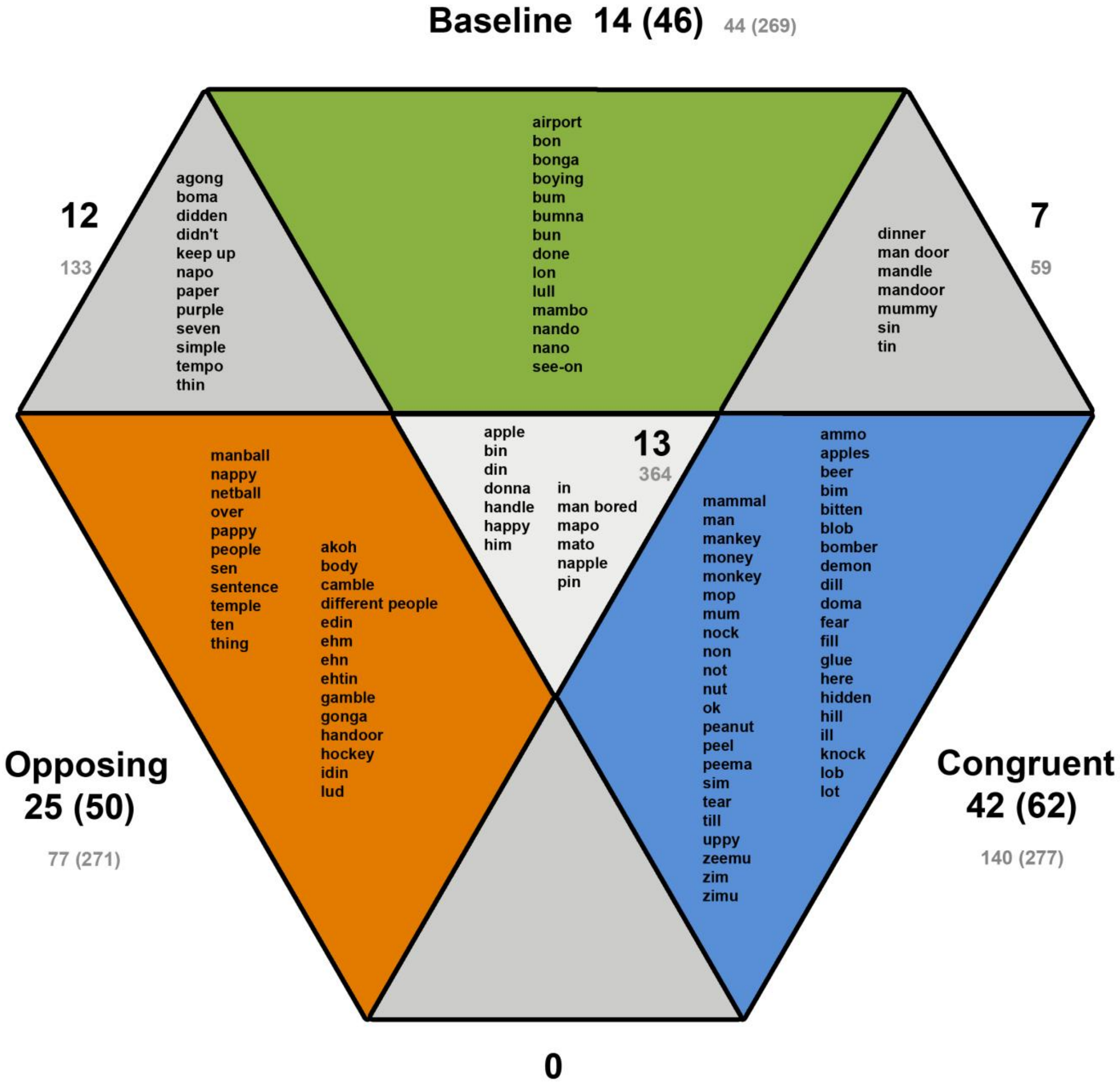


Figure 4.8 Relationship between the baseline, opposing and congruent grouping cues for Sequence 2 in Experiment 5. Values in bold represent number of unique forms heard and values in grey show the total number of VTs that occurred. Values in brackets represent total number of VTs for a given condition, unique plus shared.

In terms of the relationships between any two conditions, there seems to be less overlap between the opposing and congruent cases than there is between the baseline and opposing, or the baseline and congruent cases; this is most evident for sequence 2. Note that absence of entries in the unique area of overlap between any particular pair of conditions does not imply that there are no shared forms between those two conditions, as there may be cases which are shared by all three. These cases are represented in the middle of the Venn diagram - see Figure 4.8 – overlap between opposing and congruent. This in turn suggests that for sequence

2 there are no *unique* common forms that are shared between the opposing and congruent conditions. Therefore, the total overlap between any pair of conditions is the sum of all the unique forms from the corresponding peripheral grey area plus the area common to all three cues. For example, for sequence 2 the overlap for baseline-congruent will be 13+7, for baseline-opposing 13+12, and for opposing-congruent 13+0. Nonetheless, for both sequences, the total number of forms shared by the opposing and congruent conditions is the smallest.

4.3 Experiment 6 – PTE and ITD cues

ITD cues can be a very effective sequential grouping cue, and so introducing inconsistency between pairs of vowels using ITD cues might affect the type of forms that are reported on either side. This was explored in terms of the PTE, where the experimental sequence of vowels was presented to both ears but with each of the two pairs of vowels lateralised to opposite ears using the maximum natural ITD possible (see below).

As for the previous experiment, sequences of four vowels - [iː], [uː], [ae], and [ɒ] as in the words: ‘heat’, ‘hoot’, ‘hat’ and ‘heart’ - were chosen such that the default grouping (according to spectral similarity) would be the [iː] and [ae] vowels in one stream and the [uː] and [ɒ] in the other. ITD cues were used either to ‘support’ the timbral cue ([iː] and [ae] presented to one ear, e.g. right, and [uː] and [ɒ] presented to the other, e.g. left) or to favour an alternative grouping – [iː] & [uː] presented to one ear, e.g. right, and [ae] & [ɒ] presented to the other, e.g. left. For the alternative arrangement, if grouping occurs on the basis of ITD cues, listeners will report different transformations to the ones heard in a ‘supported’ sequence. Furthermore, as the two cues, timbre and ITD, will be in competition with each other, hence offering more possibilities for within and across ear perceptual re-grouping, it was hypothesised that listeners will report more VTs and forms in the alternative arrangement rather than when ITD cues supported the timbral cues.

4.3.1 Methods

Participants

Twelve listeners (3 males, 9 females) took part in the experiment. They were all native speakers of English and reported normal hearing. At the end of the study, they were either paid cash or received course credit. The listeners' mean age was 23.2 years old (s.d. = 5.32).

Stimuli and Conditions

The general procedure was the same as in Experiment 5. The difference between the two studies was in the way the stimuli were manipulated. Like before, two sequence permutations of the four vowels were used. If [iː] is 1, [uː] is 2, [ae] is 3 and [ɒ] is 4, **sequence 1** was: 1-2-3-4 and **sequence 2** was 2-4-1-3. No acoustical mixing or transitional stages (e.g., amplitude ramps) from one vowel to the next were used.

There was a total of six experimental conditions in the study. For the labelling of the conditions *TP* stands for *tongue position* while *Right* and *Left* refer to lateralization of the vowels. Hence the sequence *High(TP)Left/Low(TP)Right* will be interpreted as 'vowels with a high tongue position were presented to the left ear while vowels with a low tongue position were presented to the right ear'.

The first two conditions included all vowels lateralised on the left – *All Left*, or all vowels lateralised on the right – sequence *All Right*. The remaining four conditions were constructed using the same principle of congruent and opposing sequences as for Experiment 5. This time, however, rather than one pairing being synthesised on a different F0 frequency from the other, they were lateralised either to the left or right. All the experimental conditions are summarized in Table 4.6.

To produce the lateralised versions of the two sequences, the maximum natural ITD of 680 µs was used (this value is based on the size of the average adult male head). Using MITSYN (Henke, 1997) left and right lateralised versions for each condition were created by introducing appropriate delays (see below). All sequences were generated on an F0 frequency of 120 Hz. Each pulse was iterated 10 times (10 x 8.35 ms), hence the four-vowel sequence was 334-ms long. Again, note that there were congruent and opposing pairings of cues with respect to the baseline (timbre-based) pairings.

Table 4.6 The conditions and its vowels arrangements in Experiment 6.

Type	Condition	Vowel arrangement
Baseline (natural timbre only)	***All Left***	All 4 heard on the left
	All Right	All 4 heard on the right
Congruent	***Front(TP)Left/Back(TP)Right***	[iː] & [ae] to the left, [uː] & [ɒ] to the right
	Front(TP)Right/Back(TP)Left	[iː] & [ae] to the right, [uː] & [ɒ] to the left
Opposing	***High(TP)Left/Low(TP)Right***	[iː] & [uː] to the left, [ae] & [ɒ] to the right
	High(TP)Right/Low(TP)Left	[iː] & [uː] to the right, [ae] & [ɒ] to the left

For each condition, there were two 3-minute presentations (*All Left seq*1 and *All Left seq*2, *All Right seq*1 and *All Right seq*2, and so on), one for each sequence permutation. Within each session, the order of 3-minute sequences was randomised. Participants were assigned to either the ‘Odds’ or ‘Evens’ group (see Figure 4.9) and attended two sessions - either with conditions *All Left* and *All Right* in the first session and the remaining four conditions in the second, or vice versa. Listeners were required to report all transformations that they heard. In addition, for conditions *All Left* and *All Right*, they were asked to indicate (using key presses) on which voice (higher or lower) the change occurred. For the remaining conditions, participants instead indicated whether the change occurred on the left- or the right-hand side of space. Due to the nature of the experimental design, it was important to provide clear instructions to the participants. Although there was no F0 difference between vowels on any presentation in the current experiment, for conditions *All Left* and *All Right* listeners were told that on any trial they should be able to identify two voices, typically one which sounds lower and one which sounds higher.

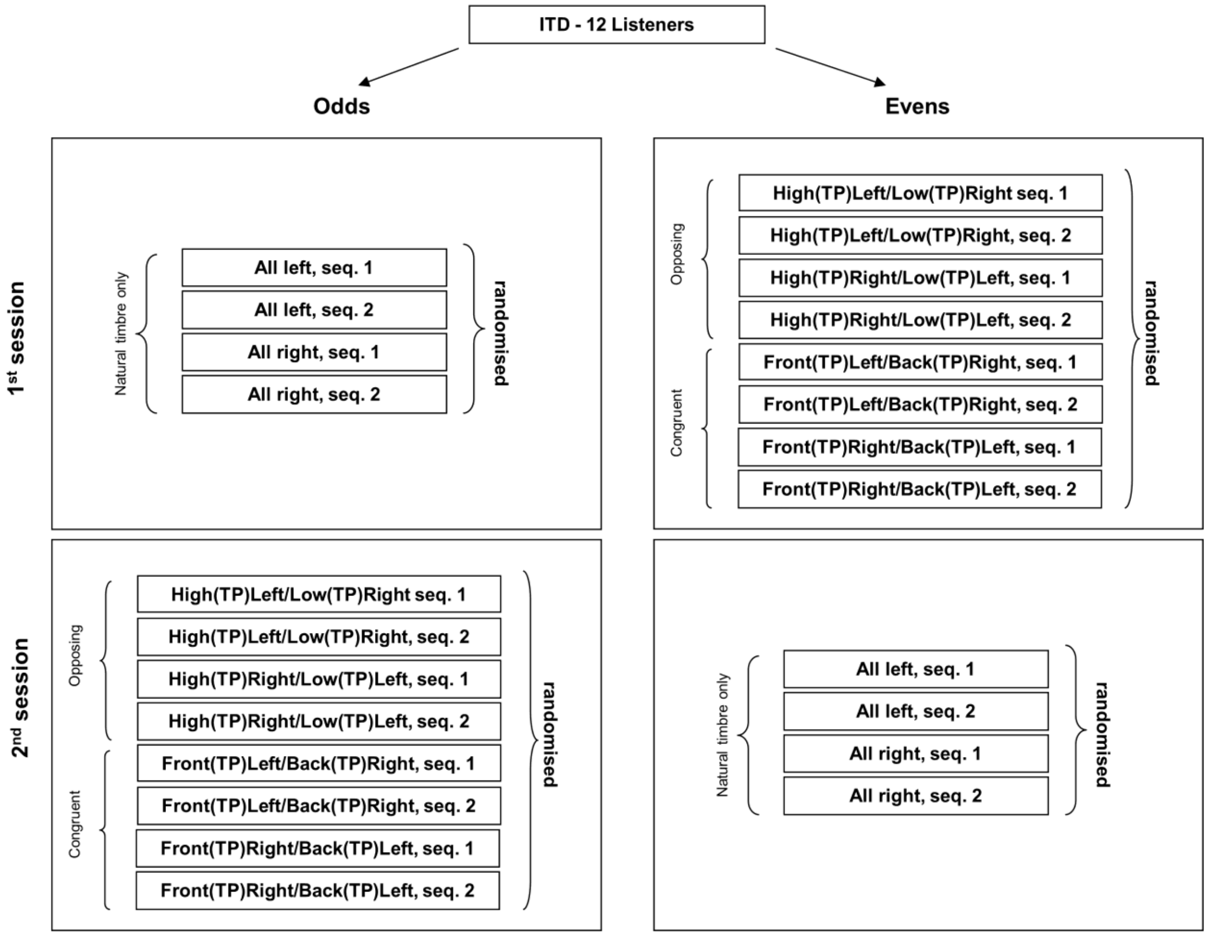

ITD - 12 Listeners
Odds
Evens
1st session
2nd session
Natural timbre only
All left, seq. 1
All left, seq. 2
All right, seq. 1
All right, seq. 2
randomised
Opposing
High(TP)Left/Low(TP)Right seq. 1
High(TP)Left/Low(TP)Right, seq. 2
High(TP)Right/Low(TP)Left, seq. 1
High(TP)Right/Low(TP)Left, seq. 2
Congruent
Front(TP)Left/Back(TP)Right, seq. 1
Front(TP)Left/Back(TP)Right, seq. 2
Front(TP)Right/Back(TP)Left, seq. 1
Front(TP)Right/Back(TP)Left, seq. 2
randomised

Figure 4.9 Experimental design for Experiment 6.

Further information on the lateralisation of the stimulus sequences.

The following describes in detail the procedure used to apply lateralisation cues to the sequences. In the example below (Figure 4.10), initially the right ear is delayed by 680 µs; hence, a listener experiences the first vowel as coming from the left (see first red area). The second vowel, however, is right lateralized as the first glottal pulse to the left ear is delayed by 680 µs. In order to compensate for this switch between the leading ears, a silence of 680 µs has been added after the last glottal pulse of the first vowel (see second red area). As a result, there is a silent gap of 1360 µs (2 x 680 µs) between the last pulse of the first vowel and the first pulse of the second vowel in the left ear, whereas in the right ear, those glottal pulses meet at the zero crossing.

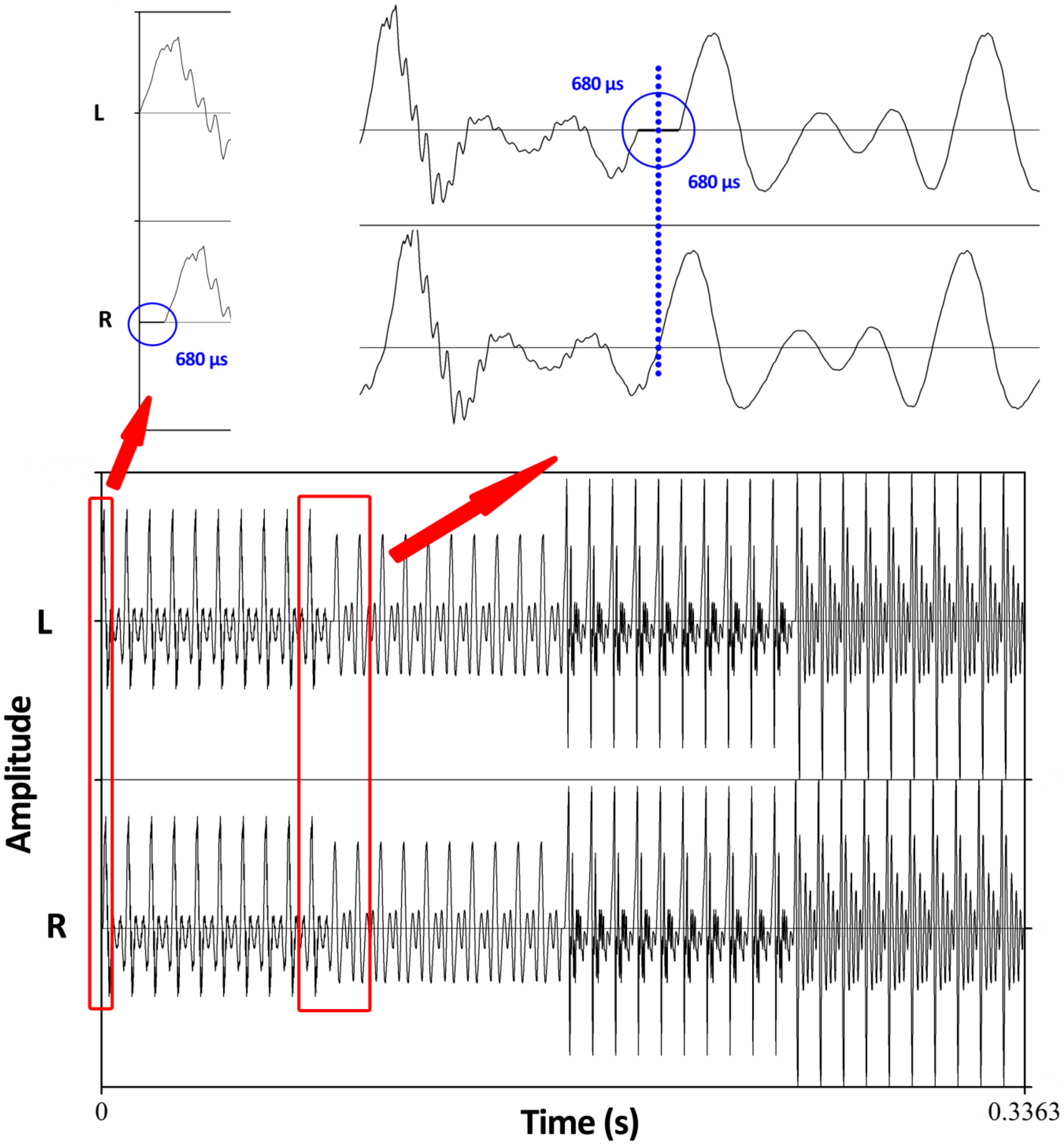


Figure 4.10 Lateralisation technique used in Experiment 6. The lower panel shows time waveform of one cycle of a four-vowel sequence and the upper details how the ITD cues were implemented.

4.3.2 Results and Discussion

Verbal Transformations

Just as in Experiment 5, for the baseline conditions (*All Left* and *All Right*) there was no effect of either sequence permutation, different condition or the voice reported by participants. Table 4.7 shows the mean number of VTs for each condition and it suggests that the sequence permutation has an effect for conditions *High(TP)Left/Low(TP)Right* and *High(TP)Right/ Low(TP)Left* (opposing case). This was confirmed by the corresponding three-way ANOVA (permutation x condition x voice), for which the main effect of permutation was significant [F(1,11)=5.92, p=.03, η^2=.35]. Listeners reported more forms for sequence 1 (5.85 VTs per 3 min, s.e.= 0.67) than for sequence 2 (4.46 VTs per 3 min, s.e.=0.61). However, there was also a significant interaction between permutation, condition and voice. From Table 4.7, it seems that the three-way interaction is driven by the following patterns: (a) for condition *High(TP)Right/Low(TP)Left* there was a significant difference between the two sequences (7.50 VTs per 3 minutes for sequence 1 vs. 2.75 for sequence 2) in the right ear and a similarly large difference between the sequences for condition *High(TP)Left/Low(TP)Right* in the left ear (7.17 for sequence 1 and 3.50 for sequence 2). (b) This pattern was reversed for each condition in the opposite ear. Namely, there were fewer responses to sequence 1 than sequence 2 (4.75 vs. 6.17 respectively) for condition *High(TP)Right/Low(TP)Left* in the left ear, and for condition *High(TP)Left/Low(TP)Right* there were fewer VTs for sequence 1 compared to sequence 2 (4.00 vs. 5.42) in the right ear.

As conditions *High(TP)Left/Low(TP)Right* and *High(TP)Right/Low(TP)Left* are mirror images of each other in terms of lateralization, the results of the interaction suggest that they are not due to an ear effect – listeners are not showing a general preference for either side. It is worth noting that there was a corresponding trend in Experiment 5, although in that case the three-way interaction did not reach significance [F(1,11)=2.53, p>.1, η^2=.19]. Although the general character of the trend was the same across the two experiments, in the current study this was due to lateralization of the stimuli rather than to the F0 manipulation used in the previous one. This relationship between separation of the vowels by F0 difference in Experiment 5 is demonstrated more clearly by the lateralization of the vowels in the current study.

Table 4.7 Average number of VTs reported in Experiment 6 across all conditions. For the baseline conditions, listeners classified each VT as spoken either by a bright voice or a dull voice. For the opposing and congruent conditions, each VT was classified as either coming from the left or the right side of space.

	Condition	**Average no. of VTs reported in 3min** *(±SE)*		
		Cumulated	**Bright Voice**	**Dull Voice**
Baseline	***All Left*** **seq1**	10.92 *(1.69)*	6.33 *(1.86)*	4.58 *(0.72)*
	All Left **seq2**	10.83 *(1.52)*	5.58 *(1.46)*	5.25 *(1.13)*
	***All Right seq*1**	10.00 *(1.31)*	5.50 *(1.40)*	4.50 *(0.76)*
	***All Right seq*2**	8.75 *(1.14)*	4.92 *(0.85)*	3.83 *(0.61)*
			Right	**Left**
Congruent	***Front(TP)Left/Back(TP)Right*** **seq1**	9.92 *(1.94)*	6.00 *(0.73)*	3.92 *(1.60)*
	Front(TP)Left/Back(TP)Right **seq2**	9.50 *(1.10)*	5.92 *(1.15)*	3.58 *(0.87)*
	***Front(TP)Left/Back(TP)Right seq*1**	8.92 *(1.53)*	4.00 *(1.05)*	4.92 *(1.03)*
	Front(TP)Left/Back(TP)Right **seq2**	9.75 *(1.16)*	4.08 *(0.71)*	5.67 *(1.30)*
Opposing	***High(TP)Left/Low(TP)Right*** **seq1**	11.17 *(1.20)*	4.00 *(0.84)*	7.17 *(0.83)*
	High(TP)Left/Low(TP)Right **seq2**	8.92 *(1.33)*	5.42 *(1.01)*	3.50 *(0.70)*
	High(TP)Right/Low(TP)Left **seq1**	12.25 *(1.53)*	7.50 *(1.22)*	4.75 *(0.96)*
	High(TP)Right/Low(TP)Left **seq2**	8.92 *(1.30)*	2.75 *(0.57)*	6.17 *(1.43)*

For conditions *Front(TP)Left/Back(TP)Right* and *Front(TP)Left/Back(TP)Right* (congruent), although there was no main effect of either of the three factors, there was a significant interaction between condition and voice [F(1,11)=6.81, p=.02, η^2=.38]. For condition *Front(TP)Left/Back(TP)Right*, there were more VTs reported on the right side than on the left (5.96 *(0.87)* vs. 3.75 *(0.81)* respectively) while the opposite was the case for condition *Front(TP)Left/Back(TP)Right*, where more VTs were reported on the left side than on the right (5.29 *(0.77)* vs. 4.04 *(0.52)* respectively). This further emphasised the fact that this effect was most likely due to the different regroupings within conditions rather than being attributable to a particular ear effect.

The additional (superordinate) two-way ANOVA, including the factors permutation (sequence 1, sequence 2) and condition type (baseline, opposing, and congruent), was not significant in any of its terms.

Forms

The results of the three separate analyses for VTs for pairs of conditions *All Left* and *All Right*, *High(TP)Left/Low(TP)Right* and *High(TP)Right/Low(TP)Left*, and *Front(TP)Left/ Back(TP)Right* and *Front(TP)Left/Back(TP)Right* were very similar for the number of Forms. For opposing conditions *High(TP)Left/Low(TP)Right* and *High(TP)Right/Low(TP)Left*, the main effect of permutation did not quite reach significance [F(1,11)=4.10, p=.07, η^2=.27], but there was a trend in the same direction, with fewer forms being reported for sequence 2 (6.88 Forms per 3 min) than for sequence 1 (7.32 Forms per 3 min).

Two of the interaction terms were significant – (i) the three-way interaction between sequence permutation, condition, and voice for opposing cues *High(TP)Left/Low(TP)Right* & *High(TP)Right/Low(TP)Left* [F(1,11)=11.22, p=.01, η^2=.51]; (ii) the two-way, condition x voice interaction for the congruent cues E & F [F(1,11)=12.44, p=.01, η^2=.53].

Table 4.8 Average number of forms reported in Experiment 6 across all conditions. For the baseline conditions, listeners classified each VT as spoken either by a bright voice or a dull voice. For the opposing and congruent conditions, each VT was classified as either coming from the left or the right side of space.

Condition		**Average no. of Forms reported in 3min** *(±SE)*		
		Cumulated	**Bright Voice**	**Dull Voice**
Baseline	***All Left seq*1**	8.08 *(0.82)*	4.08 *(0.66)*	4.00 *(0.55)*
	***All Left seq*2**	7.75 *(1.05)*	4.25 *(0.89)*	3.50 *(0.50)*
	***All Right seq*1**	8.25 *(1.17)*	4.42 *(0.97)*	3.83 *(0.68)*
	***All Right seq*2**	6.67 *(0.82)*	3.42 *(0.50)*	3.25 *(0.49)*
Congruent			**Right**	**Left**
	***Front(TP)Left/Back(TP)Right* seq1**	6.67 *(0.79)*	4.33 *(0.61)*	2.33 *(0.47)*
	***Front(TP)Left/Back(TP)Right* seq2**	7.17 *(1.01)*	4.33 *(0.86)*	2.83 *(0.52)*
	***Front(TP)Left/Back(TP)Right seq*1**	6.58 *(1.06)*	2.75 *(0.70)*	3.83 *(0.67)*
	***Front(TP)Left/Back(TP)Right seq*2**	6.92 *(0.61)*	3.00 *(0.43)*	3.92 *(0.53)*
Opposing	***High(TP)Left/Low(TP)Right seq*1**	8.08 *(0.78)*	3.17 *(0.61)*	4.92 *(0.47)*
	***High(TP)Left/Low(TP)Right* seq2**	6.67 *(0.86)*	4.00 *(0.64)*	2.67 *(0.51)*
	***High(TP)Right/Low(TP)Left* seq1**	7.92 *(0.83)*	4.75 *(0.58)*	3.17 *(0.51)*
	***High(TP)Right/Low(TP)Left* seq2**	6.75 *(0.89)*	2.33 *(0.36)*	4.42 *(0.97)*

It is noteworthy that there is a much lower ratio of total responses (VTs) to forms i.e. a greater number of forms for the number of VTs in Experiments 5 and 6 compared to the previous experiments in this thesis. Comparable data is not available from Chalikia and Warren (1991) as they focused almost entirely on the first response. For Experiments 1 to 4 this ratio was almost twice as big (2.28 to 3.71) as for Experiments 5 and 6 (1.40 and 1.43 respectively). Table 4.9 shows average numbers of VTs and forms in 3 minutes for each experiment along with the VTs to forms ratio.

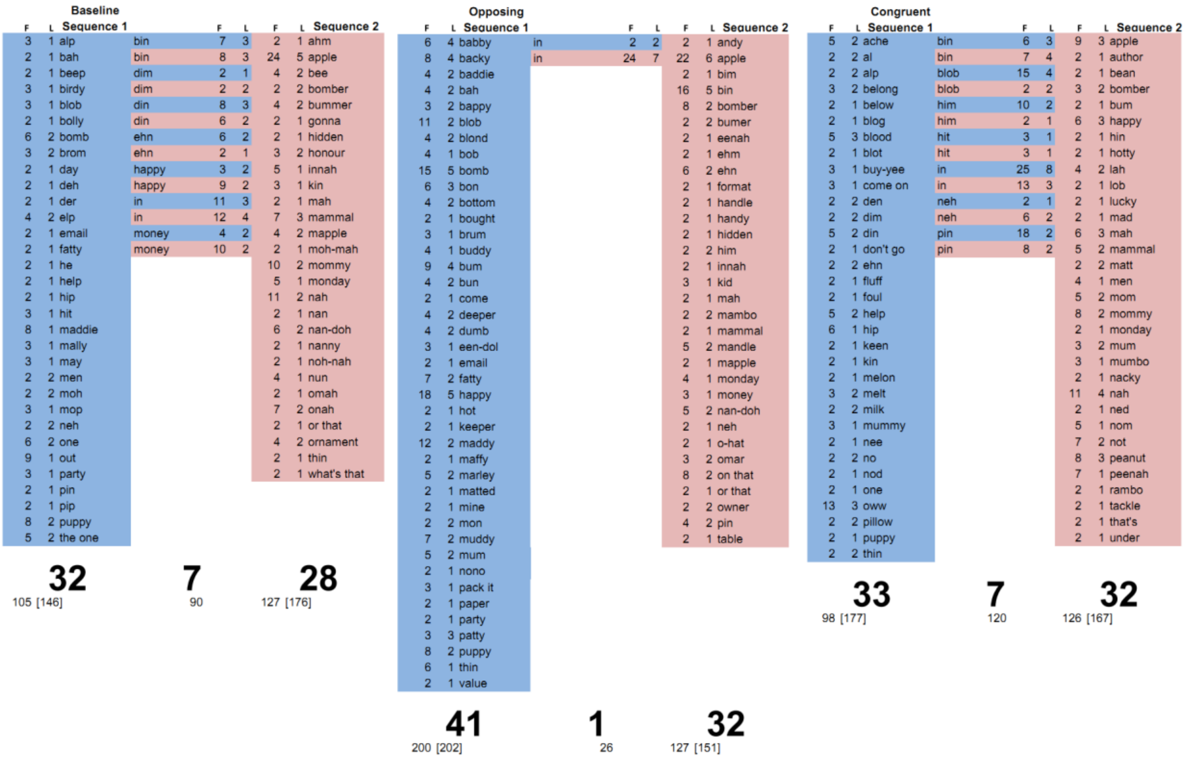
Table 4.9 Average number of VTs and Forms in all 6 experiments.

Experiment no.	1	2	3	4	5	6
VTs (in 3 min)	13.17	10.71	15.76	19.94	15.58	9.98
Forms (in 3 min)	3.52	4.23	6.73	7.61	10.38	7.30
Forms to VTs ratio	3.74	2.53	2.34	2.62	1.50	1.37

4.3.3 Additional analyses

Comparison of responses to sequences 1 and 2

As in Experiment 5, the diagrams below (see Figure 4.11) show the extent to which responses to sequence 1 and 2 were similar or different. It is noticeable that, in the opposing case (the middle diagram), there is only one form that is shared between the two sequences. It can be viewed as a magnified effect of the corresponding relationship from the previous experiment (see Fig 4.6 on p. 109). The other two conditions are very similar in terms of the number of unique and shared forms across the two sequences. In general, this represents a very similar pattern to that seen in the previous experiment.



Baseline

F	L	Sequence 1		F	L	F	L	Sequence 2
3	1	alp	bin	7	3	2	1	ahm
2	1	bah	bin	8	3	24	5	apple
2	1	beep	dim	2	1	4	2	bee
3	1	birdy	dim	2	2	2	2	bomber
3	1	blob	din	8	3	4	2	bummer
2	1	bolly	din	6	2	2	1	gonna
6	2	bomb	ehn	6	2	2	1	hidden
3	2	brom	ehn	2	1	3	2	honour
2	1	day	happy	3	2	5	1	innah
2	1	deh	happy	9	2	3	1	kin
2	1	der	in	11	3	2	1	mah
4	2	elp	in	12	4	7	3	mammal
2	1	email	money	4	2	4	2	mapple
2	1	fatty	money	10	2	2	1	moh-mah
2	1	he				10	2	mommy
2	1	help				5	1	monday
2	1	hip				11	2	nah
3	1	hit				2	1	nan
8	1	maddie				6	2	nan-doh
3	1	mally				2	1	nanny
3	1	may				2	1	noh-nah
2	2	men				4	1	nun
2	2	moh				2	1	omah
3	1	mop				7	2	onah
2	2	neh				2	1	or that
6	2	one				4	2	ornament
9	1	out				2	1	thin
3	1	party				2	1	what's that
2	1	pin						
2	1	pip						
8	2	puppy						
5	2	the one						
		32	**7**					**28**
		105 [146]	90					127 [176]

Opposing

F	L	Sequence 1		F	L	F	L	Sequence 2
6	4	babby	in	2	2	2	1	andy
8	4	backy	in	24	7	22	6	apple
4	2	baddie				2	1	bim
4	2	bah				16	5	bin
3	2	bappy				8	2	bomber
11	2	blob				2	2	bumer
4	2	blond				2	1	eenah
4	1	bob				2	1	ehm
15	5	bomb				6	2	ehn
6	3	bon				2	1	format
4	2	bottom				2	1	handle
2	1	bought				2	1	handy
3	1	brum				2	1	hidden
4	1	buddy				2	2	him
9	4	bum				2	1	innah
4	2	bun				3	1	kid
2	1	come				2	1	mah
4	2	deeper				2	2	mambo
4	2	dumb				2	1	mammal
3	1	een-dol				5	2	mandle
2	1	email				2	1	mapple
7	2	fatty				4	1	monday
18	5	happy				3	1	money
2	1	hot				5	2	nan-doh
2	1	keeper				2	1	neh
12	2	maddy				2	1	o-hat
2	1	maffy				3	2	omar
5	2	marley				8	2	on that
2	1	matted				2	1	or that
2	1	mine				2	2	owner
2	2	mon				4	2	pin
7	2	muddy				2	1	table
5	2	mum						
2	1	nono						
3	1	pack it						
2	1	paper						
2	1	party						
3	3	patty						
8	2	puppy						
6	1	thin						
2	1	value						
		41	**1**					**32**
		200 [202]	26					127 [151]

Congruent

F	L	Sequence 1		F	L	F	L	Sequence 2
5	2	ache	bin	6	3	9	3	apple
2	2	al	bin	7	4	2	1	author
2	2	alp	blob	15	4	2	1	bean
3	2	belong	blob	2	2	3	2	bomber
2	1	below	him	10	2	2	1	bum
2	1	blog	him	2	1	6	3	happy
5	3	blood	hit	3	1	2	1	hin
2	1	blot	hit	3	1	2	1	hotty
3	1	buy-yee	in	25	8	4	2	lah
3	1	come on	in	13	3	2	1	lob
2	2	den	neh	2	1	2	1	lucky
2	2	dim	neh	6	2	2	1	mad
5	2	din	pin	18	2	6	3	mah
2	1	don't go	pin	8	2	5	2	mammal
2	2	ehn				2	2	matt
2	1	fluff				4	1	men
2	1	foul				5	2	mom
5	2	help				8	2	mommy
6	1	hip				2	1	monday
2	1	keen				3	2	mum
2	1	kin				3	1	mumbo
2	1	melon				2	1	nacky
3	2	melt				11	4	nah
2	2	milk				2	1	ned
3	1	mummy				5	1	nom
2	1	nee				7	2	not
2	2	no				8	3	peanut
2	1	nod				7	1	peenah
2	1	one				2	1	rambo
13	3	oww				2	1	tackle
2	2	pillow				2	1	that's
2	1	puppy				2	1	under
2	2	thin						
		33	**7**					**32**
		98 [177]	120					126 [167]

Figure 4.11 Overlap for the three condition groups between two sequence permutations in Experiment 6. Values at the bottom of each diagram represent the number of unique forms heard (values in bold) and the total number of VTs that occur (values in grey). Values in brackets represent total number of VTs for a given sequence, unique plus shared. 'F' is a total number of responses for a given form and 'L' is a total number of listeners who reported that particular form.

Comparison of shared and unique forms across condition types

For the comparison of overlap in responses across the three condition types (shown separately for sequences 1 and 2 in Figures 4.12 and 4.13, respectively), there was a similar trend in terms of shared forms between conditions. There were fewer forms shared between the opposing and congruent cues (3+1 for sequence 1, and 7+3 for sequence 2) than between the other two pairings (with the exception of sequence 2, where for the baseline-congruent pair this number was equal). The introduction of ITD cues seems to restrict the number of forms that the opposing case shares with other conditions. By presenting the sequences to different ears, this effect is enhanced compared to Experiment 5 where conditions were differentiated by F0 frequency. Even though the ITD lateralisation was not a dichotic presentation, and so both sequences were present in two ears, listeners found it easier to distinguish between right- and left-lateralised voices than between higher and lower voices in the previous experiment.

This was also confirmed anecdotally by the author, who noted listeners' comments suggesting that the procedure for Experiment 5 was more challenging.

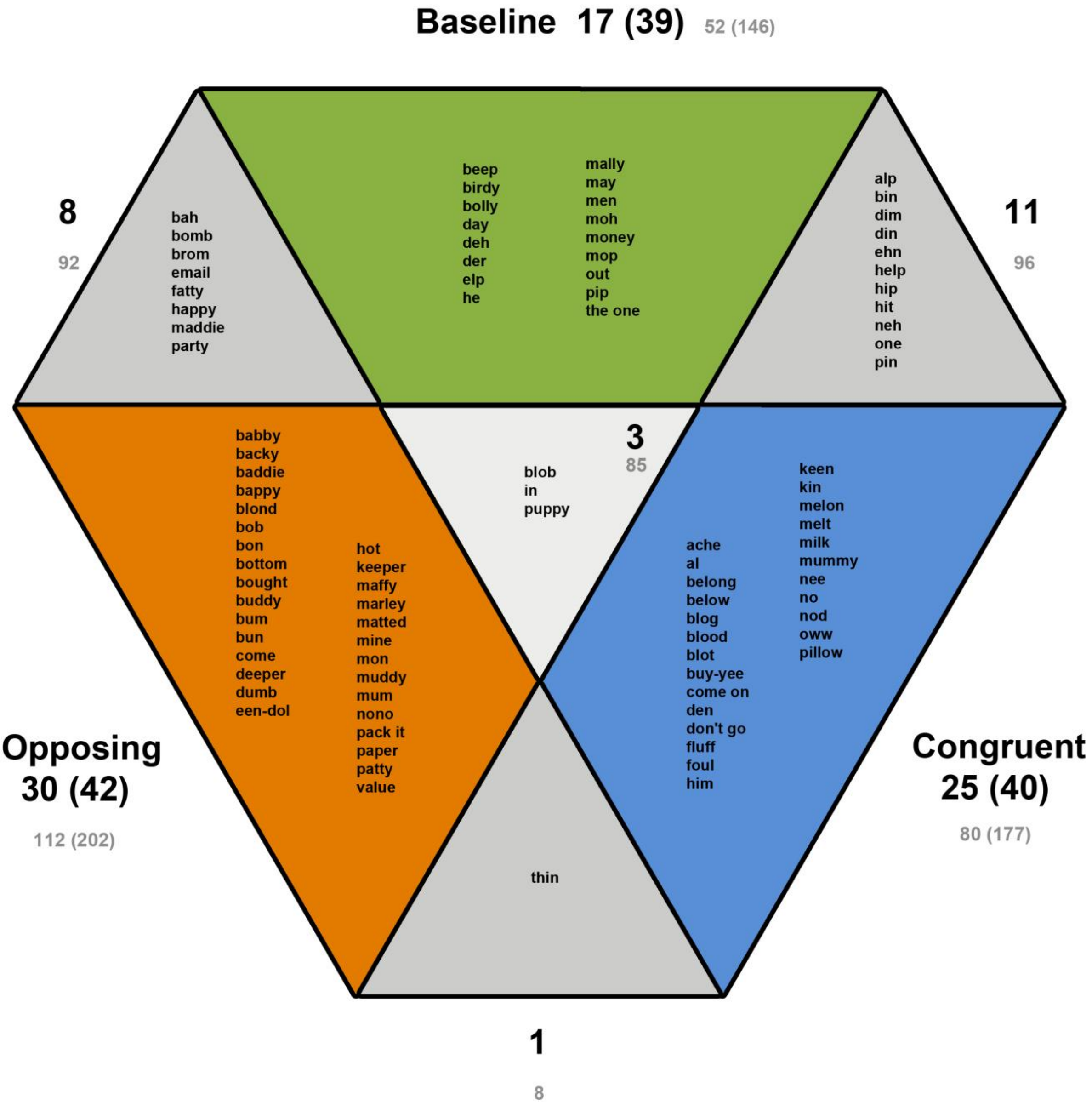


Figure 4.12 Relationship between the baseline, opposing and congruent grouping cues for Sequence 1 in Experiment 6. Values in bold represent number of unique forms heard and values in grey show the total number of VTs that occurred. Values in brackets represent total number of VTs for a given condition, unique plus shared.

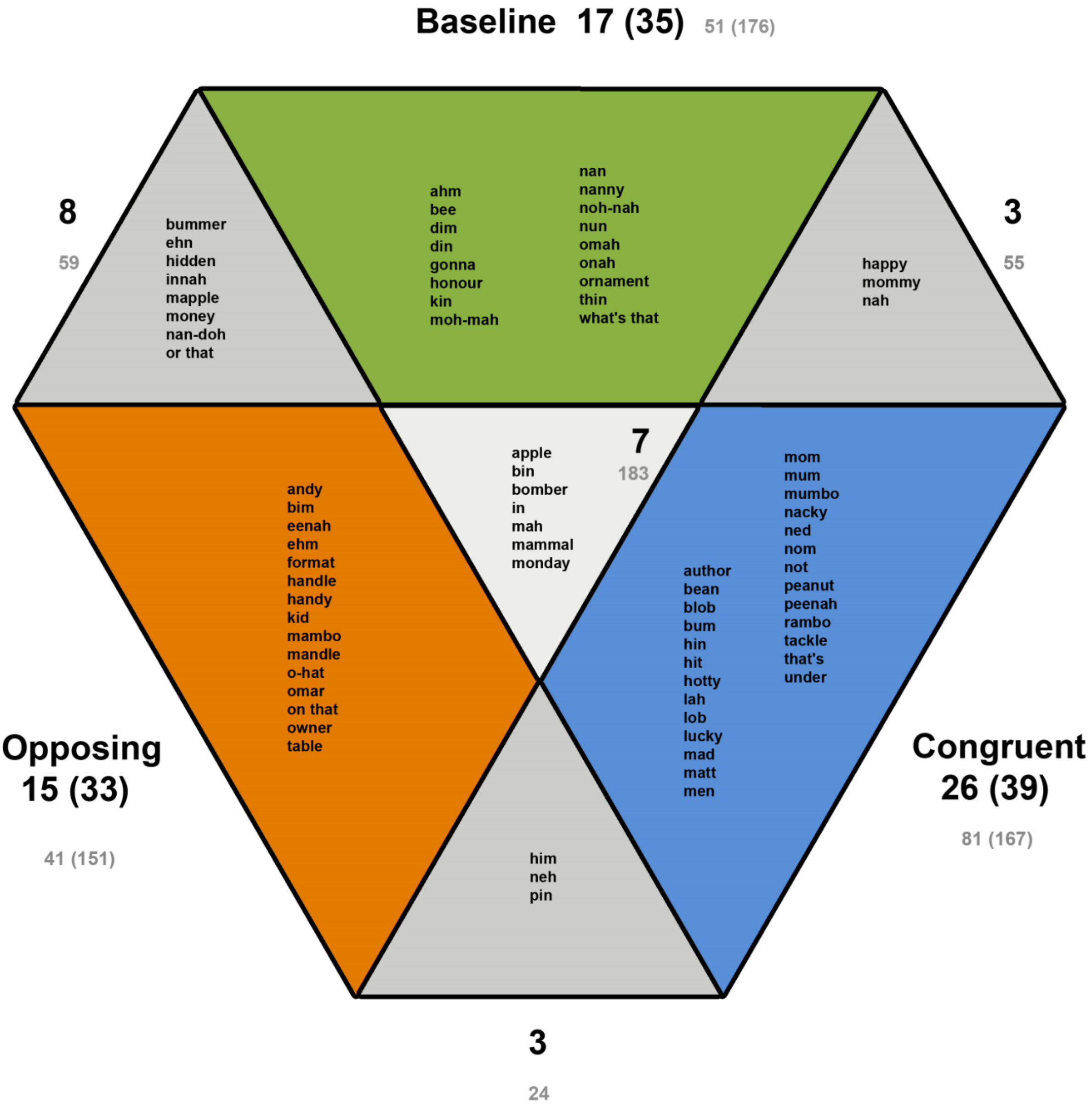


Figure 4.13 Relationship between the baseline, opposing and congruent grouping cues for Sequence 2 in Experiment 6. Values in bold represent number of unique forms heard and values in grey show the total number of VTs that occurred. Values in brackets represent total number of VTs for a given condition, unique plus shared.

Compared to Experiment 5, two main differences emerged from the descriptive inspection of the distribution and the type of responses from the current study. Firstly, supporting the average numbers from the statistical analyses, there are fewer unique forms reported by single participants for each condition. In other words, there are fewer forms reported by a single listener in the current study compared to Experiment 5. This could relate to the subjective experience of a different group of listeners which has been demonstrated in previous studies on VTE, and in our laboratory where the variation between the average number of responses for separate experiments can be considerable. Secondly, the differences in Forms reported between right- and left-lateralised voices are more distinct as they were between the higher and lower voice in Experiment 5. There tends to be an association between the forms for the particular voice and the vowels that are included in those forms, e.g. front or back. That pattern is much more obvious when the two voices were heard as left and right lateralised than they were when they were distinguished simply by their pitch or timbre. Responses were very rarely duplicated between the right- and left-lateralised voices, suggesting lesser opportunity for regrouping to take place both within and between the sequences.

It is worth noting that the very first response made by listeners here, and in Experiment 5, cannot be classified as a verbal transformation. Compared to the VTE, there is no *veridical* percept of a word (however controversial this concept is in itself, see earlier discussion in Chapter 1). As participants hear repeated sequences of vowels, their first response involving one or more consonants is an illusory percept (for a full list of first responses for Experiment 5 and 6 refer to Appendix 7.1 and 7.2 respectively). Only then does a participant experience a syllable or word which can in turn transform into other syllables, words or phrases. Therefore, the current experiment involved participants in two phases of responding. Firstly, the illusory consonant(s) arises from the sequencing of the vowels and the possible types of perceptual regroupings. Secondly, the modulation of the pattern of subsequent responses by factors such as F0 or ITD cues occurs. Even though most of the responses are idiosyncratic, some patterns in listeners' responses can be observed. In general, these initial illusory percepts agree with the type of VTs and Forms that will come out of a given condition. In other words, they give a good impression of the type of regroupings that will subsequently take place, e.g. for sequence *High(TP)Left/Low(TP)Right seq*1 reports on the right ear only, the majority of responses will be based around the vowel [ɒ]. Likewise for *High(TP)Left/Low(TP)Right seq*1 reports on the left ear only, many transforms will include the vowel [iː]. It is also clear from the distribution of the first responses that there was a propensity for nasal and plosive consonants being reported rather than fricative sounds. The few occasions when consonant 'f' was reported seems to cluster in the opposing conditions. In general, first responses confirm the distribution of consonants reported in the study usually indicate the form that will be reported most often in terms of subsequent transformations.

In summary, for both experiments, there were no significant differences in the number of forms reported by listeners. However, the descriptive analysis in which the conditions were rearranged by type into baseline, opposing, and congruent groups (after removing the single-report responses), showed that there are substantial changes in the particular forms heard by participants across conditions. Nonetheless, the small number of forms which fall into the region of overlap between the different conditions represents a relatively high proportion of the number of transformations reported. In other words, whilst there is relatively limited overlap compared to the forms that are uniquely heard for any sequence, there is greater overlap in terms of the total number of responses to particular forms (as they are reported more often). Additionally, in general there is evidence of a tendency for a smaller number of forms to be shared between opposing and congruent conditions than between other pairs.

Chapter 5

General Discussion

Experiments summary

The six experiments presented in this thesis set out to investigate the influence of various grouping cues on the Verbal Transformation Effect (VTE) and on the related phenomenon known as the Phonemic Transformation Effect (PTE). In both phenomena, upon listening to a repeated sequence of the same stimuli – either a word (VTE) or a series of concatenated vowels (PTE) – participants report hearing changes to the initial percept. Although it has been widely accepted that grouping cues, both general and speech specific, contribute to the perceptual organisation of speech within the framework of Auditory Scene Analysis (ASA), their relative contribution to that process is still debated (see, e.g., Remez et al., 1994; Darwin, 2008; Roberts, Summers & Bailey, 2010). While there is a considerable ASA research utilising simple sounds, relatively little has been done with more complex and dynamic signals, such as speech. Revisiting the VTE and PTE within the framework of ASA allowed this issue to be addressed. Both phenomena can be described as auditory illusions which can be used to investigate the normally inaccessible mechanisms underlying speech perception.

Two characteristics of the VTE/PTE approach are: (a) that participants are exposed to the stimuli for prolonged periods of time and (b) that they are essentially open ended tasks where in principle there is a limitless number of items into which the initial percept of a word (or sequence of vowels) can transform. The transformations are quite volatile and the frequency of change can be quite high within a short period of time; nonetheless, it was still possible to observe significant differences between various experimental conditions in the experiments reported in this thesis. As it was the intention to use stimuli derived from recordings of natural speech signals in the present series of experiments, the nature of natural speech is as such that it was necessary to use repeated exposure to build up sufficiently the tendency for stream segregation to occur. This requirement is also evident from less open ended, yet similar tasks such as the study by Cole and Scott (1973), where the presented material (which was restricted to repeating sequences of CV syllables) still required relatively long exposure for the segregation to occur. Although a closed-ended approach might be more controlled than the free report used in the current experiments, it inevitably restricts the number of new forms that listeners can report, hence potentially providing less information on the underlying processes of speech perception.

Pitt and Shoaf (2002) showed that extended repetition can reduce the perceptual coherence of the rapidly changing and diverse elements of speech. It was therefore argued that the VTE can be utilised to highlight the processes by which diverse elements of speech are grouped during speech perception. The influence of fundamental frequency (F0) and lateralization cues on the type and pattern of VTs was investigated in Experiment 1, using a modification of the paradigm introduced by Warren & Ackroff (1976). Using six words resynthesized on F0s of 100 Hz (low pitch) and 178 Hz (high pitch), two repeating sequences were presented concurrently, one on each pitch, but offset by half a cycle. It was found that even in the absence of differences in lateralization, listeners reported VTs on both sequences and these were mainly independent of one another. Additionally, the responses were significantly less independent when there was no separation of the two sequences by ear (i.e., where both sequences were presented to both ears). As the lateralization difference increased, the number of forms was reduced, with the fewest reported in the dichotic condition. The total number of VTs showed a similar trend. On average, the first VT occurred significantly later for the dichotic case and this tendency for later first VTs and fewer forms in the dichotic condition probably reflects a greater degree of perceptual re-grouping when each ear was stimulated by both sequences. Differences observed between particular words imply that the number of ways in which the elements of a given stimulus can recombine also depends on the acoustic variation of its phones. Finally, the similarity in the results for the no-ITD and ±680-μs ITD conditions suggested that a large difference in the apparent lateralization of the two sequences per se has little impact on perceptual re-grouping. Overall, the results are consistent with the hypothesis that verbal transformations are facilitated by the increased possibility of across-sequence re-groupings offered by conditions, allowing within-ear interactions between the two sequences.

In Experiment 1, there was a general preference for participants to provide more VTs on the high pitch. Responses to the high-pitched sequence were more numerous, displayed more forms, and occurred earlier than responses to the low-pitched sequence. These effects of sequence pitch (high vs. low) on verbal transformations, which were evident throughout the analyses, were explored further in Experiment 2. Whilst Warren and Ackroff (1976) used physical separation of the two sequences (i.e., dichotic presentation), in Experiment 2 diotic presentation was used and the only cue for the segregation of the two sequences was the difference in F0. Overall, the results suggested a tendency for responses (VTs and forms) to increase, and for the time to the first response to fall, when the second sequence was present. These changes are offset, in part or in whole, when listeners are presented with both sequences at once but are required to monitor only one of them. The fact that the number of VTs and forms declined in conditions where listeners monitored both sequences compared to when they had to monitor one or the other suggests a constraint arising from listeners trying to monitor both streams at the same time. In essence, the difference in response patterns between conditions where single (Low & High) or two concurrent sequences (High/Low & High/Low) were played seems to be primarily driven by the stimulus difference, whereas the difference between conditions with both sequences present (High/Low & High/Low and High/**LOW** & **HIGH**/Low) is primarily driven by the limitations of the response strategy. This further indicates that the particular combination of stimuli and task used in High/Low and High/Low is the most effective in terms of eliciting a greater number of reported VTs and forms. Additionally, stimulus context seems to be affecting the outcomes of the study. Comparing conditions Low & High with High/Low & High/Low, even though listeners are only reporting one of the pitches, the addition of another pitch (as in conditions High/Low & High/Low) resulted in an increase of the number of VTs and forms reported. This suggests a different type of regrouping of the speech sounds between the two pairs of conditions, and is likely to be influenced by the nature of the two sequences, where both were present in both ears at the same time (unlike for a dichotic condition). It can be concluded that the effect of sequence pitch observed in Experiment 1 was not attributable to the resynthesis of the stimulus words per se, but rather to the demand characteristics of the task itself. While Experiments 1 and 2 used two concurrent sequences played at the same time, Experiments 3 and 4 used single-sequence presentations.

The connection between the VTE and auditory stream segregation was suggested by evidence that formant transitions facilitate the integration of speech segments into a single coherent stream in a rapidly repeating sequence of CV syllables (Cole & Scott, 1973). Experiment 3 looked at the role of formant transitions in the context of the VTE, using precisely controlled digital editing to manipulate the formant transitions between the initial segments of monosyllabic words. Six CVC words with strong formant transitions between the initial consonant and vowel were paired with another set which had weak formant transitions. Each set of six words was used to derive another in which the CV transitions were edited out and replaced with samples selected from the neighbouring steady-state portions. VTs obtained for 3-minute sequences of the edited and unedited versions were compared. Listeners reported more Forms in the edited than in the unedited case for the strong-transition words, but not for those with weak transitions. The results supported the notion that perceptual re-grouping influences the VTE and indicate that the effect of removing formant transitions reported by Cole and Scott was not due to an artefact of analogue tape splicing. The findings support earlier studies suggesting that formant transitions play an important role in binding disparate speech segments together into a coherent whole – e.g., the study by Dorman et al. (1975), which showed that sequences of vowels linked by smooth transitions tended to fuse into a single stream. Additionally, results from Experiment 3 confirm the earlier indications from Experiments 1 and 2 that Forms are more likely to reveal changes related to grouping. Forms appears to provide a more stable measure with a smaller variance compared to VTs, which in turn might be affected by listeners not always reporting every change in the stimulus (e.g. during rapid oscillation between two forms, see Ditzinger, Tuller & Kelso, 1997).

The Gestalt principle of good continuity and its role in the cohesiveness of speech was further investigated in Experiment 4. Listeners were exposed to single-sequence recordings of words with rising or falling pitch contours, arranged such that the sequence had either a continuous or discontinuous pitch contour across the boundaries between adjacent tokens. The results were consistent with the idea that the pitch contour contributes to the perceptual cohesion of speech. There were significantly fewer forms reported in the continuous pitch case in comparison to the discontinuous condition. In summary, for experiments 3 and 4 there is clear evidence that manipulation of strong formant transitions and smoothness of change in the pitch contour influence the number of forms heard. Hence, the results are consistent with the hypothesis that formant transitions between phonetic segments and the continuity of the pitch contour both influence the regrouping of phonetic segments.

Experiments 5 and 6 explored the effects of primitive grouping cues on a phenomenon closely related to the VTE, called the PTE. In the PTE, listeners experience vowel sequences as verbal forms – syllables and words. Here, four-vowel sequences were used. Using a similar method to that of Bregman et al. (1990), the relationship between timbre and F0 cues, and between timbre and ITD cues, were investigated in Experiments 5 and 6, respectively. Specifically, F0 or ITD cues were introduced that either supported (congruent) or opposed within-sequence pairings of vowels based on timbre cues. Statistical analyses of the differences between baseline, opposing, and congruent conditions revealed relatively little impact of these grouping cues on the number of VTs and forms. However, a more descriptive analysis of the data indicated that there were differences in the particular forms reported between conditions. Nonetheless, it was also apparent that the small number of specific forms that were common to any two conditions accounted for a relatively high proportion of the total number of responses.

Limitations, future research, and concluding remarks

It is important to note that the VTE, and to a lesser extent the PTE, are quite complex linguistic phenomena. The stimulus words themselves can be controlled precisely using digital manipulation with respect to various grouping cues such as F0 differences or lateralization by ITD cues. The reported VTs and Forms, however, will be influenced by additional higher order factors that are unavoidable when dealing with a speech signal. Inevitably, isolating the influence of primitive grouping factors from higher order linguistic processes in the VTE would pose a considerable challenge. Nevertheless, the current set of experiments has shown that the VTE does appear to respond to primitive factors that are known to influence auditory grouping. The use of prolonged stimulus exposure enables the process of streaming to occur. Although the open ended nature of VTE tasks can in principle result in any number of percepts occurring, there is some evidence that the switching between lexical interpretations shows properties in common with the perception of visual reversible figures (Ditzinger, Tuller & Kelso, 1997; Ditzinger, Tuller, Haken & Kelso, 1997). Rather than reporting a large number of transformations with equal frequency, studies by Ditzinger et al. show that listeners often tend to switch between one pair of percepts. This, however, still contrasts with the classic bi-stability examples in vision, such as the face-vase illusion (the Rubin’s vase, see Schwartz et al, 2012). Compared with the VTE visual examples often involve a relatively short exposure time leading to switching between a very limited number of percepts (usually two). In their early paper, Warren and Gregory (1958) highlighted the differences in both approaches, saying that (i) VTs occur over a wide range of stimuli like syllables, words or phrases; (ii) they sometimes involve considerable distortions from the original percept; (iii) responses vary considerably between participants, and (iv) they generally produce many forms in 2 or 3 minutes, whereas with reversible figures there are typically only two forms possible.

The concept of multistability of perception is an interesting one in the context of the VTE. In principle, listeners can come up with an infinite number of VTs, and so it is a clear that the percepts are multistable in nature. However, as mentioned earlier, there is a suggestion that for prolonged periods of time participants experience flipping between two dominant forms from among the total set of forms reported (Ditzinger, Tuller & Kelso, 1997). This would imply that the VTE is mainly a bi-stable phenomenon despite the possibility of multistable percepts occurring. For example a listener may hear six different forms (A, B, C, D, E, and F) when hearing a particular stimulus repeated. While responses to A and F can be relatively rare, for long periods listeners might experience back and forth alternation between B and C. This in turn may change at some point to back and forth alternation between D and E suggesting a "fatigue" of the previous lexical items B and C. Hence, listeners can spend most of their time hearing bi-stable pairs but the multistability can be manifested in a shift to a different pair as unlike simple high and low tones there are many different ways in which the acoustically more complex speech segments can be re-grouped.

Within the concept of the build-up of stream segregation, there is an initial period when listeners perceive the 'veridical' percept. This build-up lasts for a period of around 30 s, after which segregation of speech occurs and listeners report VTs. In their study of the bi-stability of stream segregation, Pressnitzer & Hupé (2006) suggested that after the initial build-up, there is no significant difference between the duration of successive perceptual states. In the context of VTE it would mean that for any set of responses to sequence of speech sounds, after the identification of the two most dominant forms, it should be possible to investigate whether the proportion of time spent experiencing the two forms is the same.

To further develop the current approach, future investigations might involve a more in-depth analysis of the types of forms reported by listeners in different experimental conditions. More qualitative analysis in terms of the phonetic structure of the VT forms reported might prove illuminating. As an example, in order to elucidate the difference between the opposing and congruent conditions in the last two experiments, making phonetic representations of the forms reported and analysing them using principal components analysis (e.g., McAdams, 1999) may help in finding the underlying dimensions and relationships that describe the variability in the data, e.g. that certain vowel sounds tend only to pair up with particular consonants (cf. Chalikia & Warren, 1991).

In conclusion, the VTE and PTE can be used experimentally to successfully investigate the processes involved in auditory grouping. In particular, the results of the experiments reported in this thesis have supported and extended the studies by Ditzinger, Tuller and Kelso (1997) and Pitt and Shoaf (2002), who suggested that the perceptual regrouping of speech sounds plays a key role in the VTE.

References

Anstis, S., & Saida, S. (1985). Adaptation to auditory streaming of frequency-modulated tones. *Journal of Experimental Psychology: Human Perception & Performance, 11*, 257-271.

Assmann, P. (1999). Fundamental frequency and the intelligibility of competing voices. *Proceedings of the 14th International Congress of Phonetic Sciences, San Francisco, Aug. 1-7* (pp. 179-182).

Assmann, P., & Summerfield, Q. (2004). Perception of speech under adverse conditions. In S. Greenberg, A. Popper, W. Ainsworth, & R. Fay (Eds.), *Speech Processing in the Auditory System* (Vol. 10, pp. 231-308). Springer-Verlag.

Atal, B. S., & Hanauer, S. L. (1971). Speech analysis and synthesis by linear prediction of the speech wave. *Journal of the Acoustical Society of America*, *50(2)*, 637-655.

Bailey, P. J., Summerfield, Q., & Dorman, M. F. (1977). On the identification of sine-wave analogues of certain speech sounds. *Haskins Laboratories Status Report on Speech Research, SR-51/52*, 1-25.

Bashford, J. A., Warren, R. M., & Lenz, P. W. (2006). Polling the effective neighbourhoods of spoken words with the verbal transformation effect. *Journal of the Acoustical Society of America: Express Letters, 119*, EL55-EL59.

Bashford, J. A., Warren, R. M., & Lenz, P. W. (2009). The spread and density of the phonological neighbourhood can strongly influence the verbal transformation illusion. *Proceedings of Meetings on Acoustics*, *6(1)*, 060002 (8 pages).

Basirat, A., Schwartz, JL., & Sato, M. (2012). Perceptuo-motor interactions in the perceptual organization of speech: evidence from the verbal transformation effect. *Philosophical Transactions of The Royal Society B, 367*, 965-976.

Beauvois, M. W., & Meddis, R. (1991). A computer model of auditory stream segregation. *The Quarterly Journal of Experimental Psychology*, *43A(3)*, 517-541.

Bench, J., Kowal, A., & Bamford, J. (1979). The BKB (Bamford-Kowal-Bench) sentence lists for partially-hearing children. *British Journal of Audiology*, *13*, 108-112.

Bird, J., & Darwin, C. J. (1998). Effects of a difference in fundamental frequency in separating two sentences. In A. R. Palmer, A. Rees, A. Q. Summerfield, & R. Meddis (Eds.), *Psychophysical and physiological advances in hearing*. London: Whurr.

Boersma, P., & Weenink, D. (2009). PRAAT: doing phonetics by computer (Version 5.1.02) [Computer program]. Retrieved from http://www.praat.org/

Bregman, A.S. (1978). Auditory streaming is cumulative. *Journal of Experimental Psychology: Human Perception & Performance, 4,* 380-387.

Bregman, A. S. (1990). *Auditory scene analysis: The perceptual organization of sound* . Cambridge, Massachusetts: The MIT Press.

Bregman, A. S. (2004). Auditory scene analysis. In N. J. Smelzer & P. B. Baltes (Eds.), *International Encyclopedia of the Social and Behavioral Sciences* (pp. 940-942). Amsterdam: Pergamon (Elsevier).

Bregman, A. S., & Ahad, P. (1996). *Demonstrations of auditory scene analysis: the perceptual organization of sound*. MIT Press.

Bregman, A. S., & Campbell, J. (1971). Primary auditory stream segregation and perception of order in rapid sequences of tones. *Journal of Experimental Psychology, 89(2),* 244-249.

Bregman, A. S., & Dannenbring, G. (1973). The effect of continuity on auditory stream segregation. *Perception & Psychophysics*, *13*, 308-312.

Bregman, A. S., Liao, C., & Levitan, R. (1990). Auditory grouping based on fundamental frequency and formant peak frequency. *Canadian Journal of Psychology*, *44(3),* 400-413.

Bregman, A. S., & Pinker, S. (1978). Auditory streaming and the building of timbre. *Canadian Journal of Psychology*, *32*, 19-31.

Broadbent, D. E., & Ladefoged, P. (1957). On the fusion of sounds reaching different sense organs. *Journal of the Acoustical Society of America1*, *29*, 708-710.

Brokx, J. P. L., & Nooteboom, S. G. (1982). Intonation and the perceptual separation of simultaneous voices. *Journal of Phonetics, 10*, 23-36.

Bronkhorst, A. W., & Plomp, R. (1992). Effect of multiple speechlike maskers on binaural speech recognition in normal and impaired hearing. *Journal of the Acoustical Society of America*, *92*, 3132-3139.

Brunstrom, J. M., & Roberts, B. (1998). Profiling the perceptual suppression of partials in periodic complex tones: Further evidence for a harmonic template. *Journal of the Acoustical Society of America*, *104(6)*, 3511-3519.

Carlyon, R. P., Cusack, R., Foxton, J.M., & Robertson, I.H. (2001). Effects of attention and unilateral neglect on auditory stream segregation. *Journal of Experimental Psychology: Human Perception & Performance*, *27*, 115-127.

Chalikia, M. H., & Warren, R. M. (1991). Phonemic transformations: Mapping the illusory organisation of steady-state vowel sequences. *Language and Speech*, *34(2),* 109-143.

Chalikia, M. H., & Warren, R. M. (1994). Spectral fissioning in phonemic transformations. *Perception and Psychophysics*, *55(2)*, 218-226.

Cherry, C. (1953). Some experiments on the recognition of speech, with one and two ears. *Journal of the Acoustical Society of America*, *25*, 975-979.

Clegg, J. M. (1971). Verbal transformations on repeated listening to some English consonants. *British Journal of Psychology*, *62(3)*, 303-309.

Cole, R. C., & Scott, B. (1973). Perception of temporal order in speech: The role of vowel transitions. *Canadian Journal of Psychology*, *27(4)*, 441-449.

Culling, J. F., & Summerfield, Q. (1995). Binaural grouping of complex sounds: absence of across frequency grouping by common inter-aural delay. *Journal of the Acoustical Society of America*, *98*, 785-797.

Cusack, R., Deeks, J., Aikman, G., & Carlyon, R. P. (2004). Effects of location, frequency region, and time course of selective attention on auditory scene analysis. *Journal of Experimental Psychology: Human Perception & Performance, 30(4),* 643-656.

Darwin, C.J. (1984). Perceiving vowels in the presence of another sound: constraints on formant perception. *Journal of the Acoustical Society of America, 76,* 1636-1647.

Darwin, C. J. (1997). Auditory grouping. *Trends in Cognitive Sciences, 1(9),* 327-333.

Darwin, C. J. (2008). Listening to speech in the presence of other sounds. *Philosophical Transactions of The Royal Society B, 363(1493),* 1011-1021.

Darwin, C. J., & Bethell-Fox, C. E. (1977). Pitch continuity and speech source attribution. *Journal of Experimental Psychology: Human Perception and Performance, 3(4),* 665-672.

Darwin, C. J., & Hukin, R. W. (1999). Auditory objects of attention: The role of interaural time differences. *Journal of Experimental Psychology: Human Perception and Performance, 25(3),* 617-629.

Demany, L. (1982). Auditory stream segregation in infancy. *Infant Behavior and Development, 5,* 261-276.

Denham, S.L., & Winkler, I. (2006). The role of predictive models in the formation of auditory streams. *Journal of Physiology – Paris, 100,* 154-170.

Deutsch, D. (1979). Binaural integration of melodic patterns. *Perception and Psychophysics, 25,* 399-405.

Ditzinger, T., Tuller, B., Haken, H., & Kelso, J. A. S. (1997). A synergetic model for the verbal transformation effect. *Biological Cybernetics, 77,* 31-40.

Ditzinger, T., Tuller, B., & Kelso, J. A. S. (1997). Temporal patterning in an auditory illusion: the verbal transformation effect. *Biological Cybernetics, 77,* 23-30.

Dorman, M. F., Cutting, J. E., & Raphael, L. J. (1975). Perception of temporal order in vowel sequences with and without formant transitions. *Journal of Experimental Psychology: Human Perception and Performance, 104(2),* 121-129.

Evans, C. R., & Wilson, J. (1968). Subjective changes in the perception of consonants when presented as “stabilized auditory images.” *Division of Computer Science Publication No. 41*, National Physical Laboratory, England.

Fant, G. (1960). *Acoustic theory of speech production*. Mouton, The Hague.

Fenelon, B., & Blayden, J. A. (1968). Stability of auditory perception of words and pure tones under repetitive stimulation in neural and suggestibility conditions. *Psychonomic Science, 13,* 285-286.

Goldstein, L. M., & Lackner, J. R. (1973). Alterations of the phonetic coding of speech sounds during repetition. *Cognition*, *2,* 279-297.

Guilford, J. P., & Nelson, H. M. (1936). Changes in the pitch of tones when melodies are repeated. *Journal of Experimental Psychology*, *19*, 193-202.

Henke, W. L. (1997). MITSYN: A coherent family of high-level languages for time signal processing, software package. Belmont, MA.

Hukin, R. W., & Darwin, C. J. (1995). Effects of contralateral presentation and of interaural time differences in segregating a harmonic from a vowel. *Journal of the Acoustical Society of America*, *98*, 1380-1387.

Kaminska, Z., & Mayer, P. (2002). Changing words and changing sounds: A change of tune for verbal transformation theory? *European Journal of Cognitive Psychology, 14(3),* 315-333.

Kaminska, Z., Pool, M., & Mayer, P. (2000). Verbal transformation: habituation or spreading activation? *Brain and Language, 71*, 285-298.

Kashino, M., & Kondo, H. M. (2012). Functional brain networks underlying perceptual switching: auditory streaming and verbal transformations. *Philosophical Transactions of The Royal Society B, 367,* 977-987.

Keppel, G. (1991). *Design and analysis: A researcher’s handbook* (3rd ed.). Englewood Cliffs, N.J.: Prentice Hall.

Kimura, D. (1961). Cerebral dominance and the perception of verbal stimuli. *Canadian Journal of Psychology*, *15(3)*, 166-171. University of Toronto Press.

Koffka, K. (1935). *Principles of Gestalt Psychology* (1st ed.). New York: Harcourt.

Lackner, J. R., Tuller, B., & Goldstein, L. M. (1977). Some aspects of the psychological representation of speech sounds. *Perceptual and Motor Skills*, *45*, 459-471.

Lass, N. J., & Gasperini, R. M. (1973). The verbal transformation effect: a comparative study of the verbal transformations of phonetically trained and nonphonetically trained listeners (A). *The Journal of the Acoustical Society of America*, *53(1)*, 369.

Lass, N. J., & Golden, S. S. (1971). The use of isolated vowels as auditory stimuli in eliciting the verbal transformation effect. *Canadian Journal of Psychology*, *25*, 349-359.

Lass, N. J., West, L. K., & Taft, D. D. (1973). A non-verbal analogue of the verbal transformation effect. *Canadian Journal of Psychology*, *27*, 273-279.

McAdams, S. (1999). Perspectives on the contribution of timbre to musical structure. *Computer Music Journal*, *23(3)*, 85-102.

Miller, G. A., & Heise, G. A. (1950). The trill threshold. *The Journal of the Acoustical Society of America*, *22*, 637-638.

Moulines, E., & Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, *9*, 453-467.

Naeser, M. A., & Lilly, J. C. (1970). Preliminary evidence for a universal feature detector system: perception of the repeating word (A). *Journal of the Acoustical Society of America*, *48*, 85.

Natsoulas, T. (1965). A study of the verbal-transformation effect. *American Journal of Psychology*, *78*, 257-263.

Ohde, R. N., & Sharf, D. J. (1979). Relationship between adaptation and the percept and transformation of stop consonant voicing: Effects of number of repetitions and intensity of adaptors. *Journal of the Acoustical Society of America*, *66*, 30-45.

Ortmann, O. (1926). On the melodic relativity of tones. *Psychological Monographs*, *35*, 1-47.

Perl, N. T. (1970). The application of the verbal transformation effect to the study of cerebral dominance. *Neuropsychologia, 8*, 259-261.

Pitt, M. A., & Shoaf, L. (2002). Linking verbal transformations to their causes. *Journal of Experimental Psychology, 28(1),* 150-162.

Pressnitzer, D., & Hupé, J.M. (2006). Temporal dynamics of auditory and visual bistability reveal common principles of perceptual organization. *Current Biology, 16,* 1351-1357.

Remez, R. E., Rubin, P. E., Berns, S. M., Pardo, J. S., & Lang, J. M. (1994). On the perceptual organisation of speech. *Psychological Review*, *101*, 129-156.

Remez, R. E., Rubin, P. E., Pisoni, D. B., & Carrell, T. D. (1981). Speech Perception without Traditional Speech Cues. *Science*, *212(4497),* 947-950.

Roberts, B., Summers, R.J., & Bailey, P.J. (2010). The perceptual organization of sine-wave speech under competitive conditions. *Journal of the Acoustical Society of America, 128(2),* 804-817.

Shackleton, T. M., & Meddis, R. (1992). The role of interaural time difference and fundamental frequency difference in the identification of concurrent vowel pairs. *Journal of the Acoustical Society of America*, *91*, 3579-3581.

Shoaf, L., & Pitt, M. A. (2002). Does node stability underlie the verbal transformation effect? A test of node structure theory. *Perception & Psychophysics*, *64(5)*, 795-803.

Schwartz, J., Grimault, N., Hupé, J-M., Moore, B. C. J. & Pressnitzer, D. (2012) Multistability in perception: binding sensory modalities, an overview. *Philosophical transactions of the Royal Society of London Series B Biological sciences*, *367(1591)*, p. 896-905.

Snedecor, G. W., & Cochran, W. G. (1967). *Statistical methods*. Iowa State University Press.

Tuller, B., Ding, M., & Kelso, J. A. S. (1997). Fractal timing of verbal transforms. *Perception*, *26*, 913-928.

Van Noorden, L. P. A. S. (1975). *Temporal coherence in the perception of tone sequences*. Eindhoven University of Technology, The Netherlands.

Warren, R. M. (1961a). Illusory changes of distinct speech upon repetition - the verbal transformation effect. *British Journal of Psychology*, *52(3)*, 249-258.

Warren, R. M. (1961b). Illusory changes in repeated words: differences between young adults and the aged. *The American Journal of Psychology*, *74(4)*, 506-516.

Warren, R. M. (1968). Verbal transformation effect and auditory perceptual mechanisms. Psychological Bulletin, 70*(4)*, 261-270.

Warren, R. M. (1996). Auditory illusions and perceptual processing of speech. In N. J. Lass (Ed.), *Principles of Experimental Phonetics* (pp. 435-466). St. Louis: Mosby.

Warren, R. M. (2008). *Auditory perception: An analysis and synthesis* (3rd ed.). New York: Cambridge University Press.

Warren, R. M., & Ackroff, J. M. (1976). Dichotic verbal transformations and evidence for separate processors for identical stimuli. *Nature*, *259*, 475-477.

Warren, R. M., Bashford, J. A., & Gardner, D. A. (1990). Tweaking the lexicon: Organization of vowel sequences into words. *Perception & Psychophysics*, *47(5)*, 423-432.

Warren, R. M., & Gregory, R. L. (1958). An auditory analogue of the visual reversible figure. *The American Journal of Psychology*, *71(3)*, 612-613.

Warren, R. M., Healy, E. W., & Chalikia, M. H. (1996). The vowel-sequence illusion: Intrasubject stability and intersubject agreement of syllabic forms. *Journal of the Acoustical Society of America*, *100(4)*, 2452-2461.

Warren, R. M., Obusek, C. J., Farmer, R. M., & Warren, R. P. (1969). Auditory sequence: confusion of patterns other than speech or music. *Science*, *164(3879)*, 586-587.

Warren, R. M., & Warren, R. P. (1966). A comparison of speech perception in childhood, maturity, and old age by means of the verbal transformation effect. *Journal of Verbal Learning and Verbal Behaviour*, *5*, 142-146.

Warren, R. M., & Warren, R. P. (1970). Auditory illusions and confusions. *Scientific American*, *223*, 30-36.

Wessel, D. (1979). Timbre space as a musical control structure. *Computer Music Journal*, *3,* 45-52.

Zuck, D. (1992). The verbal transformation effect: auditory illusions as an index of lexical processing and homolog activation. *Brain and Language, 43*, 323-335.

Appendices

Appendix 1 – *Forms reported by participants for each stimulus word in Experiment 1.*

"noise", high pitch
annoy, annoys, do you know where he is, nice, night, nine, no, no his, no use, no yes, norris, noy, snore-yeeu, yes no

"noise", low pitch
annoy, annoys, die, nay, nice, night, nine, no, no use, norris, noy

"flame", high pitch
bates bake, bates bay, betley, blame, brain, brain-flame, delaying, faced, fame, flame flewn, flane, flay, flying, flying by, flying-brain, frame, frying, lame, lay-in, my, pain, paste, paying, plaim, plane, plane flame, play, playing, same, sane, slain, slay, stain, thank vee, thank view, train

"flame", low pitch
blame, bloody, brain, christ-man, cried, fame, faying, flame flame, flame flewn, flane, fley, flying, lay-in, my, plane, playing, same, same flame, slave, slay, slaying, staim, stain, st-lain, thank for you

"face", high pitch
bake, bay, bleet, by, dave, envy, fade-in, fair, faith, fame, fear, fey, fire, fleet, grey, hi, khey, paste, pay, prayer, prior, pry, safe, same, save, say, science, science a, supree, they pay, train, tray, try, vague, vey

"face", low pitch
bake, base, bay, de-face, faced, faith, fake, fang, fate, fav, fent, fey, hate, hi, pace, paste, pasted, pay, supree, taste, thing, tray, vague, vapour, vase

"sleep", high pitch
asleep, beep, beep beep, belongs to me, bleak, bleep, blink, see, clean, clee, delete, eat, feed, fleet, immensely, leap, let's sleep, pea, pleh, plea, please, please speak, pleat, seat, seed, slee, sleepy, sleet, speak to me, squeeze, stee, sweep, three

"sleep", low pitch
asleep, beak, beep, bleep, blink, delete, eat, feet, fleet, late, leap, lee, link, me, minced, minced meat, plea, please, please speak, pleat, seed, seep, slee, sleepy, sleet, snake, speak, speak link, speak to me, sweet, sweet sleep, vee, weak

"see", high pitch
bee, baby, beer, dee, dear, easy, pee, sear, seat, see her, seed, seeing, seem, seen, sig, sing, tee, team, tear, tee-in, theme, vee, zee

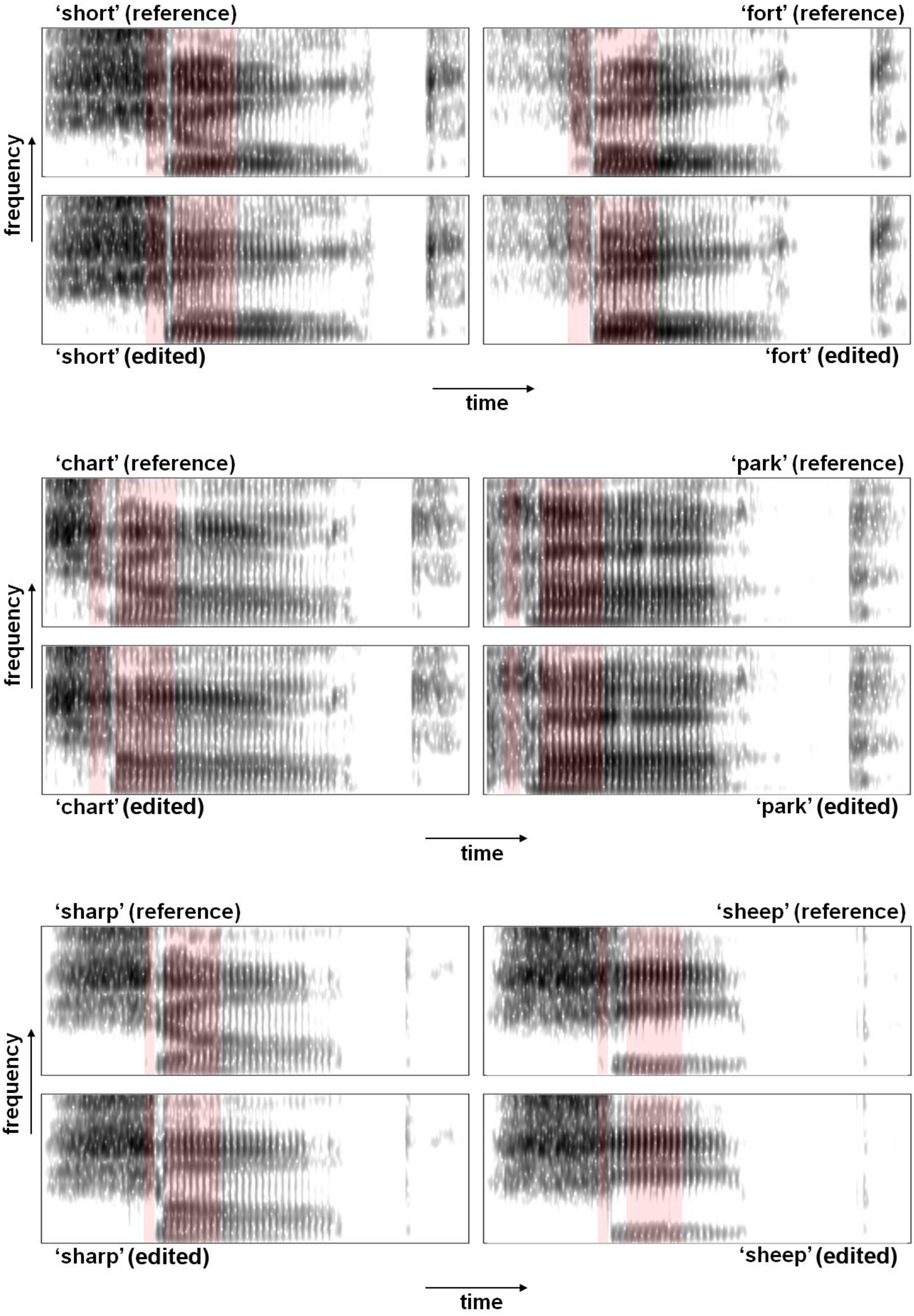
"see", low pitch

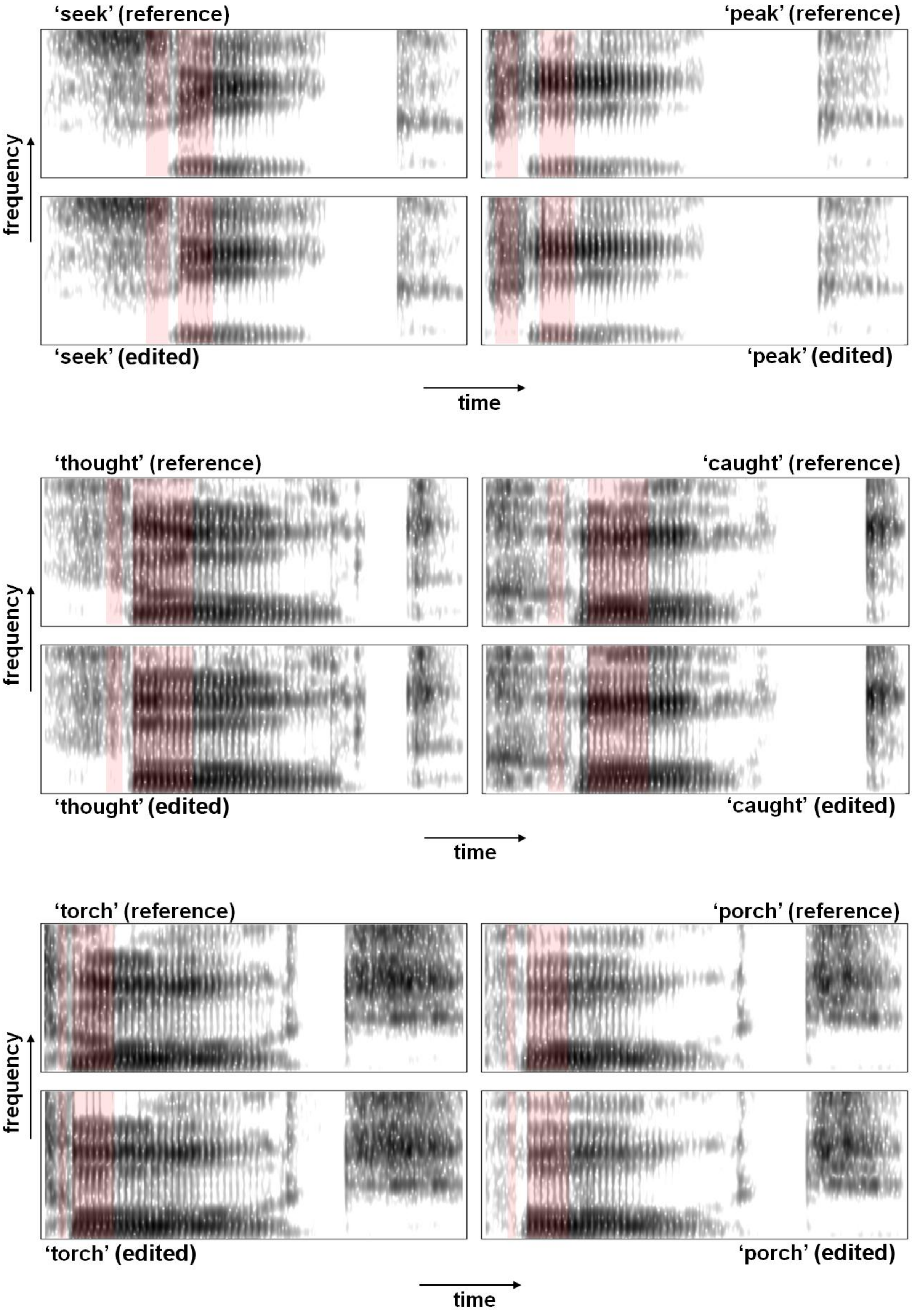
bee, been, beer, dee, deeds, fee, fee fee, feet, pee, seap, sear, seat, see id, seed, seeing, studying, tee, tear, twenty three, twenty two, vee, zee

“right”, high pitch
blank, bright, dry, light, ride, rider, ripe, rye, to write, tright, try, white

“right”, low pitch
blank, blanked, bright, dry, light, night, ride, ripe, rye, to write, try

Appendix 2 – *Spectrograms of the stimulus words used in Experiment 3. Edited regions are indicated in red, one for the initial consonant and one for the vowel.*




‘seek’ (reference)
‘peak’ (reference)
frequency
‘seek’ (edited)
‘peak’ (edited)
time
‘thought’ (reference)
‘caught’ (reference)
frequency
‘thought’ (edited)
‘caught’ (edited)
time
‘torch’ (reference)
‘porch’ (reference)
frequency
‘torch’ (edited)
‘porch’ (edited)
time

Appendix 3 – *Raw data from Experiment 3 – Formant Transitions. 'R' is a total number of responses for a given form and 'L' is a total number of listeners who reported that particular form.*

		REFERENCE
R	L	short
1	1	bed
2	2	caught
2	1	fill
2	1	filter
1	1	float
1	1	gont
4	1	got
9	2	salt
6	1	salts
3	2	sholl
4	2	shore
61	10	short
2	2	shorter
3	1	shorts
8	3	shoulder
12	3	show
2	1	show what
1	1	showed
4	2	slaughter
1	1	slope
5	1	slow
1	1	sort short
2	2	sought
1	1	take it all
5	1	taught
4	2	thought
1	1	tickle
6	2	to saw
4	3	to shore
5	1	to show
3	1	to slo
2	1	too short
1	1	too sure
1	1	what
1	1	your
1	1	you're short

		EDITED
R	L	short
1	1	assault
1	1	bed
1	1	belot
1	1	bored
9	2	bought
1	1	builter
4	1	daughter
2	2	door
2	1	door to door
1	1	dow-what
1	1	fault
1	1	i thought
3	1	it up
6	1	it's up
4	1	persil
1	1	pest-ee-o
20	4	salt
9	1	salts
2	1	salty
1	1	saw-it
4	1	shalot
1	1	shaloter
1	1	shautar
1	1	shawater
1	1	shollty
1	1	shoot
1	1	shorp
53	11	short
13	4	shorter
4	1	shorts
1	1	shorty
2	2	shot
1	1	shots
20	3	shoulder
1	1	shoulter
1	1	shout
2	1	shoutar
1	1	show
1	1	show id
1	1	show up
1	1	show-oot
1	1	slaughter
1	1	so
1	1	so old
1	1	sort
1	1	sorter
1	1	sought
15	5	thought
5	2	to door
1	1	to saw
1	1	to shaw
2	2	to sholl
5	2	to shore
8	3	to short
2	1	to shovel
6	1	to show
2	1	to sol
1	1	too short
1	1	tusle

		REFERENCE
R	L	fort
1	1	ball
1	1	bought
1	1	clock
1	1	dee four
9	4	default
4	1	effany
1	1	faggot
5	1	fall
41	8	fault
2	1	faulty
1	1	feed
1	1	feel
1	1	feeling
1	1	feet
1	1	fill
4	1	fill-lat
1	1	fill-lod
2	1	filter
2	1	flaw
9	1	float
5	1	flood
6	1	flow
3	1	flower
1	1	foe
1	1	fold
1	1	follow on
1	1	follow up
1	1	foot
5	1	for
3	1	forgot
1	1	fork
15	3	fort
8	1	forth
1	1	forts
5	2	fought
2	2	full
1	1	phone
1	1	salt
1	1	the floor
3	2	the fort
1	1	the thought
6	6	thought
2	1	tiffany
1	1	to fault
1	1	to foil
2	1	to fold
1	1	to fork
1	1	to full
1	1	volt
7	1	vote
10	1	walt
5	2	what
1	1	wonderful

		EDITED
R	L	fort
1	1	afeeny
1	1	below
1	1	big
2	1	big thing
3	1	bloat
1	1	blow
1	1	boat
7	1	bought
9	1	builta
6	3	default
1	1	defloor
2	1	fall
29	6	fault
2	1	faults
1	1	feel
1	1	feeny
1	1	fillip
1	1	finger
1	1	flaw
1	1	flawed
5	1	float
3	1	flont
5	3	floor
1	1	flop
4	2	flow
3	1	flower
1	1	foe
1	1	fold
1	1	folder
1	1	font
2	2	for
9	4	fort
7	1	forth
3	2	fought
2	2	full
1	1	fulont
2	1	guilta
3	1	loat
1	1	pink
10	1	salt
1	1	slow
1	1	t-feeny
1	1	thank you
3	1	the floor
3	1	the flow
1	1	the forth
3	1	thing
10	2	think
6	3	thought
5	1	tiffany
1	1	ting
6	1	tink
1	1	to fall
1	1	to fault
1	1	to fold
1	1	to fort
4	2	vault
1	1	vill
1	1	vo
12	1	volt
1	1	vot
12	1	vote
1	1	walt
1	1	what

		REFERENCE
R	L	chart
4	1	bar
1	1	bar chart
1	1	barch
1	1	bart
1	1	canarl
4	1	cannot
8	1	cant
4	1	car
1	1	carl
2	2	cart
1	1	caught
2	1	cha-heart
5	1	chalk
1	1	chant
6	3	char
1	1	chargoo
2	1	chark
4	2	charl
63	11	chart
1	1	che-art
2	1	chee-ba-heart
1	1	chew
33	7	child
1	1	chillot
1	1	chin up
1	1	geheart
1	1	go home
2	1	mark
1	1	not
3	1	park
2	1	part
1	1	patrol
1	1	pile
1	1	rot
1	1	see-ba-heart
1	1	sharlet
4	1	sha-up
3	1	talk
1	1	taught
5	1	the child
1	1	the heart
1	1	the hot
1	1	to chant
2	1	to char
1	1	to charl
1	1	to sharlet
1	1	trial

		EDITED
R	L	chart
1	1	arrow
2	1	ball
1	1	balt
3	1	bar
2	1	bar chart
3	1	bar charts
1	1	barrel
1	1	barrot
1	1	bart
3	1	beep
1	1	bold
8	2	cant
4	3	car
1	1	carlet
4	2	cart
1	1	chance
1	1	chap
2	1	char
2	1	chark
2	1	charl
1	1	charlt
43	11	chart
1	1	charts
24	5	child
1	1	chillout
1	1	current
1	1	cut
1	1	dark
2	1	hard
1	1	hark
3	3	heart
4	1	i cant
1	1	it's dark
1	1	k-char
1	1	pard
3	1	park
1	1	p-hark
1	1	scarlet
1	1	sharlet
1	1	star
1	1	starlet
5	1	the child
1	1	to chart
1	1	to trial
1	1	to try out
1	1	trial
1	1	troll
2	1	try out
3	1	what to do

		REFERENCE
R	L	park
1	1	a car
2	1	a park
1	1	acant
2	2	ark
3	1	beep
4	1	cannot
4	2	can't
6	2	car
1	1	car par
4	2	car park
1	1	card
1	1	carrot
1	1	caught
5	1	clark
2	1	clock
1	1	ha-ha-ha-ha
4	2	har
4	2	hark
1	1	honk
1	1	how long
1	1	i can't
1	1	onk
4	2	pack
2	2	par
43	11	park
2	1	park car
3	2	park the car
2	1	parker
1	1	pellunk
1	1	pok
2	1	pork
1	1	punk
3	2	the car
1	1	the park
2	1	to park the car

		EDITED
R	L	park
1	1	acan
3	1	acant
2	1	acar
2	1	bark
1	1	beep
4	2	can't
11	3	car
1	1	car par
1	1	car park
2	1	cark
2	1	clark
1	1	haha
1	1	har
7	3	hark
1	1	honk
1	1	how long
4	1	hug
1	1	i can't
5	1	k-par
2	1	pack
5	3	par
56	11	park
1	1	park car
2	2	park the car
7	2	parker
1	1	pub
3	1	puck
5	1	the car
1	1	the park
1	1	to park

		REFERENCE
R	L	sharp
3	3	chop
1	1	darp
1	1	drop
1	1	forgot
1	1	garber
10	2	garp
2	1	get up
1	1	give up
1	1	it's sharp
1	1	jarp
1	1	jump
1	1	pesarr
2	1	pixie bra
1	1	pixie mail
1	1	p-shar
1	1	push up
3	1	saap
2	1	scarper
1	1	sha
23	3	shark
70	11	sharp
10	4	sharper
1	1	sha-up
10	3	shop
2	1	shot
1	1	shut
4	2	shut up
1	1	slow
1	1	stop
1	1	sunk
1	1	top shop
1	1	yarp

		EDITED
R	L	sharp
1	1	ab up
3	1	arp
1	1	barber
3	1	barp
1	1	barper
4	1	beep
1	1	bizzare
2	1	bop
1	1	de sa
1	1	de sar
7	4	harp
1	1	hope
3	1	it up
2	1	it's up
4	1	pahper
1	1	piss ah
3	1	piss ar
4	1	pub
1	1	push
1	1	push them up
1	1	push up
2	2	shap
1	1	shark
59	10	sharp
8	3	sharper
3	1	sha-up
2	1	ship
28	4	shop
1	1	short
11	2	shut up
2	1	slop
1	1	snap
1	1	sop
1	1	stand up
1	1	starbucks
6	3	stop
1	1	stop her
4	1	stop it
1	1	sum-up
2	1	sup
1	1	taught
1	1	teh sa
1	1	tom
1	1	tomp
1	1	top shop
1	1	up
1	1	what's up

		REFERENCE
R	L	sheep
1	1	buffy
22	6	cheap
2	2	cheaper
3	1	chip
2	2	deep
1	1	deep ship
2	1	deeper
1	1	dep
1	1	feet
5	2	heap
4	1	hep
1	1	jeep
1	1	keep
1	1	pea
2	1	perceive
1	1	persue
2	1	persuit
7	1	press u
2	1	proceed
6	1	pushy
1	1	see
6	4	seep
1	1	she
1	1	shed
56	11	sheep
1	1	sheep asleep
2	2	sheeper
4	1	ship
7	2	sleep
2	1	sleeper
1	1	suit

		EDITED
R	L	sheep
1	1	achieve
3	1	ashleep
4	1	asleep
1	1	beep
35	7	cheap
2	1	cheaper
3	1	cheat
1	1	chin
3	3	chip
5	2	deep
3	1	deep ship
3	1	deeper
1	1	flu
4	2	heap
1	1	heaper
2	1	hep
3	1	perceive
8	1	persue
1	1	persuit
1	1	proceed
1	1	p-she
7	2	pushy
1	1	seal
2	2	she
76	11	sheep
27	4	ship
2	1	shleep
1	1	silk
1	1	sip
5	4	sleep
1	1	sleeper
1	1	slip
1	1	syoo

		REFERENCE
R	L	seek
1	1	can't see
3	1	could see
1	1	courtesy
1	1	feak
2	1	good to see
4	1	gutsee
11	3	kah-see
2	2	kah-seek
4	1	khasee
8	1	k-seat
2	1	sea seek
2	1	seat
12	4	see
71	11	seek
9	4	seeker
27	3	sick
5	2	sink
1	1	sint
1	1	sy-heek
1	1	sy-he-hack
6	2	think
1	1	thint
2	1	zieg

		EDITED
R	L	seek
1	1	aseek
1	1	cause he
1	1	conseal
4	1	could see
1	1	dick
3	1	faik
10	2	feak
1	1	feel
1	1	ghasier kasee
1	1	good to see
11	1	gusee
3	1	gutsee
5	1	kasee
1	1	kha-seal
8	3	kha-see
1	1	kha-seek
1	1	kha-seem
1	1	khe khe
1	1	khe-seek
1	1	kwick
4	1	meak
5	1	milk
1	1	peak
6	1	saik
3	2	see
1	1	see her
76	11	seek
4	3	seeker
1	1	seeyek
15	3	sick
2	1	silk
8	1	sink
2	1	sneaker
3	1	soap-kha
4	1	thin
8	1	think
1	1	zeeg
3	1	zeek
1	1	zyeg

		REFERENCE
R	L	peak
3	1	a kick
1	1	a peak
1	1	ateyo
6	2	beak
1	1	big
1	1	blake
2	1	bleak
2	1	blee
1	1	come cute
3	1	compete
2	1	complete
1	1	compute
8	1	cookie
1	1	could be you
1	1	dizzy
1	1	drink
3	1	duke
1	1	d-zeek
1	1	ee
2	1	eek
3	1	eek beak
3	1	flake
2	1	flee
1	1	fleet
2	1	gleek
2	1	hate
4	1	he
1	1	heat
1	1	heek
1	1	hey
1	1	ick
1	1	ink
3	1	jake
2	1	kah-pea
2	1	keekee
1	1	kha-peak
1	1	khapee
1	1	kha-pee-yek
16	2	kick
1	1	leak
57	10	peak
1	1	peaker
2	1	phyk
1	1	pick
1	1	p-leek
2	1	ribbit [frog sound]
2	1	silk
3	1	teeck
5	1	tick

		EDITED
R	L	peak
3	1	ache
1	1	appeal
6	1	bake
11	5	beak
1	1	beep
5	2	big
1	1	big beak
1	1	bin
5	2	bleak
1	1	bleep
13	1	cheek
1	1	cleek
3	1	click
4	1	commute
4	1	compete
1	1	compute
14	2	cookie
5	1	could be
1	1	could be you
1	1	dick
1	1	din
4	2	eek
1	1	eelk
1	1	fleak
1	1	gleek
1	1	he
2	2	heak
3	1	ink
3	1	jake
1	1	keew
4	2	kha peak
2	1	kha peat
1	1	kha-pea
3	1	kick
5	2	leak
1	1	lee
1	1	monique
54	11	peak
8	2	pick
1	1	pill
3	1	pink
3	1	puke
2	1	quickly
1	1	think
1	1	tick
1	1	till

		REFERENCE
R	**L**	**thought**
1	1	amount
2	1	bed
6	1	default
1	1	defoe
2	1	de-rot
1	1	doll
2	1	dolled
1	1	don't
14	7	fault
2	1	faulter
3	1	felt
12	3	filter
13	1	float
6	4	fold
4	1	folder
4	1	go out
3	2	goat
1	1	gold
1	1	i thought
1	1	into-o
1	1	itel
1	1	it's old
1	1	kesso
11	1	salt
4	1	salts
1	1	salty
4	1	saltzer
9	2	sold
2	1	sold her
1	1	solter
7	1	teso
2	1	te-thou
5	1	though
61	10	thought
4	2	thout
1	1	throat
4	1	throt
1	1	throter
1	1	throw
1	1	thud
3	1	tiffle
2	1	to doll
3	2	to fold
1	1	to thou
1	1	told
9	1	volt
7	1	vote
1	1	wrote

		EDITED
R	**L**	**thought**
1	1	a night
1	1	abort
2	1	beep
1	1	bold
1	1	defaul
1	1	default
1	1	defaulter
1	1	different
1	1	faught
40	9	fault
2	1	faulter
1	1	faults
1	1	feeling
6	1	felt
5	1	fill-lod
2	2	filter
6	2	float
2	1	flood
1	1	flot
1	1	fod
7	5	fold
1	1	fold it
6	2	folder
1	1	fooled
10	2	fort
1	1	forth
2	1	fout
1	1	i thought
2	1	i won't
1	1	ignite
1	1	low
1	1	note
1	1	old
1	1	paint
1	1	plate
6	2	salt
1	1	salt fault
1	1	salts
1	1	salty
1	1	some more
3	2	sought
11	1	te-thyl
2	1	the lot
1	1	thermometer
1	1	though
38	8	thought
1	1	thoughts
1	1	thoughts fold
7	2	throat
3	1	throt
2	1	throw
7	1	thwart
5	1	tiffle
3	2	to fold
2	1	tonight
9	1	volt
7	2	vote
1	1	walt
4	1	wolt

		REFERENCE
R	**L**	**caught**
2	1	acont
4	1	beep
4	1	boat
1	1	bonk
1	1	called
30	10	caught
5	2	coal
10	2	coat
1	1	cocked
10	4	cold
1	1	come
5	2	come on
2	2	cook
5	1	corked
1	1	could
8	4	court
1	1	cult
2	1	cup
6	3	cut
1	1	decol
1	1	deecor
1	1	fault
1	1	float
2	1	got
2	1	hall
2	1	hauk
3	1	hawk
2	1	hoh
1	1	hold
3	1	honk
17	6	hot
1	1	hot cup
1	1	hou
2	1	hut
6	1	kaut
1	1	kho
1	1	klaut
1	1	o
3	2	talk
1	1	to kho
1	1	too hot

		EDITED
R	**L**	**caught**
5	1	beep
8	1	boat
2	1	bolt
4	4	call
1	1	called
37	10	caught
1	1	caught cut
2	1	clote
1	1	coach
8	3	coat
3	2	come
2	1	come along
1	1	come on
1	1	cop
7	4	court
4	1	cup
4	2	cut
7	1	goat
2	2	got
7	1	hall
6	2	hawk
1	1	holt
2	1	honk
2	1	hor
1	1	horn
3	1	hort
7	4	hot
1	1	hou
1	1	hut
1	1	kho
1	1	pop
1	1	put
8	1	quote
5	3	talk
1	1	teecall
1	1	too hot

		REFERENCE
R	L	torch
1	1	bloach
1	1	co watch
19	4	coach
2	2	coat
1	1	could
1	1	dhu-or
2	1	dorch
10	1	echo
1	1	forgot
1	1	garch
2	1	gloach
1	1	gloat
1	1	go
1	1	go on
2	1	go watch
4	2	goal
23	3	goat
1	1	goats
1	1	god
2	1	goer
3	1	going
1	1	good
2	2	gorch
1	1	gort
1	1	got
1	1	gotch
1	1	gotta watch
5	3	scorch
2	1	she talked
2	1	talk
1	1	talked
1	1	talking
19	4	taught
1	1	t-hor
2	1	thought
2	1	t-o
1	1	to watch
1	1	told
2	1	tor
45	10	torch
1	1	torn
2	1	touch
1	1	triple
2	1	wacko
4	1	watch
1	1	what
1	1	whicheeta

		EDITED
R	L	torch
1	1	a watch
3	1	beep
9	2	coach
6	2	door
1	1	dorch
1	1	dou
1	1	doyle
1	1	gain
2	1	gal
1	1	gauge
1	1	gloach
4	1	gloat
2	1	glow
1	1	go
2	1	go on
1	1	goach
3	1	goal
12	2	goat
3	1	god
3	1	goer
4	1	golach
4	1	good
6	3	gorch
1	1	great
1	1	moocha
3	1	oocha
1	1	orch
3	1	ouch
1	1	poach
6	2	scorch
3	1	scotch
1	1	sport
1	1	talk
2	1	tall
6	2	taught
2	1	throat
1	1	to let
2	2	to watch
1	1	toat
1	1	tod
4	1	toe
1	1	toet
1	1	toil
1	1	told
1	1	too much
8	2	tor
39	11	torch
1	1	torch coach
2	1	torch gorch
1	1	torwatch
1	1	touch
4	1	tow-e
2	1	what

		REFERENCE
R	L	porch
1	1	beep
1	1	boat
3	1	bought
1	1	bro
1	1	brought
1	1	call
1	1	caught
1	1	cheaper
4	1	cheepou
1	1	cheerful
18	2	coach
1	1	default
1	1	deport
1	1	de-port
4	1	float
1	1	for
2	1	hull
1	1	or
2	1	paid
2	1	pain
1	1	paint
4	2	paul
2	1	peel
2	1	ping
2	1	pink
1	1	po
24	3	poach
1	1	poet
1	1	pooch
2	1	pool
2	1	poor
4	1	pope
26	10	porch
2	1	pore
3	1	pork
13	5	port
1	1	pou
1	1	pouch
15	7	pull
1	1	pull it
4	1	she porked
1	1	to pull
1	1	triple
1	1	would ya
2	1	would you pull

		EDITED
R	L	porch
1	1	beep
2	1	boach
3	1	boat
11	1	cheeple
17	2	coach
2	1	deport
1	1	filling
4	1	float
4	1	flow
2	1	hawk
1	1	hull
1	1	paint
2	1	paul
1	1	paw
1	1	pill
3	1	pillow
3	2	poach
1	1	po-at
3	1	poe
2	1	poet
1	1	point
1	1	pool
1	1	poor
1	1	por
54	10	porch
1	1	porch pull
3	3	pork
13	4	port
4	1	pouch
16	6	pull
1	1	pull porch
1	1	pull up
1	1	pully
2	1	put
3	1	she porked
1	1	tiffle
2	1	triple
2	1	would you pull

Appendix 4 – *Raw responses from the pitch contour study. Condition 1 (FF, all falling) is where each repetition of the word token in a 3-min sequence followed a pitch contour from high to low, Condition 2 (RR, all rising) is where each token in a 3-min sequence followed a pitch contour from low to high, and Condition 3 (RF, alternating) is where the pitch contours of successive tokens in a 3-min sequence alternated between rising and falling. 'R' is a total number of responses for a given form and 'L' is a total number of listeners who reported that particular form.*

Condition 1 FF

first response included

R	vows	L
20	avowers	1
	balvs	1
	barrels	1
	bau oz	1
11	bawls / bauls	6
5	baws / bows	2
5	bell	3
9	bells	5
	belt	1
	blouse	1
2	bounce	2
3	bow	3
	bow thau vows bauz	1
2	bowers / bawers	2
	bow-policy	1
2	clothes	1
	dau oz vows	1
5	dau/daw	3
3	daws	2
	dell	1
	down	1
	down smells	1
	eefozeefoz	1
	eewovs	1
	fause	1
3	fell	2
5	fells	2
5	flow	1
	flower	1
39	flowers	10
	flows	1
	followers	1
	fouzee	1
	fouzee bau	1
	he vows	1
23	isabel	2
4	it's a bell	1
	it's ours	1
	kous	1
	nahoos	1
	nahoosenah	1
2	now whos	1
	of ours	1
	oo-sa	1
	oo-sells	1
	oo-smell	1
6	ours	4
	ouths	1
4	o-vowers	1
	policy	1
	powers	1

Condition 2 RR

first response included

R	vows	L
4	barrels	2
	bau zee was	1
	bauz	1
	bauz vows	1
5	baw	1
3	baws / baus	2
7	bell	1
11	bells	4
8	belooskee	1
	bow	1
3	bowels	1
	colors	1
	cows	1
	daus	1
2	dow / dau	2
	falls	1
2	fauls	2
3	fell	1
2	firewalls	2
3	float	1
5	flow	1
30	flowers	6
9	isabel	1
5	it's a bell	1
	it's a vow	1
	kee was vowels vow	1
3	mouse	2
45	of ours	4
5	oozenah	1
	oozk	1
6	ours	2
4	o-vowers	1
	plawers	1
	sebawah	1
	seh-vows	1
	smells	1
2	sparrows	1
	splowel	1
3	spouse	1
	svawes	1
1	swell	1
	that colors	1
5	those	3
22	thou	6
	thou bau	1
	thou bow	1
	thou dow	1
	thou ee zee was	1
	thou fouls kee was	1
	thou kee was	1
	thou thauers	1

Condition 3 RF

first response included

R	vows	L
5	awawers	2
	barrel	1
3	bats	2
	bau vow thou vows	1
2	bawers	1
3	bell	1
7	bells	5
3	belooskee	1
	booskee	1
	bow	1
	bowels	1
	bows	1
	clauthes	1
6	clothes	3
	colders and nose	1
	colers knows	1
	cose nose	1
3	dau	1
	firewalls	1
2	flow	1
2	flower	1
33	flowers	11
	foul smells	1
3	fouled	2
2	fouls	2
	goes goes	1
	goes nose	1
2	hours and hours	1
16	isabel	1
12	it's a bell	1
	it's a val	1
15	it's a vowel	2
	it's a werewolf	1
	it's of our vows	1
	it's vowels	1
	kee was kee buzz thou vows	1
	kee was kee buzz vow thou	1
	kee was kee buzz vowers	1
	kee was thauers vows	1
	kee was vows	1
	lots of hours	1
24	of ours	2
	quotes	1
	sevel	1
	sit down	1
	smell	1
8	smells	6
2	spell	1
	spell smell	1
2	spells	1
6	spouse	2

R	vows	L
	say vowel	1
	scales	1
4	scattle	2
2	scettle	1
	seh vowel	1
2	smell	1
9	smells	5
	smoohoo	1
	snails	1
	some owls	1
	spall	1
	sparrows	1
	spau	1
18	spell	5
3	spells	1
	spiral	1
	splowers	1
14	spouse	3
3	spowels	1
5	spowers	3
	stall	1
4	stalls	2
	stator	1
2	staus	2
	svels	1
	teh-fowel	1
2	the vowel	1
7	those	2
13	though	2
	toes	1
	towels	1
7	val	3
14	vals	5
2	valve	2
12	valves / valvs	4
	va-oot	1
2	vaos	1
3	vels	2
4	vo-u-ws / va-ooz	2
8	vow	6
	vow bau	1
	vow bau thau eefoz	1
	vow oz bau z	1
	vow thau	1
	vow thaus	1
4	vow z	2
3	vowel	2
	vowel he knows	1
56	vowels	11
18	vowers	11
	vow-ooz	1
115	vows	23
	vows he knows	1
	vows nose	1
	vows us	1
	vows z	1
4	vow-us / vow oz	2
2	zbower	1
	zbowers	1

R	wave	L
	always	1
	ave	1
	beep	1
	bees	1
3	cheese	2
2	clave	2
2	eef	1
	foo-e-ee	1
2	foo-way-ee	1
	fooweya	1
	fooweyou	1
2	for me	1
	glave	1
	glaze	1
2	heeth	1
	if way	1
2	leaf	1
18	leave	1
2	oyee / oee	1
5	peace	4
	people	1
	play	1

R	vows	L
	thou ugh	1
4	though	1
	thou-oz	1
6	thou-s / thous	53
	thous vows kee was	1
2	twos	1
	us us us	1
	vah-wah	1
5	val	2
9	vals	4
3	valve	2
10	valves	4
2	va-ool	1
	vel	1
7	vels	1
	vil	1
9	vow	5
	vow kooz	1
	vow vow-ah	1
	vow-ah	1
4	vowel	1
76	vowels	15
	vowels thauls	1
27	vowers	9
3	vow-oz	1
142	vows	24
	vows is	1
	vows kee was	1
4	vows those	1
	vows vows those ours ours	1
	vow-zee-woz	1
	who's a bell	1
3	wow	1
6	z-vah-wah	1
2	zvowers	1

R	wave	L
	ave	1
	away	1
4	awayes	2
	aways	1
	bave	1
	bees	1
2	beyeth	1
2	clog	1
	disc	1
	ethor	1
6	eve	2
	heap	1
	heave	1
2	heeth	1
	keith	1
8	king	5
11	leave	2
	my ear phone	1
2	play	1
6	please	1
	police	1
	pum pum pum pum	1

R	vows	L
	spowers	1
	stall-val-stalls	1
	swell	1
10	tha-oos / thaus / thous	5
	thau us vowels	1
2	thauers vows	1
6	those	3
4	those flowers	2
14	though	3
	though bau kee was	1
5	though bau vows kee was kee buzz	2
	though vow	1
	though vows	1
	va va vows	1
	val smells	1
	valors	1
23	vals	8
3	val-spells	1
	val-stalls	1
	valve	1
15	valves	4
	vase	1
2	veils	1
	vels smells	1
10	vow	4
	vow kee buzz vows	1
	vow zee was	1
	vow zee was vows	1
3	vowel	2
	vowel bowel ask us	1
450	vowels	13
	vowels smells	1
	vowels vows	1
39	vowers	11
4	vowers vows	1
123	vows	23
	vows bows	1
	vows bows ask us	1
	vows fouls	1
	werewolf	1
4	who's in the	1
3	wow	2

R	wave	L
6	are u useful	1
	either way	1
8	for me	1
4	give	2
	heave	1
6	if	3
3	if way	1
	kwey kweev	1
2	leaf	1
7	leave	1
	leave please	1
3	play	1
6	plea	3
13	please	1
	please wave	1
	rate	1
	sway	1
	useful	1
	wah	1
6	wah wah	1
	wait	1
	waste	1

R	wave	L
4	plea	1
7	please	2
2	wait	2
132	wave	24
	wave beep	1
	wave brief	1
4	wave wave	2
	wave weyef	1
2	way	1
	way eve wave	1
24	way if / way f	7
3	way if wave	1
	way if wave weave	1
	way lyf	1
	way puh buh	1
	way way	1
3	ways	3
23	weave	9
2	weave wave	1
	weeth	1
	we-eyv	1
3	weyes	1
4	weyev / weyef / weye-f	3
7	weyf	6
	wha if	1
2	wha	1
11	what if	1
2	where	1
11	where are u	1
17	where u	2
18	where you from	4
	where you-f	1
	wife	1
2	with	1

R	maze	L
	a mate	1
	ace	1
15	amaze / a-maze	9
	ameyers	1
2	baby	1
5	bait	1
39	base	12
	base mace	1
	base-in	1
	bathe	1
	bathed	1
8	bay / bey	5
	beat	1
2	bees	2
4	beiz / baze	3
3	betties	3
5	bite	2
	can u see now	1
	clay	1
	eiz	1
	formees	1
2	glaze	1
4	haze	2
	is	1
	it's a maze	1
5	ladies	3
	leaf	1
	leaves	1
	m	1
10	mace / meis	5
	mace base	1
2	make	2
	make me	1
5	make space	2
	mason	1
3	mate	2

R	wave	L
3	rave	2
	squeze	1
	sway	1
9	th-way-you	1
5	useful	3
	value	1
6	wait	2
157	wave	24
	wave way yfa	1
	wave weave	1
	wave weave where youfa	1
	wave what	1
2	wave what if	1
13	way	5
	way eefa wave	1
37	way if	7
2	way if wave	1
	way yfa	1
2	way-eph	1
16	weyf	7
	wayne	1
13	ways	5
	way-you	1
	way-youtn	1
4	we	2
6	weave	5
3	weave wave way if	1
	weave wave where you from	1
	weave ways	1
	weef	1
	we're youthful	1
8	weyef	4
15	weyes	2
4	wha	1
3	what if	1
4	wheat	1
9	where	4
4	where are you	1
	where are you from	1
16	where you	1
2	where you f	2
23	where you from	4
	where youfa	1
	why u	1
	youthful	1

R	maze	L
4	ace	1
60	amaze	12
	amaze amaze somey somey	1
2	amaze amaze someys someys	1
	amaze smaze maze symbaee	1
11	bait / bate	5
2	bake	1
	bake some eggs	1
35	base	9
	base-in	1
	bates	1
	bath	1
3	bathe	2
5	bay	4
2	bay is	1
	bay ts symbaee mace	1
	bees	1
4	beiz / baze	3
	benny	1
7	bite	1
	day is	1
4	eat some	1
	eembee symbaee maze	1
	eggs	1
	endays	1
	fomeh	1
	fomey	1
3	lay	3
16	mace / meis / mais	6
	made	1
2	make	1
3	mate	3
2	mates	2
14	may	6
	may bay ytsa	1
3	may is	2

R	wave	L
134	wave	24
	wave away	1
	wave please	1
	wave wave puh bhup	1
	wave wayef puh bhup	1
9	wave weave	4
3	wave weef	2
14	way	5
23	way if	2
2	way wave	1
	ways	1
	we	1
2	weak	1
6	weave	3
3	weave wave	2
	were if	1
23	were u	2
2	we're useful	1
	weyd	1
12	weyef	2
9	weyef wave	1
	weyef weave wave	1
8	weyf	3
4	what if	1
9	where	3
5	where are you	1
	where from	1
14	where you from	2
	where you've been	1
	width	1
	ygif wave	1

R	maze	L
4	ace	1
	amah	1
19	amaze	12
	amaze maze	1
2	ameh	1
	aze	1
2	bait	1
4	bake	2
13	base	7
2	base-in	1
4	bay	1
	bike	1
4	bite	2
4	bites	2
2	buys	1
3	ey-may	1
5	hemeyes	1
	it's for me	1
	kehmeyez	1
	kemeyes	1
3	keneyes	1
13	mace / meis	5
7	mace maze	2
	made	1
2	make	2
8	make me	2
12	make space	1
3	mate	2
2	mates	1
22	may	7
	may is	1
	may space	1
	may zee	1
5	maybe	1
3	may-space	1
108	maze	22

R	maze	L
4	mates	2
11	may	5
2	may is	1
	may z	1
3	maybe	1
	maybeeza	1
145	maze	22
	maze veyz	1
	me me	1
	mease	1
4	meyers	3
	mm maze	1
6	phase	1
5	plate	1
6	play	1
6	plays	3
	plea	1
2	please	2
4	smate	2
3	smates	2
	smeym	1
	sombaee	1
	some base	1
	some bees	1
2	somebaee maze	1
2	somebee	1
6	somebody	2
	somebody smaze maze	1
	somey please	1
3	somey police	1
3	soonnee	1
17	space	6
	space mace	1
	spees	1
2	spey	1
	spey may	1
4	sue me / sumee	4
8	sumey	5
3	sumey somebody	2
34	sumeys / some a's (ace)	7
6	symbaee	1
2	to me	2
	to meh	1
2	to mey	1
11	to meyez	1
3	to-maze	1
2	tomee / tummy	2
6	vase	2
	zmaze	1

R	maze	L
	may its	1
3	may-it	1
124	maze	23
	mazey	1
2	meets	1
6	meh	1
37	meyers	3
15	naze	1
	peace	1
22	pehneys	1
2	plate	1
7	play	1
	plaze	1
6	please	1
	same	1
	smate	1
	smates	1
	smey	1
	some ways	1
5	somebody	3
	somey please	1
	somey police	1
	somey someys	1
3	space	2
3	spey	1
	sumate	1
2	sumbee	1
9	sumey / somey	4
	sumeyou	1
15	sumeys / someys	5
	sundee	1
3	surveyv	2
2	symbaee	1
4	symbaee(y) mace	1
	symbaees	1
	symbe-ee	1
	symbey mace	1
2	tammy	1
13	they	1
	they eat	1
	to me	1
	vaz	1
7	veyz	5

R	maze	L
	maze symbaee	1
	maze wait	1
11	me	3
	meesome mace maze	1
	meesome maze	1
3	meet	1
5	meets	2
3	meez	1
3	meh	1
	meh-is	1
15	meyers	6
3	mey-soomeys	1
2	mey-space	1
7	naze	1
	need some air	1
2	play	1
6	please	1
3	smeeze	1
	smeyz	1
27	some-ace / sumeys	8
2	some-bees	2
2	some-ees	1
	someys wait	1
	space	1
26	sumey / some-ey	8
	sumeyt	1
7	symbaee	1
4	symbaee mace maze	1
	symbaee mace maze meesome	1
	symbaee maze	1
5	tehmeyes	1
3	tehneyes	1
	temeyes	1
	they	1
2	they play	1
6	to me	4
	trays	1
	values	1
	vase	1
	veyz	1
	who's gonna	1

R	nose	L
	analys [nw]	1
	and i was	1
8	and ours	1
3	colors	1
18	covers	4
3	cut nose	2
	dawers	1
	don't know	1
	dose	1
	eewas nose	1
16	elvis	5
	elviska	1
2	gaut	1
	gawers	1
3	gnome	1
4	goes	2
3	he knows	2
18	hours / ours	6
	in on this	1
4	is it now	3
	keewas	1
	keewas no	1
2	keewas nose	1
38	knowers / nowers	12
	know-ss	1
	na-nas	1
42	no	10
	no coarse	1
2	no keewas nose	1
3	no ooz	1
	no risk	1
4	no-kuz	1

R	nose	L
	colors	1
	colors no	1
	colors no colors	1
5	cose	3
11	covers	3
	damn	1
2	do it	1
	do them	1
	dogs	1
	dome	1
	donk	1
5	don't	4
2	don't know	2
	dose	1
	dumb	1
	dumb coarse gnome coarse	1
	dumb dumb coarse nose	1
	ella	1
3	elves	1
8	elvis	1
	gaun	1
7	gnome	5
2	gnome coarse	1
2	gnome geboos gnome keboos	1
	gnomes coarse	1
4	goes	3
	he know us	1
2	he knows	2
9	hours	3
	kee was	1
	kee was nung	1
	k-neh	1

R	nose	L
5	an hours	2
3	and hours	1
	and i was	1
	damn	1
7	don't	5
18	don't know	6
3	dose	2
	dose nose	1
	elvers	1
7	elves	2
13	elvis	1
	en-elves	1
	goers goes	1
3	goes	1
	he knows	1
	kee was no-ung nose	1
	know	1
5	know it	2
	know ya	1
35	knowers / nowers	12
	lowers knows	1
	naut	1
	nayon knows	1
	neeyung nose kee was	1
8	neh neh	1
42	no	7
	no coarse	1
	no goes	1
	no kee was nung nose	1
2	no no	1
2	no no z	1
	no nose	1

R	nose	L
5	nolan	2
	noles	1
	no-nas	1
3	noone	1
5	no-oz	1
3	nos	1
9	no-s	1
180	nose	24
	nose elvis	1
	nose keewas yo nose	1
	nose myh	1
	nose nas	1
	now is it	1
2	now us	2
	nowers myh	1
	o	1
8	of ours	3
	oo-nose	1
2	shall we	1
7	she knows	1
	snose	1
8	snow	2
	snows	1
	tell us	1
3	that was	1
	that wasn't	1
8	those	5
2	toes	2
	towers	1
	ugh! keewas	1
2	ve-knows	1
	visk [p: wysk]	1
	viska	1
	visker	1
	vy-know [pol: wy]	1
	when i was	1
4	who knows	3
	windows	1
	yeh keewas nose	1
	y-know-s	1

R	lathe	L
6	aids	3
2	alathe	1
3	allow u	2
2	amaze	1
5	and ladies	1
2	and layers	1
3	and loads	1
3	ave	1
	ave lathe	1
	ave pave live lathe	1
	base	1
4	bathe	2
2	beans	3
4	bees	3
	believe	1
	below you	1
2	blade	1
	blaves	1
4	blaze	4
	blaze slaves	1
	can't use	1
	can't use that	1
	clades	1
	clays	1
2	dave	1
8	days	5
	ee-layeth	1

R	nose	L
3	know is	1
	know it	1
	know us	1
51	knowers / nowers	13
	know-e-snows	1
3	known	2
	known isk	1
	known to poos known canoos	1
2	neh	1
	neves	1
44	no	6
	no at risk	1
8	no colors	2
	no isk	1
2	no kee was	1
	no kee was nose	1
3	no nose	2
	no nose no	1
	no nowers	1
2	no one knows	1
	no skewers	1
2	no us	1
	noaz	1
	noen	1
4	noise	3
	nom	1
	none kee was	1
	nong dong kee was	1
4	no-oos	1
	noose noose	1
3	nos	2
156	nose	24
	nose coarse	1
2	noser	1
12	nosey	5
	no-ss	1
	now	1
16	now is	3
	nowers kee was	1
2	nowers nose	2
2	no-z	1
13	noze	1
	no-zs	1
4	nung	2
4	nung kee was	1
	nung no kee was	1
	skelers	1
	skelter	1
	skewers	1
8	those	4
	whiskey nung	1
	who knows	1
4	windows	2
	you know us	1

R	lathe	L
	aids	1
4	a-lathe	1
	alive	1
5	amaze	3
	and you	1
	are they yours	1
2	ave	1
	behave	1
	blaze	1
	commute	1
3	dave	3
	delayed pace	1
	eight	1
4	eighth	3
11	e-leyoo	1
4	eve	2
	ey	1
	eyef	1
2	e-you	1
2	fee-ley	1
	fee-ley-you	1
4	fee-ley-youth	1
	glades	1
	havee-as	1
	he lays	1
	heegth	1
	i'm late	1

R	nose	L
	noms	1
3	noone knows	2
	noone knows kee was	1
	noones	1
2	noose	2
	nos	1
157	nose	24
2	nose coarse	1
	nose kee was	1
	nose nose	1
	nose nung	1
	no-ung nose	1
	now	1
2	now it snows	1
	nowa nose	1
	nower knows	1
	nowes knows	1
19	noze	4
	nung knows kee was	1
	nung nose	1
2	nung nose kee was	1
7	of ours	3
7	ours	4
	out of hours	1
	she does	1
2	she don't	1
4	she goes	2
	she knows	1
	she-done	1
4	snow	2
2	toes	2
	vals	1
	when i was	1

R	lathe	L
	a is closed	1
5	alathe	1
3	allow u	3
	and layers	1
	and loads	1
14	ave	2
	below you	1
4	blade	1
	blades	1
	blave lathe iskus	1
	blaze glave	1
	bleyeth blave	1
2	days	3
	dee yskus diskus	1
3	eelay-u	1
	eleyets	1
2	en-ladies	1
	explain	1
2	flav	2
	he laze	1
3	huh huh	1
	i like u	1
	in lanes	1
2	it's below you	1
6	keh-disk	1
	keh-liss	1
	kiss	1

R	lathe	L
	ee-layoo	1
	eelays	1
2	elayoo	1
	escalay	1
3	escalay-you	1
	escalay-youth	1
5	eve	4
	explain	1
	eyers is kah	1
6	feeds	3
	fee-lace	1
	fee-lay	1
3	fee-lay-yous	1
6	feyv	1
	galyeth	1
5	glades	2
	gladies	1
7	glathes / glaves	6
	haties	1
	hayers	1
	he lays	1
6	if	4
	if tle eve	1
	i'm laze	1
3	is	2
	is amaze	1
	is bah laze	1
	is kah lathe	1
	is kah play	1
3	keneyes	1
	keys	1
	kiff	1
2	kiz	1
4	la	1
4	lace	3
6	ladies	2
	laid ace	1
	laid aids	1
5	late	3
83	lathe	20
	lathe leave	1
	lathe-kids	1
8	lathes	3
11	lay	6
3	lay is / la ees	3
	lay you	1
6	layer	3
2	layered ace	1
	layered ace laze ace	1
53	layers	11
	layeth	1
	lay-kids	1
	lay-kiss	1
	lay-spee-lay-you	1
	lay-u-kiz	1
	lay-you-t	1
	lay-youth	1
4	laze	2
	laze aze	1
4	lazy	3
15	leave	4
	leave lathe blave	1
7	leaves	4
	les-kee-lay-o-you	1
2	less	1
3	let u	1
2	leyeth	2
3	leyoo	1
10	leys / laze	6
	leyv	1
10	like u	1
	lines	1
5	loads	2
	loaf	1
	l-you	1
	mates	1
3	mave	2
3	may have	1
	mayers	1
7	maze	5
	maze lathe	1
	na	1
7	peas	1
2	peneyes	1
	peyv	1

R	lathe	L
	influen	1
	it's below you	1
7	kiss	1
13	ladies	6
	ladies ley if	1
5	laid	1
	laid ace	1
	laid pace	1
5	lanes	3
5	late	1
73	lathe	21
6	lathe e-leyoo	1
	lathe flave	1
	lathe ladies	1
	lathe lath yv	1
	lathe yfa	1
	lathes	1
26	lay	6
	lay if discus	1
	lay in	1
14	lay is / ley is	2
	lay it	1
	lay you	1
	lay you discus	1
	lay-a ts	1
9	layer	5
	layer discus	1
	layer lay if layer lathe	1
	layered delay	1
72	layers / leyez	16
	layers leys	1
	layers pace	1
	layers pace lays pace	1
48	laze / lays	9
	lazy	1
3	leads	2
8	leaf	3
15	leave	6
	leaves	1
	l-eaves	1
	lee	1
	leeh	1
8	let you	1
4	leyds	2
5	ley-eeth / ley eev	3
8	leyeth / ley-eff	6
5	ley-if	2
3	leyoo	1
6	leyv	1
	leyv lathe	1
5	ley-youth	3
	lie you	1
5	life	1
9	like you	1
4	live	1
3	loads	3
	lothes	1
	may have	1
2	may-it	1
	maze	1
3	move	2
5	naive	2
3	please	4
	pleeth	1
10	slave	2
	spee-lay-youth	1
5	value	1
	veyth	1
7	yoo-se-lah	1
	yoo-ste-lah	1
	yoo-stu-lah	1
	you	1

R	lathe	L
14	lace / laze	4
9	ladies / lay dees	4
	laeyeth blaze	1
5	late	2
	later	1
70	lathe	21
8	lathe eelay-u	1
	lathe it is kus	1
2	lathes	2
13	lay	6
7	lay is	1
	lay is lathe	1
	lay loose	1
	lay you	1
	layds	1
6	layer	3
	layer clave	1
	layered ace	1
	layered pace	1
	layered peace layered pace	1
60	layers	15
3	layers and layers	1
	layet	1
	laze blaze	1
9	lazy	2
4	leads	1
	leaf	1
31	leave	8
2	leave please	2
19	leaves	7
	lee-aids	1
	lee-ers	1
5	leh	1
2	leh leh	1
3	length	1
14	let u	3
	leyd	1
	leyeth	1
3	leyeth blaze	1
	leyeth clave	1
4	leyeth lathe	1
	leyeth lathe blaze	1
2	leyeth lay yfka	1
3	leyeth laze	1
	leyeth laze yfka	1
	leyeth please	1
	leyop ye laze	1
	leyouth clave	1
	leyskus lathe	1
17	leyz / laze	5
	lez	1
4	liar	3
	lie u	1
	life	1
	lights	1
16	like u	3
	linear	1
	loads	1
	look-a-lay-loo	1
	mave	1
4	may have	2
3	pizza	2
3	please	2
3	slave	2
	slayer	1
8	they	3
5	thou	1
	trees	1
14	value	4
2	were u	1
9	who's gonna	1
	yfka	1
	yoostella	1
2	yoo-ze-lah	1
6	you-ke-lay-loo	1
4	you-still-there	1

R	lathe	L
4	phase	3
	piers	1
2	play	2
	play iv	1
2	player	1
	plays	1
	playv is bah lathe	1
18	please	6
	pleav	1
	save	1
	skee-lay-yous	1
14	slave	5
2	slaves	2
11	some-eys	2
	spee-lace	1
	spee-lay	1
2	spee-lay-you	1
	spee-lay-yous	1
	splay-you	1
5	they use	1
	they used to	1
	use to the	1
13	value	4
2	veyv	1
	whos gonna	1
	whos in the	1

R	writhe	L
9	arrive	5
	berolliance	1
2	blith	1
	blithe what is	1
	bra is is bra brais	1
11	brian	6
	brian's	1
5	brieth	1
3	brilliant	2
	brine	1
	brise	1
11	briyez / bryez / brayes	4
4	brize / bry is	3
4	brollies	1
	brollies lies	1
	bry	1
2	bry if	1
	bry is is bra	1
	bryth	1
	didn't u	1
5	draw	3
3	drawing	1
15	drive	5
	dryes	1
2	drys	1
4	eyes	2
	fire	1
	five	1
4	flyers	1
	fry-es	1
9	go in	2
17	going	3
	goliath	1
3	grith	1
	higher	1
8	if	5
	is bra	1
	is bry is fry bryth rise	1
	is funny	1
	is rah	1
2	is rye	2
6	life	3
	life writhe	1
2	night	2
	poa hith	1
	poa-yez	1
	prian	1
7	prize	4
	prolliez	1
2	rah-his	2
25	rah-hiv	1
	rah-if	1
5	rah-is	1
16	rah-you	1
	raw-eez	1

R	lathe	L

R	writhe	L
4	arise	3
	arriai	1
6	arrive	4
3	bike	2
	blith	1
	bliyeth	1
	bollies	1
7	brian	3
11	brieth	1
	brive	1
	bro expo	1
2	bro if	1
	bro if pry if writhe rides yfa	1
	bro if writhe	1
	bro if yfa	1
	bro is bro if writhe rise	1
	brohith	1
3	brollies	1
	bry	1
	brythe bra	1
5	cliff	4
	clive	1
12	drive	5
	expa	1
3	flies	2
2	flyers	2
	flying spry	1
	glai-eth	1
8	going	1
5	grith	1
	heef	1
	hees	1
	heeth	1
	hill	1
	his	1
	if	1
	if kah	1
	ipska bro if yfsa	1
3	is rye	1
14	knife	5
10	life	4
2	live	2
	ly-eth	1
	nice	1
	no ifs	1
	ny-eth	1
	pirollies	1
4	please	3
	plys	1
	plyz [p: plyz]	1
	prollies	1
	pry if	1
2	rah	1
	rah is	1
	rah-eev	1

R	lathe	L

R	writhe	L
6	arrive	4
3	bias	2
	big ears	1
2	blind	2
2	blithe	1
	bra exprise	1
	bra if writhe	1
	brayent	1
19	brayes / bryez	2
8	brian	2
	brithe writhe	1
	brix rise	1
	bro if brithe	1
	bro if rise	1
	bro is bra is	1
7	brollies	2
	bry ex prise	1
	bry ex rise	1
	bry if	1
	bry if buy	1
	bry if right	1
3	bry if rise	1
2	bry its rise	1
	bryeth rise	1
	buy it writhe	1
6	drawing	3
17	drive	5
2	dry	1
3	flies	3
4	flyers	1
	fry	1
	glieth	1
	glithe	1
6	go	2
7	going	2
	grive	1
2	his	2
5	if	1
	ifrow	1
4	is rye	1
2	kite	2
7	life	5
2	live	1
	noise	1
	perollies	1
	poa-yez	1
	police	1
3	rah	1
	rah and rise	1
	rah in rise	1
2	rah is	1
	rah its rise	1
2	rah-hid rye	1
26	rahiv	3
	rahry flies	1

R	writhe	L
3	raw-is	1
14	rayeth	5
	rayth	1
	ray-you	1
17	rice	6
3	rides	2
2	rife	2
5	right	2
	rights	1
6	riot	2
	riots	1
62	rise	10
	rise writhe	1
	row is	1
6	ryan	4
9	rye	4
11	rye if	3
	rye if bry if	1
	rye is	1
	rye-eh	1
	rye-yh	1
2	sparoli	1
	sparoliz	1
2	sprize	1
	sprolli	1
2	tries	2
7	value	1
8	values	3
	wave	1
	wayef tss	1
	wayef what if	1
	wayef what is	1
	weyef	1
18	where are you	4
8	will u	1
16	wrieth	2
88	writhe	21
	writhe life	1
2	writhe rye if	1
5	you	2

R	writhe	L
	rahere	1
11	rah-hiv	1
7	rah-u	1
2	rah-u writhe	1
19	rahyet	1
3	rahyeth	1
2	rahyets	1
	rallee-ar	1
2	raw if	1
	raw if bro if rise writhe	1
13	ray is	2
2	ray is rise	1
2	rayez	2
3	rayth	1
3	rice	2
9	ride	5
71	rieth / wrieth / rayeth / rye-eth	13
3	rife	2
9	right	5
13	riot	5
25	rise	8
	rise ray is	1
	round here	1
	row if	1
	row is	1
	rows [rauz]	1
6	royeth	1
7	ryan	4
12	rye	4
8	rye if	4
	rye if writhe	1
	rye-ou yfa	1
3	sperollee	1
	spliff	1
3	sprollee	1
	spron	1
	thrive bro if	1
3	throw it	2
7	value	2
5	values	1
2	what if	2
9	where are you	5
	where-youth	1
4	white	2
	why if	1
2	whyef	1
	whyef what if	1
	whyefs life	1
	whyefs life i knew	1
	whyefs wife	1
	whyefs wife what if why is life	1
4	wife	2
2	will-you	1
	wires wife	1
120	writhe	22
	writhe rise	1
	writhes	1

R	writhe	L
3	rahyeth	1
4	rah-you	1
	raw if writhe	1
3	ray	2
5	rayes / ryez	4
5	rice	2
4	ride	3
9	rieth / wrieth / rye-eth	6
6	rife	2
11	right	5
	right rye	1
2	rights	2
7	riot	3
	riot rye	1
6	riots	2
39	rise	8
	rollies	1
	row is [rau iz]	1
	royeth	1
4	ryan	3
	ryan brian	1
4	rye	4
5	rye if	1
	rye if writhe	1
	rye if writhe yfka	1
2	rye is	2
3	rye rye	1
	rye writhe	1
6	sperollee	1
	sprayes	1
4	sprollee	1
	sprollies	1
	thrive	1
4	ties	2
2	vaif / vayv	2
14	value	3
7	values	2
	vayef vayv	1
	wayef live	1
2	wayef live wah	1
6	ways	3
13	where are you	6
	wife	1
4	will u	1
2	wise	2
91	writhe	20
	writhe rice	1
	writhe rise	1
	writhed	1

Appendix 5 – *ANOVA tables for Experiments 5 and 6*

EXPERIMENT 5 VTs – Three-way 2 perm. (seq.1, seq. 2) x 2 cond. (All Low(F0), All High(F0)) x 2 voice (bright, dull) ANOVA

Source	***df***	**F**	**p**	**η²**
Permutation (P)	1,11	0.10	=.76	=.01
Condition (C)	1,11	0.03	=.87	<.01
Voice (V)	1,11	1.48	=.25	=.12
P x C	1,11	4.48	=.06	=.29
P x V	1,11	1.06	=.33	=.09
C x V	1,11	0.21	=.73	=.01
P x C x V	1,11	2.61	=.13	=.19

EXPERIMENT 5 VTs – Three-way 2 perm. (seq.1, seq. 2) x 2 cond. (High(TP)Low(F0)/Low(TP)High(F0), High(TP)High(F0)/Low(TP)Low(F0)) x 2 voice (bright, dull) ANOVA

Source	***df***	**F**	**p**	**η²**
Permutation (P)	1,11	0.45	=.52	=.04
Condition (C)	1,11	0.01	=.95	<.01
Voice (V)	1,11	0.07	=.80	=.01
P x C	1,11	0.08	=.79	=.01
P x V	1,11	0.17	=.69	=.02
C x V	1,11	2.18	=.17	=.17
P x C x V	1,11	2.53	=.14	=.19

EXPERIMENT 5 VTs – Three-way 2 perm. (seq.1, seq. 2) x 2 cond. (Front(TP)Low(F0)/Back(TP)High(F0), Front(TP)High(F0)/Back(TP)Low(F0)) x 2 voice (bright, dull) ANOVA

Source	***df***	**F**	**p**	**η²**
Permutation (P)	1,11	0.16	=.70	=.01
Condition (C)	1,11	0.51	=.49	=.04
Voice (V)	1,11	0.39	=.55	=.03
P x C	1,11	0.14	=.72	=.01
P x V	1,11	0.15	=.71	=.01
C x V	1,11	0.24	=.63	=.02
P x C x V	1,11	0.90	=.36	=.08

EXPERIMENT 5 VTs – Two-way 2 permutation (seq. 1, seq. 2) x 3 condition (baseline, opposing, congruent) ANOVA

Source	***df***	**F**	**p**	**η²**
Permutation (P)	1,11	0.01	=.91	<.01
Condition (C)	2,22	1.04	=.37	=.09
P x C	2,22	0.59	=.56	=.05

EXPERIMENT 5 Forms – Three-way 2 perm. (seq.1, seq. 2) x 2 cond. (All Low(F0), All High(F0)) x 2 voice (bright, dull) ANOVA

Source	***df***	**F**	**p**	**η²**
Permutation (P)	1,11	0.20	=.67	=.02
Condition (C)	1,11	0.07	=.80	=.01
Voice (V)	1,11	0.39	=.54	=.03
P x C	1,11	0.06	=.81	=.01
P x V	1,11	0.22	=.65	=.02
C x V	1,11	0.12	=.73	=.01
P x C x V	1,11	1.37	=.27	=.11

EXPERIMENT 5 Forms – Three-way 2 perm. (seq.1, seq. 2) x 2 cond. (High(TP)Low(F0)/Low(TP)High(F0), High(TP)High(F0)/Low(TP)Low(F0)) x 2 voice (bright, dull) ANOVA

Source	***df***	**F**	**p**	**η²**
Permutation (P)	**1,11**	**6.87**	**=.02***	**=.38**
Condition (C)	1,11	0.27	=.61	=.02
Voice (V)	1,11	0.20	=.66	=.02
P x C	1,11	0.38	=.55	=.03
P x V	1,11	<.01	=.97	<.01
C x V	1,11	0.45	=.52	=.04
P x C x V	1,11	1.18	=.30	=.10

EXPERIMENT 5 Forms – Three-way 2 perm. (seq.1, seq. 2) x 2 cond. (Front(TP)Low(F0)/Back(TP)High(F0), Front(TP)High(F0)/Back(TP)Low(F0)) x 2 voice (bright, dull) ANOVA

Source	***df***	**F**	**p**	**η²**
Permutation (P)	1,11	1.93	=.19	=.15
Condition (C)	1,11	0.01	=.94	<.01
Voice (V)	1,11	0.07	=.80	=.01
P x C	1,11	1.03	=.33	=.09
P x V	1,11	0.03	=.86	<.01
C x V	1,11	0.33	=.58	=.03
P x C x V	1,11	2.60	=.14	=.19

EXPERIMENT 5 Forms – Two-way 2 permutation (seq. 1, seq. 2) x 3 condition (baseline, opposing, congruent) ANOVA

Source	***df***	**F**	**p**	**η²**
Permutation (P)	**1,11**	**7.18**	**=.02***	**=.40**
Condition (C)	2,22	1.24	=.31	=.10
P x C	2,22	1.15	=.34	=.09

EXPERIMENT 6 VTs – Three-way 2 perm. (seq.1, seq. 2) x 2 cond. (All Left, All Right) x 2 voice (bright, dull) ANOVA

Source	***df***	**F**	**p**	**η²**
Permutation (P)	1,11	0.48	=.50	=.04
Condition (C)	1,11	3.21	=.10	=.23
Voice (V)	1,11	0.57	=.47	=.05
P x C	1,11	1.00	=.34	=.08
P x V	1,11	0.20	=.66	=.02
C x V	1,11	0	=1	0
P x C x V	1,11	0.31	=.59	=.03

EXPERIMENT 6 VTs – Three-way 2 perm. (seq.1, seq. 2) x 2 cond. (High(TP)Left/Low(TP)Right, High(TP)Right/Low(TP)Left) x 2 voice (right, left) ANOVA

Source	***df***	**F**	**p**	**η²**
Permutation (P)	**1,11**	**5.92**	**=.03***	**=.35**
Condition (C)	1,11	1.26	=.29	=.10
Voice (V)	1,11	0.37	=.56	=.03
P x C	1,11	0.60	=.46	=.05
P x V	1,11	0.56	=.47	=.05
C x V	1,11	0.06	=.81	=.01
P x C x V	**1,11**	**8.36**	**=.02***	**=.43**

EXPERIMENT 6 VTs – Three-way 2 perm. (seq.1, seq. 2) x 2 cond. (Front(TP)Left/Back(TP)Right, Front(TP)Right/Back(TP)Left) x 2 voice (right, left) ANOVA

Source	***df***	**F**	**p**	**η²**
Permutation (P)	1,11	0.02	=.88	<.01
Condition (C)	1,11	0.12	=.74	=.01
Voice (V)	1,11	0.64	=.44	=.06
P x C	1,11	0.44	=.52	=.04
P x V	1,11	0.04	=.86	<.01
C x V	**1,11**	**6.81**	**=.02***	**=.38**
P x C x V	1,11	0.04	=.85	<.01

EXPERIMENT 6 VTs – Two-way 2 permutation (seq. 1, seq. 2) x 3 condition (baseline, opposing, congruent) ANOVA

Source	***df***	**F**	**p**	**η²**
Permutation (P)	1,11	2.01	=.18	=.15
Condition (C)	2,22	0.42	=.66	=.04
P x C	2,22	2.00	=.16	=.15

EXPERIMENT 6 Forms – Three-way 2 perm. (seq.1, seq. 2) x 2 cond. (All Left, All Right) x 2 voice (bright, dull) ANOVA

Source	***df***	**F**	**p**	**η²**
Permutation (P)	1,11	1.62	=.23	=.13
Condition (C)	1,11	0.83	=.38	=.07
Voice (V)	1,11	0.39	=.55	=.03
P x C	1,11	1.51	=.24	=.12
P x V	1,11	0.02	=.90	<.01
C x V	1,11	<.01	=.96	<.01
P x C x V	1,11	0.59	=.46	=.05

EXPERIMENT 6 Forms – Three-way 2 perm. (seq.1, seq. 2) x 2 cond. (High(TP)Left/Low(TP)Right, High(TP)Right/Low(TP)Left) x 2 voice (right, left) ANOVA

Source	***df***	**F**	**p**	**η²**
Permutation (P)	1,11	4.10	=.07	=.27
Condition (C)	1,11	0.01	=.95	<.01
Voice (V)	1,11	0.29	=.60	=.03
P x C	1,11	0.06	=.81	=.01
P x V	1,11	0.21	=.66	=.02
C x V	1,11	<.01	=.97	<.01
P x C x V	**1,11**	**11.22**	**=.01***	**=.51**

EXPERIMENT 6 Forms – Three-way 2 perm. (seq.1, seq. 2) x 2 cond. (Front(TP)Left/Back(TP)Right, Front(TP)Right/Back(TP)Left) x 2 voice (right, left) ANOVA

Source	***df***	**F**	**p**	**η²**
Permutation (P)	1,11	0.60	=.45	=.05
Condition (C)	1,11	0.09	=.77	=.01
Voice (V)	1,11	0.99	=.34	=.08
P x C	1,11	0.04	=.85	<.01
P x V	1,11	0.11	=.74	=.01
C x V	**1,11**	**12.44**	**=.01***	**=.53**
P x C x V	1,11	0.08	=.79	=.01

Condition x Voice interaction

		Voice		
		Right	Left	
Condition	High(TP)Left/Low(TP)Right	4.33 *(0.68)*	2.58 *(0.36)*	3.46 *(0.42)*
	High(TP)Right/Low(TP)Left	2.88 *(0.44)*	3.88 *(0.41)*	3.38 *(0.40)*
		3.60 *(0.50)*	3.23 *(0.35)*	

EXPERIMENT 6 Forms – Two-way 2 permutation (seq. 1, seq. 2) x 3 condition (baseline, opposing, congruent) ANOVA

Source	***df***	**F**	**p**	**η^2**
Permutation (P)	1,11	2.59	=.14	=.19
Condition (C)	1,11	3.23	=.06	=.23
P x C	1,11	1.98	=.16	=.15

Means for main effect of condition
Baseline – 7.69 *(0.83)*
Opposing – 7.36 *(0.68)*
Congruent – 6.84 *(0.77)*

EXPERIMENT 6 Forms – Two-way 2 permutation (seq. 1, seq. 2) x 2 condition (opposing, congruent) ANOVA

Source	***df***	**F**	**p**	**η^2**
Permutation (P)	1,11	0.80	=.39	=.07
Condition (C)	1,11	3.39	=.09	=.24
P x C	**1,11**	**6.73**	**=.03***	**=.38**

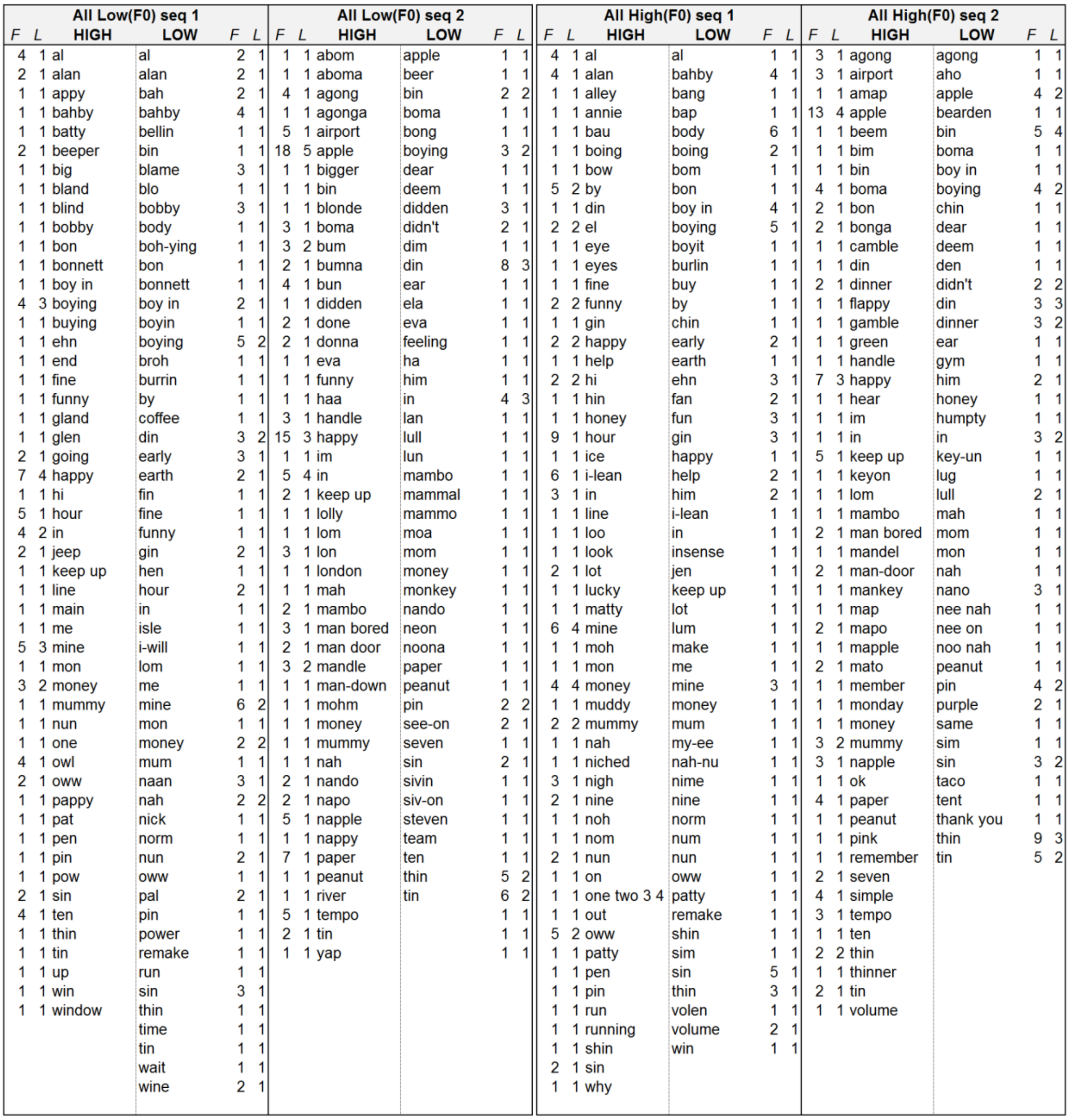
		Permutation		
		Seq. 1	Seq. 2	
Condition	Opposing	8.00 *(0.73)*	6.71 *(0.77)*	7.36 *(0.68)*
	Congruent	6.63 *(0.89)*	7.04 *(0.74)*	6.84 *(0.77)*
		7.32 *(0.78)*	6.88 *(0.73)*	

LSD Posthocs

S1-O vs S1-C, p=.01*
S1-O vs S2-O, p=.02*
S1-O vs S2-C, p=.06 [borderline]

Appendix 6 – *Collation of raw responses for all conditions in Experiments 5 and 6*

Appendix 6.1 – Distribution of raw responses for Experiment 5, Conditions *All Low(F0)* and *All High(F0)*. 'F' is a total number of responses for a given form and 'L' is a total number of listeners who reported that particular form.



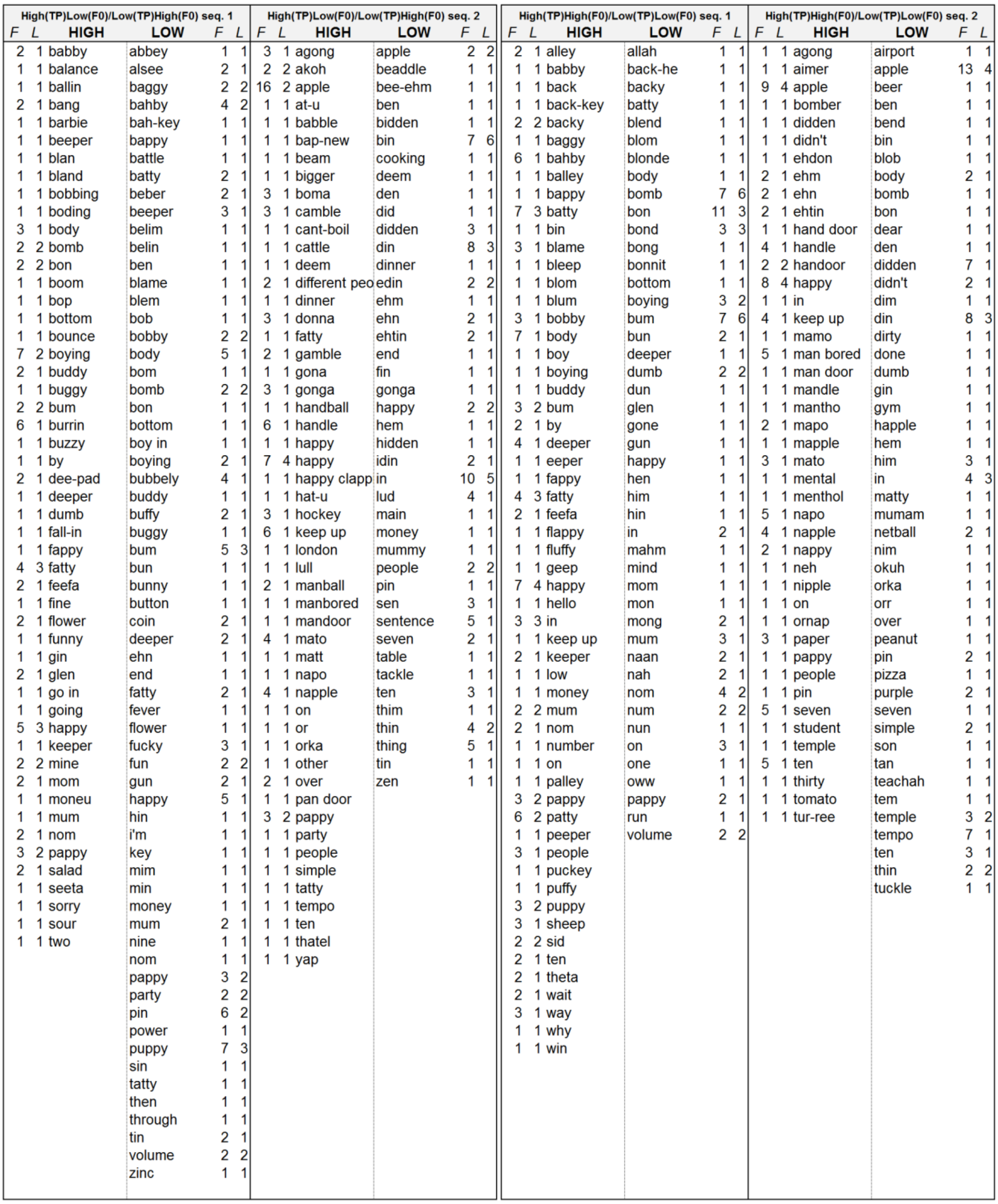
All Low(F0) seq 1					
F	*L*	HIGH	LOW	*F*	*L*
4	1	al	al	2	1
2	1	alan	alan	2	1
1	1	appy	bah	2	1
1	1	bahby	bahby	4	1
1	1	batty	bellin	1	1
2	1	beeper	bin	1	1
1	1	big	blame	3	1
1	1	bland	blo	1	1
1	1	blind	bobby	3	1
1	1	bobby	body	1	1
1	1	bon	boh-ying	1	1
1	1	bonnett	bon	1	1
1	1	boy in	bonnett	1	1
4	3	boying	boy in	2	1
1	1	buying	boyin	1	1
1	1	ehn	boying	5	2
1	1	end	broh	1	1
1	1	fine	burrin	1	1
1	1	funny	by	1	1
1	1	gland	coffee	1	1
1	1	glen	din	3	2
2	1	going	early	3	1
7	4	happy	earth	2	1
1	1	hi	fin	1	1
5	1	hour	fine	1	1
4	2	in	funny	1	1
2	1	jeep	gin	2	1
1	1	keep up	hen	1	1
1	1	line	hour	2	1
1	1	main	in	1	1
1	1	me	isle	1	1
5	3	mine	i-will	1	1
1	1	mon	lom	1	1
3	2	money	me	1	1
1	1	mummy	mine	6	2
1	1	nun	mon	1	1
1	1	one	money	2	2
4	1	owl	mum	1	1
2	1	oww	naan	3	1
1	1	pappy	nah	2	2
1	1	pat	nick	1	1
1	1	pen	norm	1	1
1	1	pin	nun	2	1
1	1	pow	oww	1	1
2	1	sin	pal	2	1
4	1	ten	pin	1	1
1	1	thin	power	1	1
1	1	tin	remake	1	1
1	1	up	run	1	1
1	1	win	sin	3	1
1	1	window	thin	1	1
			time	1	1
			tin	1	1
			wait	1	1
			wine	2	1

All Low(F0) seq 2					
F	*L*	HIGH	LOW	*F*	*L*
1	1	abom	apple	1	1
1	1	aboma	beer	1	1
4	1	agong	bin	2	2
1	1	agonga	boma	1	1
5	1	airport	bong	1	1
18	5	apple	boying	3	2
1	1	bigger	dear	1	1
1	1	bin	deem	1	1
1	1	blonde	didden	3	1
3	1	boma	didn't	2	1
3	2	bum	dim	1	1
2	1	bumna	din	8	3
4	1	bun	ear	1	1
1	1	didden	ela	1	1
2	1	done	eva	1	1
2	1	donna	feeling	1	1
1	1	eva	ha	1	1
1	1	funny	him	1	1
1	1	haa	in	4	3
3	1	handle	lan	1	1
15	3	happy	lull	1	1
1	1	im	lun	1	1
5	4	in	mambo	1	1
2	1	keep up	mammal	1	1
1	1	lolly	mammo	1	1
1	1	lom	moa	1	1
3	1	lon	mom	1	1
1	1	london	money	1	1
1	1	mah	monkey	1	1
2	1	mambo	nando	1	1
3	1	man bored	neon	1	1
2	1	man door	noona	1	1
3	2	mandle	paper	1	1
1	1	man-down	peanut	1	1
1	1	mohm	pin	2	2
1	1	money	see-on	2	1
1	1	mummy	seven	1	1
1	1	nah	sin	2	1
2	1	nando	sivin	1	1
2	1	napo	siv-on	1	1
5	1	napple	steven	1	1
1	1	nappy	team	1	1
7	1	paper	ten	1	1
1	1	peanut	thin	5	2
1	1	river	tin	6	2
5	1	tempo		1	1
2	1	tin		1	1
1	1	yap		1	1

All High(F0) seq 1					
F	*L*	HIGH	LOW	*F*	*L*
4	1	al	al	1	1
4	1	alan	bahby	4	1
1	1	alley	bang	1	1
1	1	annie	bap	1	1
1	1	bau	body	6	1
1	1	boing	boing	2	1
1	1	bow	bom	1	1
5	2	by	bon	1	1
1	1	din	boy in	4	1
2	2	el	boying	5	1
1	1	eye	boyit	1	1
1	1	eyes	burlin	1	1
1	1	fine	buy	1	1
2	2	funny	by	1	1
1	1	gin	chin	1	1
2	2	happy	early	2	1
1	1	help	earth	1	1
2	2	hi	ehn	3	1
1	1	hin	fan	2	1
1	1	honey	fun	3	1
9	1	hour	gin	3	1
1	1	ice	happy	1	1
6	1	i-lean	help	2	1
3	1	in	him	2	1
1	1	line	i-lean	1	1
1	1	loo	in	1	1
1	1	look	insense	1	1
2	1	lot	jen	1	1
1	1	lucky	keep up	1	1
1	1	matty	lot	1	1
6	4	mine	lum	1	1
1	1	moh	make	1	1
1	1	mon	me	1	1
4	4	money	mine	3	1
1	1	muddy	money	1	1
2	2	mummy	mum	1	1
1	1	nah	my-ee	1	1
1	1	niched	nah-nu	1	1
3	1	nigh	nime	1	1
2	1	nine	nine	1	1
1	1	noh	norm	1	1
1	1	nom	num	1	1
2	1	nun	nun	1	1
1	1	on	oww	1	1
1	1	one two 3 4	patty	1	1
1	1	out	remake	1	1
5	2	oww	shin	1	1
1	1	patty	sim	1	1
1	1	pen	sin	5	1
1	1	pin	thin	3	1
1	1	run	volen	1	1
1	1	running	volume	2	1
1	1	shin	win	1	1
2	1	sin			
1	1	why			

All High(F0) seq 2					
F	*L*	HIGH	LOW	*F*	*L*
3	1	agong	agong	1	1
3	1	airport	aho	1	1
1	1	amap	apple	4	2
13	4	apple	bearden	1	1
1	1	beem	bin	5	4
1	1	bim	boma	1	1
1	1	bin	boy in	1	1
4	1	boma	boying	4	2
2	1	bon	chin	1	1
2	1	bonga	dear	1	1
1	1	camble	deem	1	1
1	1	din	den	1	1
2	1	dinner	didn't	2	2
1	1	flappy	din	3	3
1	1	gamble	dinner	3	2
1	1	green	ear	1	1
1	1	handle	gym	1	1
7	3	happy	him	2	1
1	1	hear	honey	1	1
1	1	im	humpty	1	1
1	1	in	in	3	2
5	1	keep up	key-un	1	1
1	1	keyon	lug	1	1
1	1	lom	lull	2	1
1	1	mambo	mah	1	1
2	1	man bored	mom	1	1
1	1	mandel	mon	1	1
2	1	man-door	nah	1	1
1	1	mankey	nano	3	1
1	1	map	nee nah	1	1
2	1	mapo	nee on	1	1
1	1	mapple	noo nah	1	1
2	1	mato	peanut	1	1
1	1	member	pin	4	2
1	1	monday	purple	2	1
1	1	money	same	1	1
3	2	mummy	sim	1	1
3	1	napple	sin	3	2
1	1	ok	taco	1	1
4	1	paper	tent	1	1
1	1	peanut	thank you	1	1
1	1	pink	thin	9	3
1	1	remember	tin	5	2
2	1	seven			
4	1	simple			
3	1	tempo			
1	1	ten			
2	2	thin			
1	1	thinner			
2	1	tin			
1	1	volume			

Appendix 6.2 – Distribution of raw responses for Experiment 5, Conditions *High(TP)Low(F0)/Low(TP)High(F0)* and *High(TP)High(F0)/Low(TP)Low(F0)*. ‘F’ is a total number of responses for a given form and ‘L’ is a total number of listeners who reported that particular form.



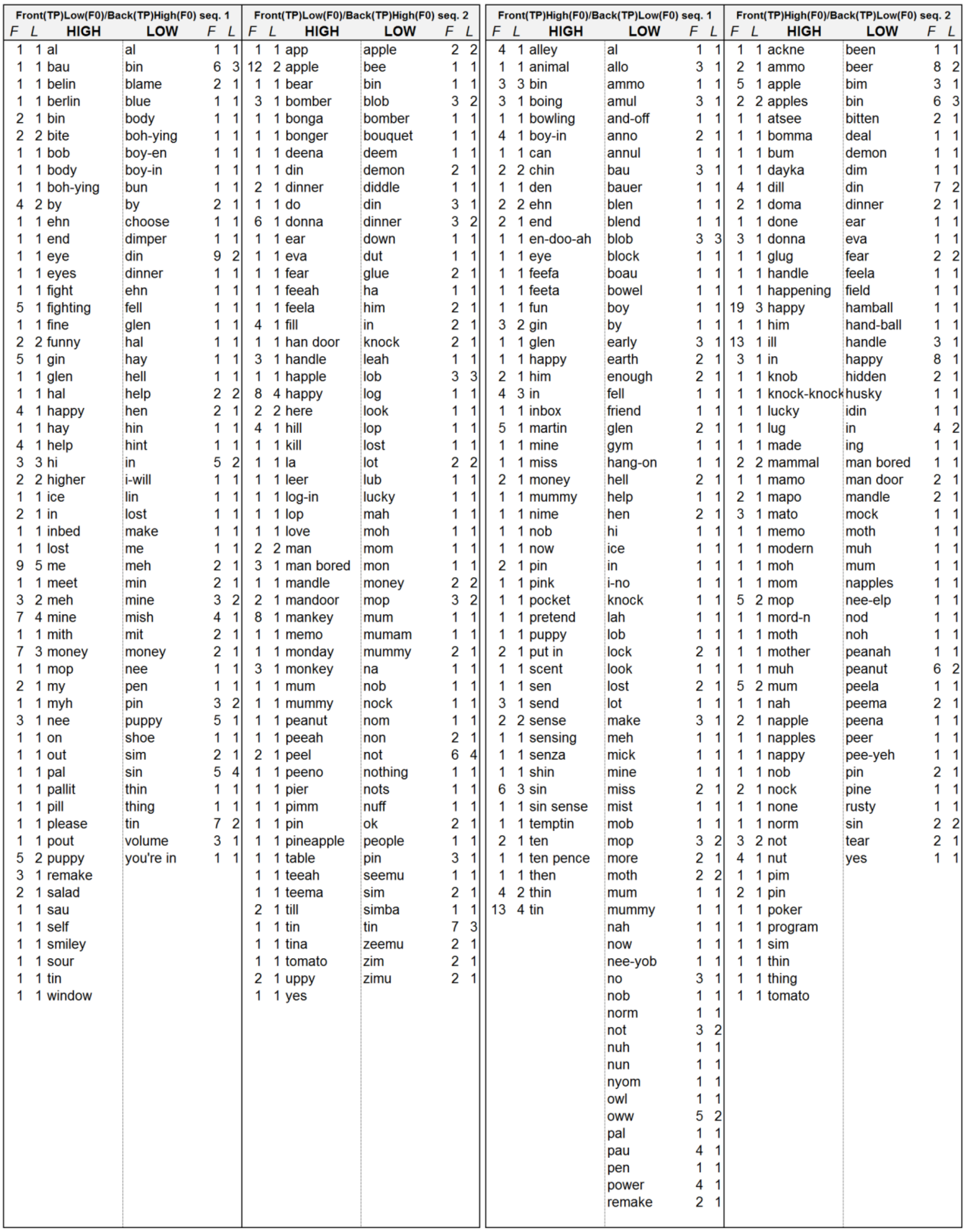
High(TP)Low(F0)/Low(TP)High(F0) seq. 1					
F	*L*	**HIGH**	**LOW**	*F*	*L*
2	1	babby	abbey	1	1
1	1	balance	alsee	2	1
1	1	ballin	baggy	2	2
2	1	bang	bahby	4	2
1	1	barbie	bah-key	1	1
1	1	beeper	bappy	1	1
1	1	blan	battle	1	1
1	1	bland	batty	2	1
1	1	bobbing	beber	2	1
1	1	boding	beeper	3	1
3	1	body	belim	1	1
2	2	bomb	belin	1	1
2	2	bon	ben	1	1
1	1	boom	blame	1	1
1	1	bop	blem	1	1
1	1	bottom	bob	1	1
1	1	bounce	bobby	2	2
7	2	boying	body	5	1
2	1	buddy	bom	1	1
1	1	buggy	bomb	2	2
2	2	bum	bon	1	1
6	1	burrin	bottom	1	1
1	1	buzzy	boy in	1	1
1	1	by	boying	2	1
2	1	dee-pad	bubbely	4	1
1	1	deeper	buddy	1	1
1	1	dumb	buffy	2	1
1	1	fall-in	buggy	1	1
1	1	fappy	bum	5	3
4	3	fatty	bun	1	1
2	1	feefa	bunny	1	1
1	1	fine	button	1	1
2	1	flower	coin	2	1
1	1	funny	deeper	2	1
1	1	gin	ehn	1	1
2	1	glen	end	1	1
1	1	go in	fatty	2	1
1	1	going	fever	1	1
5	3	happy	flower	1	1
1	1	keeper	fucky	3	1
2	2	mine	fun	2	2
2	1	mom	gun	2	1
1	1	moneu	happy	5	1
1	1	mum	hin	1	1
2	1	nom	i'm	1	1
3	2	pappy	key	1	1
2	1	salad	mim	1	1
1	1	seeta	min	1	1
1	1	sorry	money	1	1
1	1	sour	mum	2	1
1	1	two	nine	1	1
			nom	1	1
			pappy	3	2
			party	2	2
			pin	6	2
			power	1	1
			puppy	7	3
			sin	1	1
			tatty	1	1
			then	1	1
			through	1	1
			tin	2	1
			volume	2	2
			zinc	1	1

High(TP)Low(F0)/Low(TP)High(F0) seq. 2					
F	*L*	**HIGH**	**LOW**	*F*	*L*
3	1	agong	apple	2	2
2	2	akoh	beaddle	1	1
16	2	apple	bee-ehm	1	1
1	1	at-u	ben	1	1
1	1	babble	bidden	1	1
1	1	bap-new	bin	7	6
1	1	beam	cooking	1	1
1	1	bigger	deem	1	1
3	1	boma	den	1	1
3	1	camble	did	1	1
1	1	cant-boil	didden	3	1
1	1	cattle	din	8	3
1	1	deem	dinner	1	1
2	1	different peo	edin	2	2
1	1	dinner	ehm	1	1
3	1	donna	ehn	2	1
1	1	fatty	ehtin	2	1
2	1	gamble	end	1	1
1	1	gona	fin	1	1
3	1	gonga	gonga	1	1
1	1	handball	happy	2	2
6	1	handle	hem	1	1
1	1	happy	hidden	1	1
7	4	happy	idin	2	1
1	1	happy clapp	in	10	5
1	1	hat-u	lud	4	1
3	1	hockey	main	1	1
6	1	keep up	money	1	1
1	1	london	mummy	1	1
1	1	lull	people	2	2
2	1	manball	pin	1	1
1	1	manbored	sen	3	1
1	1	mandoor	sentence	5	1
4	1	mato	seven	2	1
1	1	matt	table	1	1
1	1	napo	tackle	1	1
4	1	napple	ten	3	1
1	1	on	thim	1	1
1	1	or	thin	4	2
1	1	orka	thing	5	1
1	1	other	tin	1	1
2	1	over	zen	1	1
1	1	pan door			
3	2	pappy			
1	1	party			
1	1	people			
1	1	simple			
1	1	tatty			
1	1	tempo			
1	1	ten			
1	1	thatel			
1	1	yap			

High(TP)High(F0)/Low(TP)Low(F0) seq. 1					
F	*L*	**HIGH**	**LOW**	*F*	*L*
2	1	alley	allah	1	1
1	1	babby	back-he	1	1
1	1	back	backy	1	1
1	1	back-key	batty	1	1
2	2	backy	blend	1	1
1	1	baggy	blom	1	1
6	1	bahby	blonde	1	1
1	1	balley	body	1	1
1	1	bappy	bomb	7	6
7	3	batty	bon	11	3
1	1	bin	bond	3	3
3	1	blame	bong	1	1
1	1	bleep	bonnit	1	1
1	1	blom	bottom	1	1
1	1	blum	boying	3	2
3	1	bobby	bum	7	6
7	1	body	bun	2	1
1	1	boy	deeper	1	1
1	1	boying	dumb	2	2
1	1	buddy	dun	1	1
3	2	bum	glen	1	1
2	1	by	gone	1	1
4	1	deeper	gun	1	1
1	1	eeper	happy	1	1
1	1	fappy	hen	1	1
4	3	fatty	him	1	1
2	1	feefa	hin	1	1
1	1	flappy	in	2	1
1	1	fluffy	mahm	1	1
1	1	geep	mind	1	1
7	4	happy	mom	1	1
1	1	hello	mon	1	1
3	3	in	mong	2	1
1	1	keep up	mum	3	1
2	1	keeper	naan	2	1
1	1	low	nah	2	1
1	1	money	nom	4	2
2	2	mum	num	2	2
2	1	nom	nun	1	1
1	1	number	on	3	1
1	1	on	one	1	1
1	1	palley	oww	1	1
3	2	pappy	pappy	2	1
6	2	patty	run	1	1
1	1	peeper	volume	2	2
3	1	people			
1	1	puckey			
1	1	puffy			
3	2	puppy			
3	1	sheep			
2	2	sid			
2	1	ten			
2	1	theta			
2	1	wait			
3	1	way			
1	1	why			
1	1	win			

High(TP)High(F0)/Low(TP)Low(F0) seq. 2					
F	*L*	**HIGH**	**LOW**	*F*	*L*
1	1	agong	airport	1	1
1	1	aimer	apple	13	4
9	4	apple	beer	1	1
1	1	bomber	ben	1	1
1	1	didden	bend	1	1
1	1	didn't	bin	1	1
1	1	ehdon	blob	1	1
2	1	ehm	body	2	1
2	1	ehn	bomb	1	1
2	1	ehtin	bon	1	1
1	1	hand door	dear	1	1
4	1	handle	den	1	1
2	2	handoor	didden	7	1
8	4	happy	didn't	2	1
1	1	in	dim	1	1
4	1	keep up	din	8	3
1	1	mamo	dirty	1	1
5	1	man bored	done	1	1
1	1	man door	dumb	1	1
1	1	mandle	gin	1	1
1	1	mantho	gym	1	1
2	1	mapo	happle	1	1
1	1	mapple	hem	1	1
3	1	mato	him	3	1
1	1	mental	in	4	3
1	1	menthol	matty	1	1
5	1	napo	mumam	1	1
4	1	napple	netball	2	1
2	1	nappy	nim	1	1
1	1	neh	okuh	1	1
1	1	nipple	orka	1	1
1	1	on	orr	1	1
1	1	ornap	over	1	1
3	1	paper	peanut	1	1
1	1	pappy	pin	2	1
1	1	people	pizza	1	1
1	1	pin	purple	2	1
5	1	seven	seven	1	1
1	1	student	simple	2	1
1	1	temple	son	1	1
5	1	ten	tan	1	1
1	1	thirty	teachah	1	1
1	1	tomato	tem	1	1
1	1	tur-ree	temple	3	2
			tempo	7	1
			ten	3	1
			thin	2	2
			tuckle	1	1

Appendix 6.3 – Distribution of raw responses for Experiment 5, Conditions *Front(TP)Low(F0)/Back(TP)High(F0)* and *Front(TP)High(F0)/Back(TP)Low(F0)*. 'F' is a total number of responses for a given form and 'L' is a total number of listeners who reported that particular form.



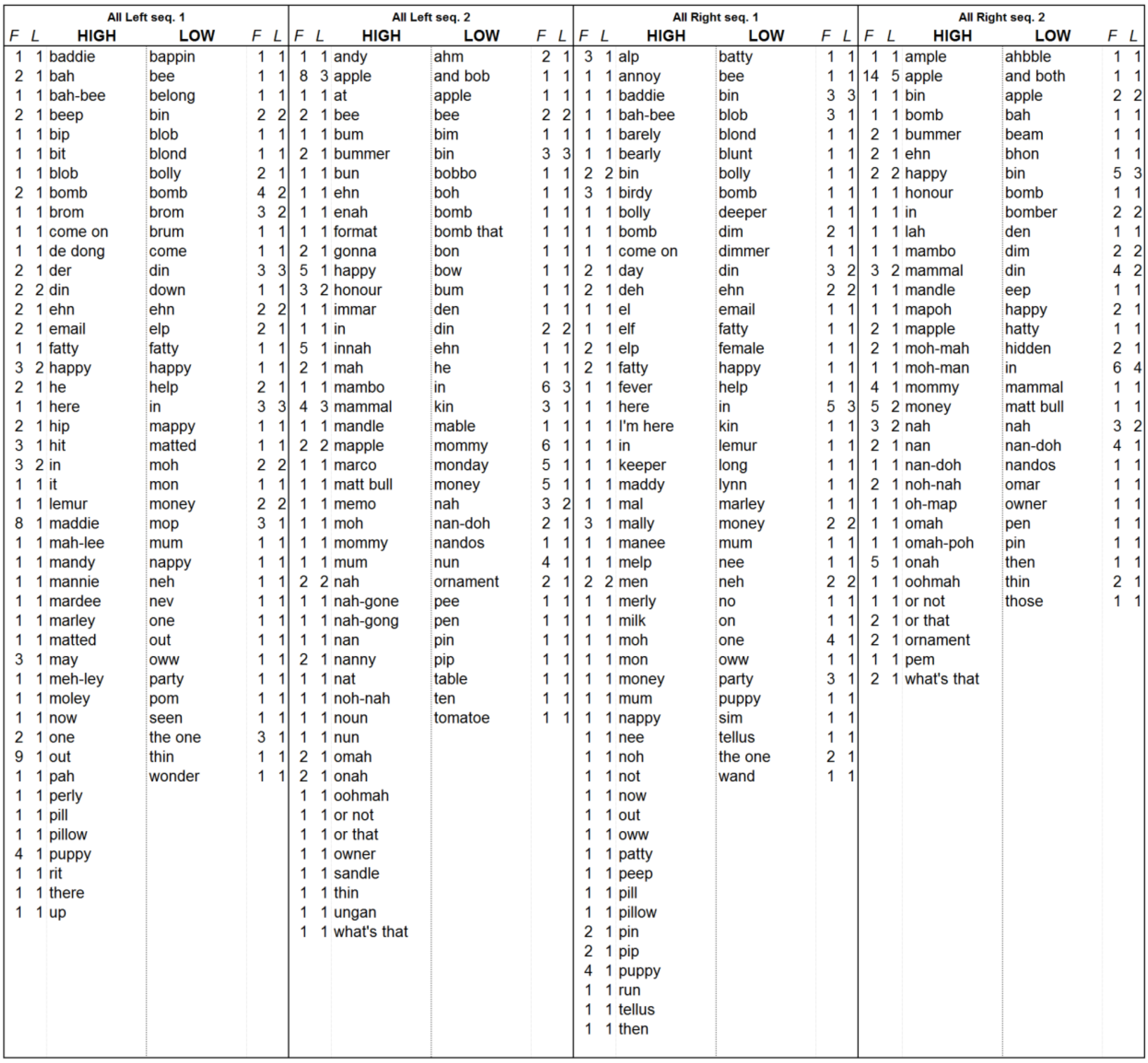
Front(TP)Low(F0)/Back(TP)High(F0) seq. 1					
F	*L*	**HIGH**	**LOW**	*F*	*L*
1	1	al	al	1	1
1	1	bau	bin	6	3
1	1	belin	blame	2	1
1	1	berlin	blue	1	1
2	1	bin	body	1	1
2	2	bite	boh-ying	1	1
1	1	bob	boy-en	1	1
1	1	body	boy-in	1	1
1	1	boh-ying	bun	1	1
4	2	by	by	2	1
1	1	ehn	choose	1	1
1	1	end	dimper	1	1
1	1	eye	din	9	2
1	1	eyes	dinner	1	1
1	1	fight	ehn	1	1
5	1	fighting	fell	1	1
1	1	fine	glen	1	1
2	2	funny	hal	1	1
5	1	gin	hay	1	1
1	1	glen	hell	1	1
1	1	hal	help	2	2
4	1	happy	hen	2	1
1	1	hay	hin	1	1
4	1	help	hint	1	1
3	3	hi	in	5	2
2	2	higher	i-will	1	1
1	1	ice	lin	1	1
2	1	in	lost	1	1
1	1	inbed	make	1	1
1	1	lost	me	1	1
9	5	me	meh	2	1
1	1	meet	min	2	1
3	2	meh	mine	3	2
7	4	mine	mish	4	1
1	1	mith	mit	2	1
7	3	money	money	2	1
1	1	mop	nee	1	1
2	1	my	pen	1	1
1	1	myh	pin	3	2
3	1	nee	puppy	5	1
1	1	on	shoe	1	1
1	1	out	sim	2	1
1	1	pal	sin	5	4
1	1	pallit	thin	1	1
1	1	pill	thing	1	1
1	1	please	tin	7	2
1	1	pout	volume	3	1
5	2	puppy	you're in	1	1
3	1	remake			
2	1	salad			
1	1	sau			
1	1	self			
1	1	smiley			
1	1	sour			
1	1	tin			
1	1	window			

Front(TP)Low(F0)/Back(TP)High(F0) seq. 2					
F	*L*	**HIGH**	**LOW**	*F*	*L*
1	1	app	apple	2	2
12	2	apple	bee	1	1
1	1	bear	bin	1	1
3	1	bomber	blob	3	2
1	1	bonga	bomber	1	1
1	1	bonger	bouquet	1	1
1	1	deena	deem	1	1
1	1	din	demon	2	1
2	1	dinner	diddle	1	1
1	1	do	din	3	1
6	1	donna	dinner	3	2
1	1	ear	down	1	1
1	1	eva	dut	1	1
1	1	fear	glue	2	1
1	1	feeah	ha	1	1
1	1	feela	him	2	1
4	1	fill	in	2	1
1	1	han door	knock	2	1
3	1	handle	leah	1	1
1	1	happle	lob	3	3
8	4	happy	log	1	1
2	2	here	look	1	1
4	1	hill	lop	1	1
1	1	kill	lost	1	1
1	1	la	lot	2	2
1	1	leer	lub	1	1
1	1	log-in	lucky	1	1
1	1	lop	mah	1	1
1	1	love	moh	1	1
2	2	man	mom	1	1
3	1	man bored	mon	1	1
1	1	mandle	money	2	2
2	1	mandoor	mop	3	2
8	1	mankey	mum	1	1
1	1	memo	mumam	1	1
1	1	monday	mummy	2	1
3	1	monkey	na	1	1
1	1	mum	nob	1	1
1	1	mummy	nock	1	1
1	1	peanut	nom	1	1
1	1	peeah	non	2	1
2	1	peel	not	6	4
1	1	peeno	nothing	1	1
1	1	pier	nots	1	1
1	1	pimm	nuff	1	1
1	1	pin	ok	2	1
1	1	pineapple	people	1	1
1	1	table	pin	3	1
1	1	teeah	seemu	1	1
1	1	teema	sim	2	1
2	1	till	simba	1	1
1	1	tin	tin	7	3
1	1	tina	zeemu	2	1
1	1	tomato	zim	2	1
2	1	uppy	zimu	2	1
1	1	yes			

Front(TP)High(F0)/Back(TP)Low(F0) seq. 1					
F	*L*	**HIGH**	**LOW**	*F*	*L*
4	1	alley	al	1	1
1	1	animal	allo	3	1
3	3	bin	ammo	1	1
3	1	boing	amul	3	1
1	1	bowling	and-off	1	1
4	1	boy-in	anno	2	1
1	1	can	annul	1	1
2	2	chin	bau	3	1
1	1	den	bauer	1	1
2	2	ehn	blen	1	1
2	1	end	blend	1	1
1	1	en-doo-ah	blob	3	3
1	1	eye	block	1	1
1	1	feefa	boau	1	1
1	1	feeta	bowel	1	1
1	1	fun	boy	1	1
3	2	gin	by	1	1
1	1	glen	early	3	1
1	1	happy	earth	2	1
2	1	him	enough	2	1
4	3	in	fell	1	1
1	1	inbox	friend	1	1
5	1	martin	glen	2	1
1	1	mine	gym	1	1
1	1	miss	hang-on	1	1
2	1	money	hell	2	1
1	1	mummy	help	1	1
1	1	nime	hen	2	1
1	1	nob	hi	1	1
1	1	now	ice	1	1
2	1	pin	in	1	1
1	1	pink	i-no	1	1
1	1	pocket	knock	1	1
1	1	pretend	lah	1	1
1	1	puppy	lob	1	1
2	1	put in	lock	2	1
1	1	scent	look	1	1
1	1	sen	lost	2	1
3	1	send	lot	1	1
2	2	sense	make	3	1
1	1	sensing	meh	1	1
1	1	senza	mick	1	1
1	1	shin	mine	1	1
6	3	sin	miss	2	1
1	1	sin sense	mist	1	1
1	1	temptin	mob	1	1
2	1	ten	mop	3	2
1	1	ten pence	more	2	1
1	1	then	moth	2	2
4	2	thin	mum	1	1
13	4	tin	mummy	1	1
			nah	1	1
			now	1	1
			nee-yob	1	1
			no	3	1
			nob	1	1
			norm	1	1
			not	3	2
			nuh	1	1
			nun	1	1
			nyom	1	1
			owl	1	1
			oww	5	2
			pal	1	1
			pau	4	1
			pen	1	1
			power	4	1
			remake	2	1

Front(TP)High(F0)/Back(TP)Low(F0) seq. 2					
F	*L*	**HIGH**	**LOW**	*F*	*L*
1	1	ackne	been	1	1
2	1	ammo	beer	8	2
5	1	apple	bim	3	1
2	2	apples	bin	6	3
1	1	atsee	bitten	2	1
1	1	bomma	deal	1	1
1	1	bum	demon	1	1
1	1	dayka	dim	1	1
4	1	dill	din	7	2
2	1	doma	dinner	2	1
1	1	done	ear	1	1
3	1	donna	eva	1	1
1	1	glug	fear	2	2
1	1	handle	feela	1	1
1	1	happening	field	1	1
19	3	happy	hamball	1	1
1	1	him	hand-ball	1	1
13	1	ill	handle	3	1
3	1	in	happy	8	1
1	1	knob	hidden	2	1
1	1	knock-knock	husky	1	1
1	1	lucky	idin	1	1
1	1	lug	in	4	2
1	1	made	ing	1	1
2	2	mammal	man bored	1	1
1	1	mamo	man door	2	1
2	1	mapo	mandle	2	1
3	1	mato	mock	1	1
1	1	memo	moth	1	1
1	1	modern	muh	1	1
1	1	moh	mum	1	1
1	1	mom	napples	1	1
5	2	mop	nee-elp	1	1
1	1	mord-n	nod	1	1
1	1	moth	noh	1	1
1	1	mother	peanah	1	1
1	1	muh	peanut	6	2
5	2	mum	peela	1	1
1	1	nah	peema	2	1
2	1	napple	peena	1	1
1	1	napples	peer	1	1
1	1	nappy	pee-yeh	1	1
1	1	nob	pin	2	1
2	1	nock	pine	1	1
1	1	none	rusty	1	1
1	1	norm	sin	2	2
3	2	not	tear	2	1
4	1	nut	yes	1	1
1	1	pim			
2	1	pin			
1	1	poker			
1	1	program			
1	1	sim			
1	1	thin			
1	1	thing			
1	1	tomato			

Appendix 6.4 – Distribution of raw responses for Experiment 6, Conditions *All Left* and *All Right*. 'F' is a total number of responses for a given form and 'L' is a total number of listeners who reported that particular form.



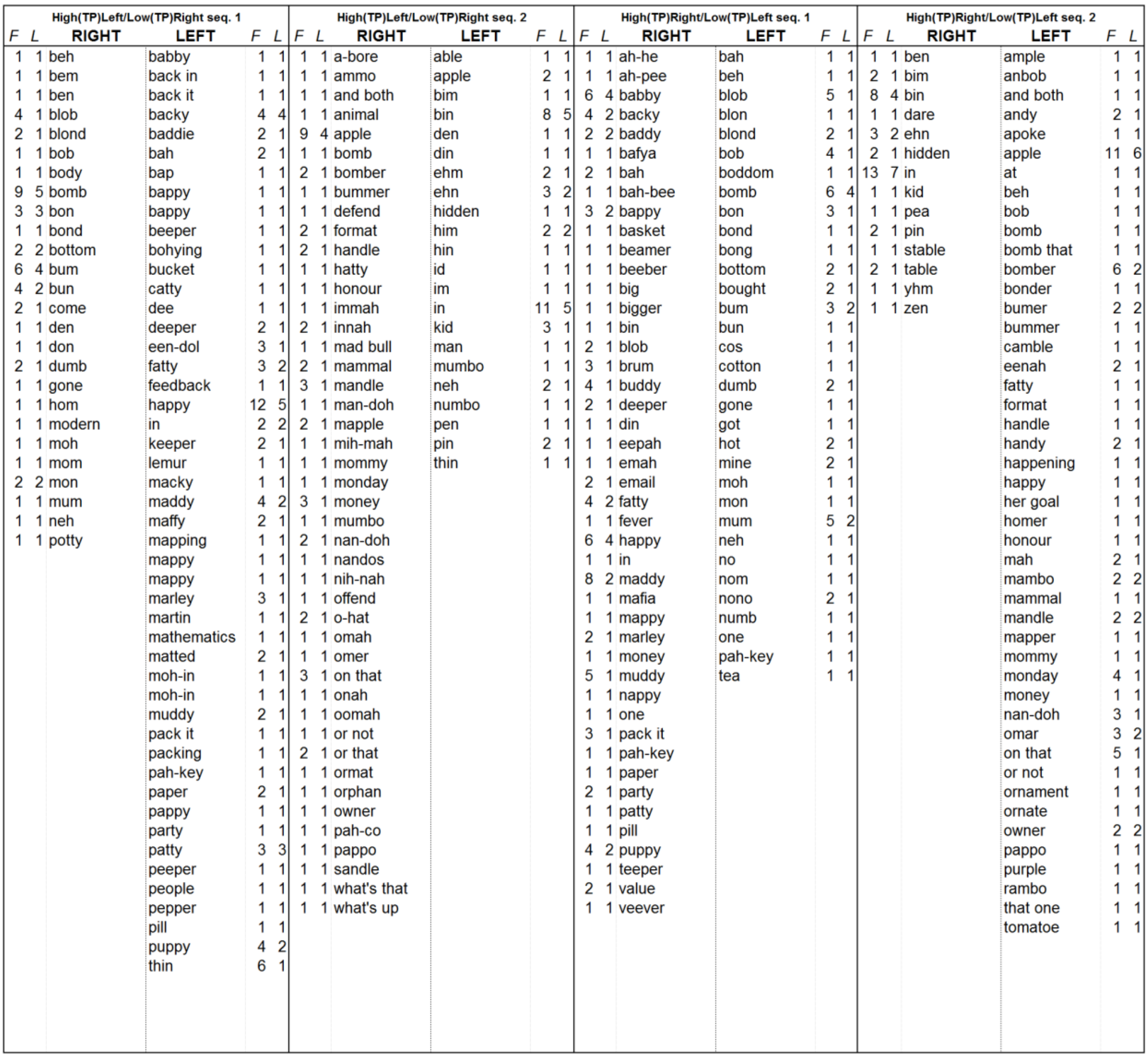
All Left seq. 1					
F	*L*	**HIGH**	**LOW**	*F*	*L*
1	1	baddie	bappin	1	1
2	1	bah	bee	1	1
1	1	bah-bee	belong	1	1
2	1	beep	bin	2	2
1	1	bip	blob	1	1
1	1	bit	blond	1	1
1	1	blob	bolly	2	1
2	1	bomb	bomb	4	2
1	1	brom	brom	3	2
1	1	come on	brum	1	1
1	1	de dong	come	1	1
2	1	der	din	3	3
2	2	din	down	1	1
2	1	ehn	ehn	2	2
2	1	email	elp	2	1
1	1	fatty	fatty	1	1
3	2	happy	happy	1	1
2	1	he	help	2	1
1	1	here	in	3	3
2	1	hip	mappy	1	1
3	1	hit	matted	1	1
3	2	in	moh	2	2
1	1	it	mon	1	1
1	1	lemur	money	2	2
8	1	maddie	mop	3	1
1	1	mah-lee	mum	1	1
1	1	mandy	nappy	1	1
1	1	mannie	neh	1	1
1	1	mardee	nev	1	1
1	1	marley	one	1	1
1	1	matted	out	1	1
3	1	may	oww	1	1
1	1	meh-ley	party	1	1
1	1	moley	pom	1	1
1	1	now	seen	1	1
2	1	one	the one	3	1
9	1	out	thin	1	1
1	1	pah	wonder	1	1
1	1	perly			
1	1	pill			
1	1	pillow			
4	1	puppy			
1	1	rit			
1	1	there			
1	1	up			

All Left seq. 2					
F	*L*	**HIGH**	**LOW**	*F*	*L*
1	1	andy	ahm	2	1
8	3	apple	and bob	1	1
1	1	at	apple	1	1
2	1	bee	bee	2	2
1	1	bum	bim	1	1
2	1	bummer	bin	3	3
1	1	bun	bobbo	1	1
1	1	ehn	boh	1	1
1	1	enah	bomb	1	1
1	1	format	bomb that	1	1
2	1	gonna	bon	1	1
5	1	happy	bow	1	1
3	2	honour	bum	1	1
1	1	immar	den	1	1
1	1	in	din	2	2
5	1	innah	ehn	1	1
2	1	mah	he	1	1
1	1	mambo	in	6	3
4	3	mammal	kin	3	1
1	1	mandle	mable	1	1
2	2	mapple	mommy	6	1
1	1	marco	monday	5	1
1	1	matt bull	money	5	1
1	1	memo	nah	3	2
1	1	moh	nan-doh	2	1
1	1	mommy	nandos	1	1
1	1	mum	nun	4	1
2	2	nah	ornament	2	1
1	1	nah-gone	pee	1	1
1	1	nah-gong	pen	1	1
1	1	nan	pin	1	1
2	1	nanny	pip	1	1
1	1	nat	table	1	1
1	1	noh-nah	ten	1	1
1	1	noun	tomatoe	1	1
1	1	nun			
2	1	omah			
2	1	onah			
1	1	oohmah			
1	1	or not			
1	1	or that			
1	1	owner			
1	1	sandle			
1	1	thin			
1	1	ungan			
1	1	what's that			

All Right seq. 1					
F	*L*	**HIGH**	**LOW**	*F*	*L*
3	1	alp	batty	1	1
1	1	annoy	bee	1	1
1	1	baddie	bin	3	3
1	1	bah-bee	blob	3	1
1	1	barely	blond	1	1
1	1	bearly	blunt	1	1
2	2	bin	bolly	1	1
3	1	birdy	bomb	1	1
1	1	bolly	deeper	1	1
1	1	bomb	dim	2	1
1	1	come on	dimmer	1	1
2	1	day	din	3	2
2	1	deh	ehn	2	2
1	1	el	email	1	1
1	1	elf	fatty	1	1
2	1	elp	female	1	1
2	1	fatty	happy	1	1
1	1	fever	help	1	1
1	1	here	in	5	3
1	1	I'm here	kin	1	1
1	1	in	lemur	1	1
1	1	keeper	long	1	1
1	1	maddy	lynn	1	1
1	1	mal	marley	1	1
3	1	mally	money	2	2
1	1	manee	mum	1	1
1	1	melp	nee	1	1
2	2	men	neh	2	2
1	1	merly	no	1	1
1	1	milk	on	1	1
1	1	moh	one	4	1
1	1	mon	oww	1	1
1	1	money	party	3	1
1	1	mum	puppy	1	1
1	1	nappy	sim	1	1
1	1	nee	tellus	1	1
1	1	noh	the one	2	1
1	1	not	wand	1	1
1	1	now			
1	1	out			
1	1	oww			
1	1	patty			
1	1	peep			
1	1	pill			
1	1	pillow			
2	1	pin			
2	1	pip			
4	1	puppy			
1	1	run			
1	1	tellus			
1	1	then			

All Right seq. 2					
F	*L*	**HIGH**	**LOW**	*F*	*L*
1	1	ample	ahbble	1	1
14	5	apple	and both	1	1
1	1	bin	apple	2	2
1	1	bomb	bah	1	1
2	1	bummer	beam	1	1
2	1	ehn	bhon	1	1
2	2	happy	bin	5	3
1	1	honour	bomb	1	1
1	1	in	bomber	2	2
1	1	lah	den	1	1
1	1	mambo	dim	2	2
3	2	mammal	din	4	2
1	1	mandle	eep	1	1
1	1	mapoh	happy	2	1
2	1	mapple	hatty	1	1
2	1	moh-mah	hidden	2	1
1	1	moh-man	in	6	4
4	1	mommy	mammal	1	1
5	2	money	matt bull	1	1
3	2	nah	nah	3	2
2	1	nan	nan-doh	4	1
1	1	nan-doh	nandos	1	1
2	1	noh-nah	omar	1	1
1	1	oh-map	owner	1	1
1	1	omah	pen	1	1
1	1	omah-poh	pin	1	1
5	1	onah	then	1	1
1	1	oohmah	thin	2	1
1	1	or not	those	1	1
2	1	or that			
2	1	ornament			
1	1	pem			
2	1	what's that			

Appendix 6.5 – Distribution of raw responses for Experiment 6, Conditions *High(TP)Left/Low(TP)Right* and *High(TP)Right/Low(TP)Left*. 'F' is a total number of responses for a given form and 'L' is a total number of listeners who reported that particular form.



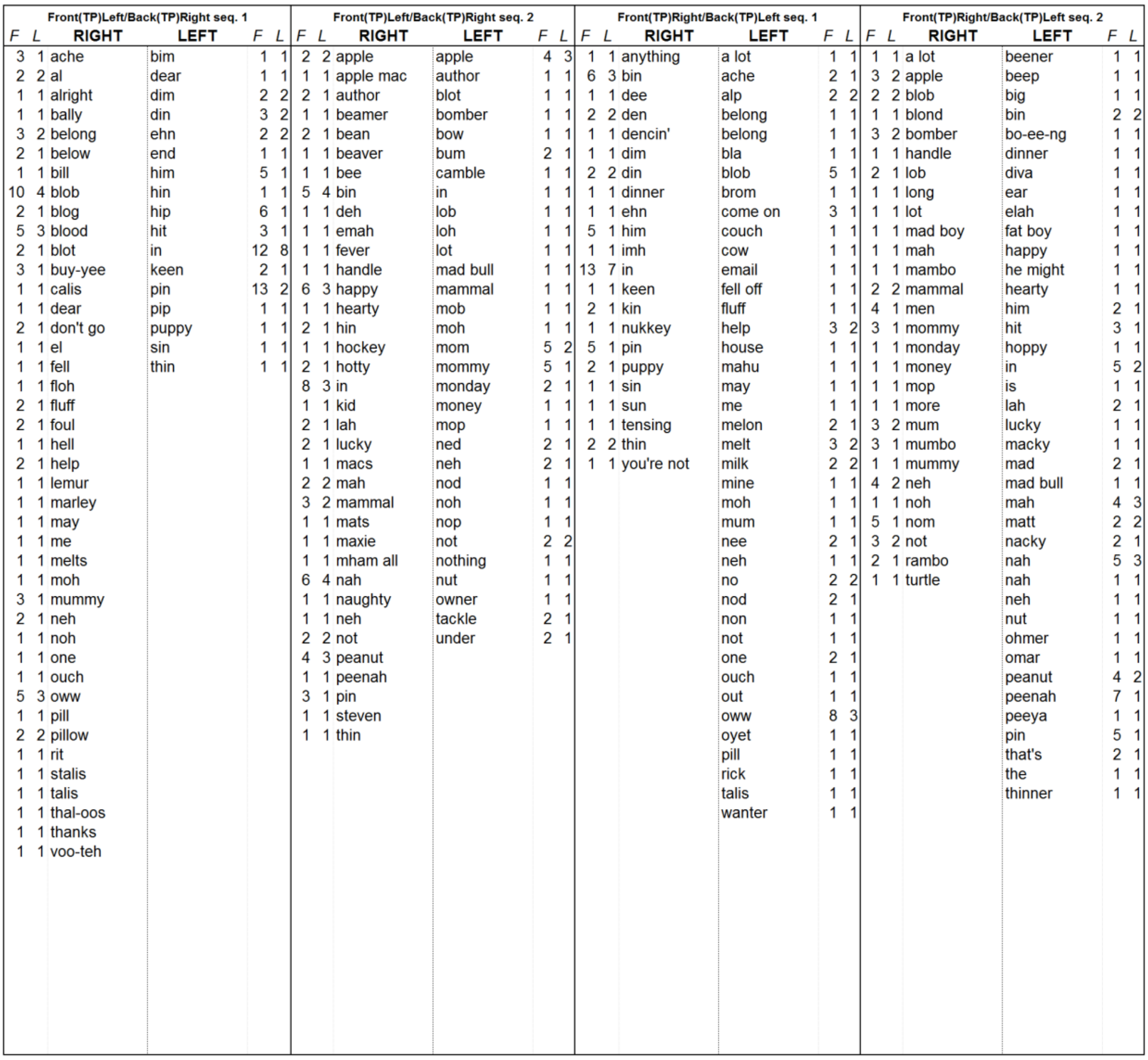
High(TP)Left/Low(TP)Right seq. 1					
F	L	RIGHT	LEFT	F	L
1	1	beh	babby	1	1
1	1	bem	back in	1	1
1	1	ben	back it	1	1
4	1	blob	backy	4	4
2	1	blond	baddie	2	1
1	1	bob	bah	2	1
1	1	body	bap	1	1
9	5	bomb	bappy	1	1
3	3	bon	bappy	1	1
1	1	bond	beeper	1	1
2	2	bottom	bohying	1	1
6	4	bum	bucket	1	1
4	2	bun	catty	1	1
2	1	come	dee	1	1
1	1	den	deeper	2	1
1	1	don	een-dol	3	1
2	1	dumb	fatty	3	2
1	1	gone	feedback	1	1
1	1	hom	happy	12	5
1	1	modern	in	2	2
1	1	moh	keeper	2	1
1	1	mom	lemur	1	1
2	2	mon	macky	1	1
1	1	mum	maddy	4	2
1	1	neh	maffy	2	1
1	1	potty	mapping	1	1
			mappy	1	1
			mappy	1	1
			marley	3	1
			martin	1	1
			mathematics	1	1
			matted	2	1
			moh-in	1	1
			moh-in	1	1
			muddy	2	1
			pack it	1	1
			packing	1	1
			pah-key	1	1
			paper	2	1
			pappy	1	1
			party	1	1
			patty	3	3
			peeper	1	1
			people	1	1
			pepper	1	1
			pill	1	1
			puppy	4	2
			thin	6	1

High(TP)Left/Low(TP)Right seq. 2					
F	L	RIGHT	LEFT	F	L
1	1	a-bore	able	1	1
1	1	ammo	apple	2	1
1	1	and both	bim	1	1
1	1	animal	bin	8	5
9	4	apple	den	1	1
1	1	bomb	din	1	1
2	1	bomber	ehm	2	1
1	1	bummer	ehn	3	2
1	1	defend	hidden	1	1
2	1	format	him	2	2
2	1	handle	hin	1	1
1	1	hatty	id	1	1
1	1	honour	im	1	1
1	1	immah	in	11	5
2	1	innah	kid	3	1
1	1	mad bull	man	1	1
2	1	mammal	mumbo	1	1
3	1	mandle	neh	2	1
1	1	man-doh	numbo	1	1
2	1	mapple	pen	1	1
1	1	mih-mah	pin	2	1
1	1	mommy	thin	1	1
1	1	monday			
3	1	money			
1	1	mumbo			
2	1	nan-doh			
1	1	nandos			
1	1	nih-nah			
1	1	offend			
2	1	o-hat			
1	1	omah			
1	1	omer			
3	1	on that			
1	1	onah			
1	1	oomah			
1	1	or not			
2	1	or that			
1	1	ormat			
1	1	orphan			
1	1	owner			
1	1	pah-co			
1	1	pappo			
1	1	sandle			
1	1	what's that			
1	1	what's up			

High(TP)Right/Low(TP)Left seq. 1					
F	L	RIGHT	LEFT	F	L
1	1	ah-he	bah	1	1
1	1	ah-pee	beh	1	1
6	4	babby	blob	5	1
4	2	backy	blon	1	1
2	2	baddy	blond	2	1
1	1	bafya	bob	4	1
2	1	bah	boddom	1	1
1	1	bah-bee	bomb	6	4
3	2	bappy	bon	3	1
1	1	basket	bond	1	1
1	1	beamer	bong	1	1
1	1	beeber	bottom	2	1
1	1	big	bought	2	1
1	1	bigger	bum	3	2
1	1	bin	bun	1	1
2	1	blob	cos	1	1
3	1	brum	cotton	1	1
4	1	buddy	dumb	2	1
2	1	deeper	gone	1	1
1	1	din	got	1	1
1	1	eepah	hot	2	1
1	1	emah	mine	2	1
2	1	email	moh	1	1
4	2	fatty	mon	1	1
1	1	fever	mum	5	2
6	4	happy	neh	1	1
1	1	in	no	1	1
8	2	maddy	nom	1	1
1	1	mafia	nono	2	1
1	1	mappy	numb	1	1
2	1	marley	one	1	1
1	1	money	pah-key	1	1
5	1	muddy	tea	1	1
1	1	nappy			
1	1	one			
3	1	pack it			
1	1	pah-key			
1	1	paper			
2	1	party			
1	1	patty			
1	1	pill			
4	2	puppy			
1	1	teeper			
2	1	value			
1	1	veever			

High(TP)Right/Low(TP)Left seq. 2					
F	L	RIGHT	LEFT	F	L
1	1	ben	ample	1	1
2	1	bim	anbob	1	1
8	4	bin	and both	1	1
1	1	dare	andy	2	1
3	2	ehn	apoke	1	1
2	1	hidden	apple	11	6
13	7	in	at	1	1
1	1	kid	beh	1	1
1	1	pea	bob	1	1
2	1	pin	bomb	1	1
1	1	stable	bomb that	1	1
2	1	table	bomber	6	2
1	1	yhm	bonder	1	1
1	1	zen	bumer	2	2
			bummer	1	1
			camble	1	1
			eenah	2	1
			fatty	1	1
			format	1	1
			handle	1	1
			handy	2	1
			happening	1	1
			happy	1	1
			her goal	1	1
			homer	1	1
			honour	1	1
			mah	2	1
			mambo	2	2
			mammal	1	1
			mandle	2	2
			mapper	1	1
			mommy	1	1
			monday	4	1
			money	1	1
			nan-doh	3	1
			omar	3	2
			on that	5	1
			or not	1	1
			ornament	1	1
			ornate	1	1
			owner	2	2
			pappo	1	1
			purple	1	1
			rambo	1	1
			that one	1	1
			tomatoe	1	1

Appendix 6.6 – Distribution of raw responses for Experiment 6, Conditions *Front(TP)Left/Back(TP)Right* and *Front(TP)Right/Back(TP)Left*. 'F' is a total number of responses for a given form and 'L' is a total number of listeners who reported that particular form.



Front(TP)Left/Back(TP)Right seq. 1						Front(TP)Left/Back(TP)Right seq. 2						Front(TP)Right/Back(TP)Left seq. 1						Front(TP)Right/Back(TP)Left seq. 2					
F	*L*	**RIGHT**	**LEFT**	*F*	*L*	*F*	*L*	**RIGHT**	**LEFT**	*F*	*L*	*F*	*L*	**RIGHT**	**LEFT**	*F*	*L*	*F*	*L*	**RIGHT**	**LEFT**	*F*	*L*
3	1	ache	bim	1	1	2	2	apple	apple	4	3	1	1	anything	a lot	1	1	1	1	a lot	beener	1	1
2	2	al	dear	1	1	1	1	apple mac	author	1	1	6	3	bin	ache	2	1	3	2	apple	beep	1	1
1	1	alright	dim	2	2	2	1	author	blot	1	1	1	1	dee	alp	2	2	2	2	blob	big	1	1
1	1	bally	din	3	2	1	1	beamer	bomber	1	1	2	2	den	belong	1	1	1	1	blond	bin	2	2
3	2	belong	ehn	2	2	2	1	bean	bow	1	1	1	1	dencin'	belong	1	1	3	2	bomber	bo-ee-ng	1	1
2	1	below	end	1	1	1	1	beaver	bum	2	1	1	1	dim	bla	1	1	1	1	handle	dinner	1	1
1	1	bill	him	5	1	1	1	bee	camble	1	1	2	2	din	blob	5	1	2	1	lob	diva	1	1
10	4	blob	hin	1	1	5	4	bin	in	1	1	1	1	dinner	brom	1	1	1	1	long	ear	1	1
2	1	blog	hip	6	1	1	1	deh	lob	1	1	1	1	ehn	come on	3	1	1	1	lot	elah	1	1
5	3	blood	hit	3	1	1	1	emah	loh	1	1	5	1	him	couch	1	1	1	1	mad boy	fat boy	1	1
2	1	blot	in	12	8	1	1	fever	lot	1	1	1	1	imh	cow	1	1	1	1	mah	happy	1	1
3	1	buy-yee	keen	2	1	1	1	handle	mad bull	1	1	13	7	in	email	1	1	1	1	mambo	he might	1	1
1	1	calis	pin	13	2	6	3	happy	mammal	1	1	1	1	keen	fell off	1	1	2	2	mammal	hearty	1	1
1	1	dear	pip	1	1	1	1	hearty	mob	1	1	2	1	kin	fluff	1	1	4	1	men	him	2	1
2	1	don't go	puppy	1	1	2	1	hin	moh	1	1	1	1	nukkey	help	3	2	3	1	mommy	hit	3	1
1	1	el	sin	1	1	1	1	hockey	mom	5	2	5	1	pin	house	1	1	1	1	monday	hoppy	1	1
1	1	fell	thin	1	1	2	1	hotty	mommy	5	1	2	1	puppy	mahu	1	1	1	1	money	in	5	2
1	1	floh				8	3	in	monday	2	1	1	1	sin	may	1	1	1	1	mop	is	1	1
2	1	fluff				1	1	kid	money	1	1	1	1	sun	me	1	1	1	1	more	lah	2	1
2	1	foul				2	1	lah	mop	1	1	1	1	tensing	melon	2	1	3	2	mum	lucky	1	1
1	1	hell				2	1	lucky	ned	2	1	2	2	thin	melt	3	2	3	1	mumbo	macky	1	1
2	1	help				1	1	macs	neh	2	1	1	1	you're not	milk	2	2	1	1	mummy	mad	2	1
1	1	lemur				2	2	mah	nod	1	1				mine	1	1	4	2	neh	mad bull	1	1
1	1	marley				3	2	mammal	noh	1	1				moh	1	1	1	1	noh	mah	4	3
1	1	may				1	1	mats	nop	1	1				mum	1	1	5	1	nom	matt	2	2
1	1	me				1	1	maxie	not	2	2				nee	2	1	3	2	not	nacky	2	1
1	1	melts				1	1	mham all	nothing	1	1				neh	1	1	2	1	rambo	nah	5	3
1	1	moh				6	4	nah	nut	1	1				no	2	2	1	1	turtle	nah	1	1
3	1	mummy				1	1	naughty	owner	1	1				nod	2	1				neh	1	1
2	1	neh				1	1	neh	tackle	2	1				non	1	1				nut	1	1
1	1	noh				2	2	not	under	2	1				not	1	1				ohmer	1	1
1	1	one				4	3	peanut							one	2	1				omar	1	1
1	1	ouch				1	1	peenah							ouch	1	1				peanut	4	2
5	3	oww				3	1	pin							out	1	1				peenah	7	1
1	1	pill				1	1	steven							oww	8	3				peeya	1	1
2	2	pillow				1	1	thin							oyet	1	1				pin	5	1
1	1	rit													pill	1	1				that's	2	1
1	1	stalis													rick	1	1				the	1	1
1	1	talis													talis	1	1				thinner	1	1
1	1	thal-oos													wanter	1	1						
1	1	thanks																					
1	1	voo-teh																					

Appendix 7 – *First responses for Experiments 5 and 6*

Appendix 7.1 – Listeners initial responses for each condition in Experiment 5. Empty cells denote no response.

	***All Low(F0)* seq.1**		***All Low(F0)* seq.2**		***All High(F0)* seq.1**		***All High(F0)* seq.2**	
	HIGH	**LOW**	**HIGH**	**LOW**	**HIGH**	**LOW**	**HIGH**	**LOW**
L1	up	blame	apple	dim	one two	volume	bim	honey
L2	me	lom	in	mammo	happy	lum	happy	lug
L3	mon	bah	bun	noona	mon	lot		nano
L4	happy	mum	happy	monkey	money	mum	mankey	mah
L5		mine	handle	tin	mine		handle	pin
L6	happy	mon	apple	neon	boying	bang	keyon	key-un
L7	hi	by	in	in	honey	by	mummy	bin
L8	pin	pin	apple	mambo	pin	remake	apple	apple
L9	one	run	lolly	pin	on	win	money	taco
L10	mummy	me	mah	moa	mummy	boying	dinna	mom
L11	main	bon	lon	eva	mine	my-ee	lom	mon
L12	pappy	norm	peanut	money	mummy	help	happy	bearden

	***High(TP)Low(F0) /Low(TP)High(F0)* seq.1**		***High(TP)Low(F0) /Low(TP)High(F0)* seq.2**		***High(TP)High(F0) /Low(TP)Low(F0)* seq.1**		***High(TP)High(F0) /Low(TP)Low(F0)* seq.2**	
	HIGH	**LOW**	**HIGH**	**LOW**	**HIGH**	**LOW**	**HIGH**	**LOW**
L1	fall-in	volume	happy	bim	blame	bomb	happy	nim
L2	bum	pappy	babble	in	blum	pappy	happy	in
L3	fatty	fun	ten	seven	baggy	bon		son
L4	happy	min	happy	bin	happy	one	happy	din
L5	fine	bah-key	handball	tin	mum	backy	handle	pin
L6	bottom	happy	matt	people	patty	bon	seven	apple
L7	feefa	ben	happy	in	theta	bomb	apple	in
L8		volume	deem	apple	wait	him	thirty	apple
L9	two	puppy	on	apple	money	volume	on	apple
L10	bobbing	puppy	apple	people	mum	mum	people	mumam
L11	boying	buggy	or	main	pukey	bon	ehtin	blob
L12	happy	bomb	pappy	happy	happy	bomb	happy	didn't

	***Front(TP)Low(F0) /Back(TP)High(F0)* seq.1**		***Front(TP)Low(F0) /Back(TP)High(F0)* seq.2**		***Front(TP)High(F0) /Back(TP)Low(F0)* seq.1**		***Front(TP)High(F0) /Back(TP)Low(F0)* seq.2**	
	HIGH	**LOW**	**HIGH**	**LOW**	**HIGH**	**LOW**	**HIGH**	**LOW**
L1	boying	volume	happy	bin	mon	blame	happy	bim
L2	meh	min	peeno	blob	in	blob	ammo	been
L3	nee	bun	din	mop	fun	not	nah	nod
L4	happy	meh	man	not	happy	nun	lucky	muh
L5	mine	tin	handle	bee	pin	enough	in	handle
L6	hal	bin	pimm	lot	tin	amul	moh	peanut
L7	hi	help	mummy	people	feefa	help	happy	in
L8	higher	hin	memo	deem		moth	mop	happy
L9	lost	lost	yes	money	pink	block	mother	deal
L10	puppy	money	man	mumam	mummy	mummy	mammal	bin
L11	mine	al	eva	lob	mine	lob	nob	eva
L12	me	in	happy	lot	bin	nob	nappy	mock

Appendix 7.2 – Listeners initial responses for each condition in Experiment 6. Empty cells denote no response.

	***All Left* seq.1**		***All Left* seq.2**		***All Right* seq.1**		***All Right* seq.2**	
	HIGH	**LOW**	**HIGH**	**LOW**	**HIGH**	**LOW**	**HIGH**	**LOW**
L1	moley	in	mammal	bim	el	money	bin	mammal
L2	din	moh	apple	pen	mon	deeper	apple	den
L3	up	down	nan	bon	pill	in	nah	bhon
L4	puppy	neh	happy	nan-doh	puppy	nee	money	nan-doh
L5	marley	money		money		marley	mommy	
L6	happy	the one	mammal	in	I'm here	wand	mammal	hidden
L7	fatty	mum	mum	bow	fatty	one	bomb	those
L8	now	in	mammal	in	milk	happy	mambo	in
L9	beep	moh	bee	nah	pip	bee	apple	bin
L10	rit	fatty	andy			party	nah	happy
L11	mah-lee	din	noh-nah	din	men	din	noh-nah	din
L12	baddy	mon	moh	bee	maddy	mum	oomah	eep
	***High(TP)Left /Low(TP)Right* seq.1**		***High(TP)Left /Low(TP)Right* seq.2**		***High(TP)Right /Low(TP)Left* seq.1**		***High(TP)Right /Low(TP)Left* seq.2**	
	RIGHT	**LEFT**	**RIGHT**	**LEFT**	**RIGHT**	**LEFT**	**RIGHT**	**LEFT**
L1	bomb	happy	animal	able	money	bomb	in	mammal
L2	moh	patty	apple	ehn	deeper	bomb	table	apple
L3	bon	puppy		in	in	bah	in	beh
L4	neh	pah-key	nan-doh	neh	pah-key	pah-key	pea	nan-doh
L5	bum	marley	mommy		blob	blob		mommy
L6	bond	keeper	what's up	hidden	happy	bomb	hidden	that one
L7	mum	fatty	apple	numbo	fatty	mum	dare	apple
L8	bomb	mappy	mammal	in	babby	mum	in	rambo
L9	blond	pack it	apple	pin	happy	moh	pin	apple
L10		patty	hatty		party			handy
L11	bem	mappy	nih-nah	bim	maddy	bun	bim	eenah
L12	mon	bap	bummer	bin	maddy	mon	in	bummer
	***Front(TP)Left /Back(TP)Right* seq.1**		***Front(TP)Left /Back(TP)Right* seq.2**		***Front(TP)Right /Back(TP)Left* seq.1**		***Front(TP)Right /Back(TP)Left* seq.2**	
	RIGHT	**LEFT**	**RIGHT**	**LEFT**	**RIGHT**	**LEFT**	**RIGHT**	**LEFT**
L1	blob	in	emah	owner	dim	email	bomber	hearty
L2	moh	in	nah	moh	din	moh	not	diva
L3	pill	in	in	mop	in	pill	mop	nah
L4	neh	in	lucky	neh	puppy	nee	neh	lucky
L5	blob	in		mommy		may	mommy	ear
L6	blot	in	not	lot	in	no	lob	matt
L7	one	dear	nah	bow	dee	one	mumbo	nah
L8	oww	pin	mammal	in	in	milk	mammal	ohmer
L9	blob	pin	bee	apple	in	oww	apple	pin
L10	me		happy		sun	me	mum	macky
L11	oww	bim	peenah	noh	imh	oww	neh	mah
L12	may	in	lah	mum	in	mum	mum	mad