A VERIFICATION EXPERIMENT OF THE SECOND FORMANT TRANSITION FEATURE AS A PERCEPTUAL CUE IN NATURAL SPEECH

BY

ROBERT MATTHEW CVENGROS

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Electrical and Computer Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2011

Urbana, Illinois

Adviser:

Associate Professor Jont Allen

ABSTRACT

Over the past century, a significant amount of of research has been devoted toward understanding the acoustic features that are used as perceptual cues in human speech perception. A brief history of this research is given, with emphasis on one important feature, the second formant (F2) transition. A review of historical arguments made for and against its role in the perception of speech, as well as theories that accentuate its significance, is provided. A verification experiment is run to evaluate the importance of this feature, along with two features within the consonantal release. Short Time Fourier Transform (STFT) modifications are made to consonant-vowel (CV) nonsense syllables to remove each of the tested acoustic features. Changes in listener response error are reported. An ANalysis Of VAriance (ANOVA) is used to determine the perceptual significance of the features in question. Perceptual scores provide strong evidence *against* the role of the F2 transition in speech perception and strong evidence *for* the role of the excited resonance frequency in the consonantal release.

To my parents, who have given unceasing love and opportunity
For my wife and unborn daughter, for whom I do all things
For my God, who has blessed me with this wonderful life

ACKNOWLEDGMENTS

I wish to thank Dr. Allen for his support and guidance, the US Army Corps of Engineers, the Department of Electrical and Computer Engineering at the University of Illinois, and Mimosa Acoustics for their financial support, and the Human Speech Recognition group for all the help provided.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 History

Nearly 60 years ago at Haskins Laboratory, landmark research on human speech perception was carried out by Alvin Liberman, Frank Cooper, Pierre Delattre, and Louis Gerstman. Using their then advanced Pattern Playback tool, they synthesized speech by converting hand-painted spectrograms to audible acoustic signals (Cooper *et al.*, 1952; Liberman *et al.*, 1954, 1967; Liberman, 1996), in order to develop a set of empirical data from which they could generate viable theories of speech perception.

By analyzing spectrograms of natural speech, they heuristically chose acoustic features as candidates for perceptual cues. One such feature was the acoustic burst (Cooper *et al.*, 1952; Liberman, 1996). They tested this feature in stop consonants by synthesizing burst plus vowel stimuli, while they varied the frequency and duration of the burst (Cooper *et al.*, 1952). Although these stimuli "were the farthest from readily recognizable speech" (Liberman, 1996, pgs. 12-14), they were surprised to see that the burst time and frequency was well correlated to the plosive consonant responses of the listeners. Moreover, the perceived consonant for a single burst in time and frequency changed given the vowel context. For example, low-frequency bursts often cued the consonant /b/, mid-frequency bursts cued /g/, and high-frequency bursts cued the consonant /d/. Figure 1.1 (left) shows the observed relationship between burst frequency, context vowel, and the perceived consonant.

They continued their investigation to find perceptual cues by researching the effect of a second acoustic feature, the F2 transition (Liberman *et al.*, 1954). In this experiment, they used the Pattern Playback tool to vary the slope of the F2 transition, and asked listeners to respond with either /b,g,d/

or /p,t,k/, depending whether they were in one of two testing groups. Based on this experiment, they determined that the F2 transition is a more robust cue for stop consonants than the burst:

> Despite the demonstration, by us and others, of extreme context sensitivity of the acoustic cues, some researchers have been concerned for many years to show that there are, nevertheless, invariant acoustic cues, implying, then, that no special theoretical exertions are necessary in order to account for invariant phonetic percepts. My own view of this matter has always been that, whatever the outcome of the seemingly never-ending search for acoustic invariants, the theoretical issue will remain largely untouched; for there is surely no question that the highly context-sensitive transitions *do* supply important information for phonetic perception – they can, indeed, be shown to be quite sufficient in many circumstances – and that incontrovertible fact must be accounted for. (Liberman, 1996, pg.16)

Later they developed the theory of *acoustic loci* to further explain how the F2 transition works as a cue, or rather, to describe what variables influence the way an F2 transition cue is perceived. The *acoustic locia* are virtual points in time and frequency prior to the syllable onset to which the F2 transition has to point for in order for listeners to preceive a given consonant, independant of the context vowel. They contended that the locus to which the F2 transition pointed is a sufficient cue if the first half of the transition is truncated; that is, if the transition pointed to but did not reach the locus. In effect, the *silence* in conjunction with the residual F2 transition works as a cue for stop consonants (Liberman *et al.*, 1955).

In both Cooper *et al.* (1952) and Liberman *et al.* (1954), the researchers at Haskins Laboratory found that there is a wide variability of acoustic features (bursts, F2 transitions) that could cue the same consonant and also that one acoustic feature can cue one of several consonants given a change in the context vowel, an effect called *coarticulation*. Figure 1.1 (left) shows that one burst at 1.5 [kHz] sounds like a /p/ in the context of some vowels but a /k/ in the context of others. Figure 1.1 (right) shows the synthetic sounds described by Liberman *et al.* (1967) to demonstrate coarticulation. For example, both the rising transition in /di/ and the falling transition in /du/ cue the /d/

stop consonant. This variability in acoustic cues caused Liberman and his colleagues to observe that speech sounds are *coarticulated* and, as a result, to assume that the acoustic cues for proximal consonants and vowels overlap in time. The variant nature of acoustic cues, combined with the assumption that *something* about speech transmission had to be invariant, caused Liberman *et al.* to theorize that the invariant part of speech is the *gestures* that formed the acoustic signal, rather than the acoustic signals themselves.

Motor Theory came to full maturity with Liberman and Mattingly (1985, 1989). Here they fully articulated the concept of *gestural transmission* of speech cues; i.e., speech is transmitted gesturally rather than acoustically; that is to say that the invariant speech cue is the place and manner of articulation rather than the actual short-time spectral properties of the acoustic waveform. Moreover, "gestures" did not necessarily designate the shape of the vocal tract, rather the more abstract neural commands that come from the brain to the vocal tract. Also articulated in these papers was the idea that speech perception requires a special cognitive process to which speech production and perception are connected.

Stevens and Blumstein assumed that perception is based on a combination of several features in the acoustic signal, or an *integrated cue*. It had been previously found that repeating one consonant to a listener before an identi-

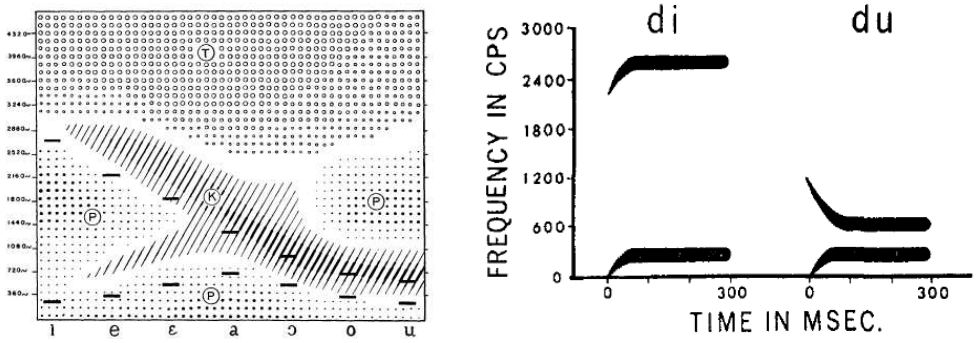

Figure 1.1: The historic Cooper *et al.* (1952) (left) and Liberman *et al.* (1967) (right) figures showing the effect of context changes on the way the burst and F2 transition are perceived. In the Cooper *et al.* (1952) figure, note that one burst at 1.4 [kHz] can be perceived as a /p/ in the context of some vowels but /k/ in the context of other vowels. In the Liberman *et al.* (1967) figure, note that the rising transition in /di/ and the falling transition in /du/ both cue the /d/ stop consonant.

fication task would cause an aversion to that consonant (Eimas and Corbit, 1973). Blumstein, Stevens, and Nigro (1977) took advantage of the stimulus fatigue (adapting) mechanism to assess the strength of some acoustic features using synthetic speech. For example, if /b/ is played many times to a listener before the task, the listener would be adverse to identifying a synthetic sound as /b/ during testing. Blumstein, Stevens, and Nigro (1977) used *adapting stimuli* for which the burst and transition were in agreement, for which the burst and transition were in disagreement, and for which there was no burst and only an F2 transition, and measured the amount of adaption in an identification task. Their hypothesis was that (i) a full-burst plus transitional adapting stimulus would cause the most fatigue for that respective consonant, (ii) an adapting stimulus for which the burst and transition were in conflict would cause moderate fatigue for each of the consonants in question, and (iii) a transition-only adapting stimulus would cause moderate fatigue for its respective consonant. Indeed, *they found this hypothesis to be true.* Stevens and Blumstein (1978) determined that the F2 transition feature is an adequate stand-alone cue for synthetic plosive sounds. Burst-only synthetic stimuli were limited to only 18% recognition, while F2 transition-only and full-cue stimuli achieved 81% and 90% respectively.

Cole and Scott (1974) followed a different explanation of how it is humans decode speech. They concluded that consonant-vowel utterances in real speech can be separated into three distinct parts: an invariant part, the transition, and the envelope. Although in some phonemes either the invariant part or the transition might be sufficient to uniquely identify the consonant, in most cases having only one of these acoustic features limited the identification to a small confusion group. Having both features would better allow listeners to easily identify the consonant. The role of the signal envelope is to tie together phonemes in conversational speech. Thus, in conversational speech, replacing one consonant for another with the same transition, regardless of a difference in the manner of articulation of the replacement, would only have a small effect on a listener's ability to interpret the speech.

Motor Theory has had a tremendous impact on the field of human speech perception as represented in the contemporary literature. Following the death of two of the strongest proponents of Motor Theory, Alvin Liberman and Ignatius Mattingly, Galantucci *et al.* (2006) reevaluated Motor Theory to determine which parts of it were viable. They broke Motor Theory down

into three parts: "(1) speech is special, (2) perceiving speech is perceiving vocal tract gestures, and (3) speech perception involves access to the speech motor system". According to the authors, (1) has either already been refuted (i.e., it is difficult to provide evidence for depending on the interpretation), but (2) and (3) are viable. For (2) they offer three arguments. First, they offer the original research by Cooper *et al.* (1952); Liberman *et al.* (1954, 1967). Second, they cite research which shows that listeners can be aided (or impeded) in the recognition task by other sensory input, such as the *McGurk effect*. Third, they offer that speech imitation is fast and resembles a *simple* task rather than a *choice* task, i.e., listeners need only 26 [ms] to interpret and reproduce speech sounds. This result is interpreted to be possible only if the listeners are perceiving gestures, since the recruitment of higher-level neural processing implies that a longer time would be needed for the listener to first interpret the sound.

Fogerty and Kewley-Port (2009) used two *glimpse* experiments to determine the strength of consonants and vowels in sentence stimuli, as well as to observe the value of transitional information to word recognition. Cooke (2003, pg. 579) defined a *glimpse* as "an arbitrary time-frequency region which contains a reasonably undistorted view of the target signal." Using labeled sentences and their phonemic boundaries specified in the TIMIT database, Fogerty and Kewley-Port (2009) replaced the consonant or vowel with noise that was 50 dB below the most intense vowel. The boundary of the replacement was varied from 10 to 50 percent of the vowel length (VP), resulting in C + VP (consonant plus vowel percentage) and V – VP (vowel minus vowel percentage) stimuli. By varying the boundary of the replacement and measuring the correct word scores, they were able to observe the relative contribution of consonants and vowels, as well as the transitional information, to recognition scores in the context of sentences. They concluded that vowels contain twice as much information as consonants in sentences. They also found that adding the transition portion of the signal did not significantly aid the recognition of vowel-only sentences; however, adding the transition information to consonant-only sentences had a significant impact on word recognition scores.

The concept of gesturally transmitted speech cues is common between Motor Theory and *Direct Realism*, although Direct Realism understands the concept in a different sense as Motor Theory (Fowler, 1986, 1996). Direct

Realism assumes knowledge of general human perception to describe perception within the auditory system. The theory argues that auditory perception is *real*, in that humans perceive distal objects *in the real world*. According to the theory, speech perception is also *direct*, in that humans directly perceive the distal object (i.e., vocal tract), rather than elements within the acoustic signal (Fowler, 1986). An analogous situation is that of sight; humans do not perceive individual photons but instead perceive the objects directly though brightness, colors, lines, etc. Likewise, they assume that our auditory system uses sound to directly infer the shape of the vocal tract. Thus the Direct Realism also adheres to the idea of gesturally transmitted cues. Whereas Motor Theory assumes that listeners perceive the neural commands issued to the vocal tract, Direct Realism states that listeners perceive the vocal tract position itself (Fowler, 1996). Often, advocates of Direct Realism use Liberman *et al.* (1967) and the concept of coarticulation to further their cause for gestural transmission, as well as the McGurk effect to emphasize the directness of speech transmission (Galantucci, Fowler, and Turvey, 2006; Fowler, 1996, 1986)(McGurk and MacDonald, 1976).

Not all theories and experiments are coherent with the 1950-1985 observations of the Haskins group. Stevens and Blumstein opposed the variable nature of the perceptual cues layed out by Liberman *et al.* (1967),

> Theories of speech perception developed in the past 20 years have been based on the notion that there is a lack of one-to-one correspondence between attributes of the acoustic signal and the phonetic percept. (Blumstein and Stevens, 1980)

and

> The motions of the formants immediately following the consonantal release, although contributing to place-of-articulation, are not essential, since eliminating movements of the second and higher formants still results in good identification performance of consonantal place of articulation. (Blumstein and Stevens, 1980)

Stevens and Blumstein (1978) continued to branch away from Motor Theory when they gave new experimental evidence that it is not the burst or transition that accounts for consonant recognition but instead the spectral

slope of the consonant onset. In this new view, the cue is due to the articulatory event (i.e., closure and release of the stop consonant) at a particular time, rather than a sequence of events over time, and is independent of the following vowel. They note that

> ... the formant transitions are *not* the primary cues signaling place of articulation. Instead their primary function... seems to be to join the onset spectrum to the vowel smoothly without introducing any additional discontinuities. Such discontinuities would, of course, signal new onsets. (Stevens and Blumstein, 1978)

The concept of the spectral onset cue was confirmed in Blumstein and Stevens (1979) with natural speech when they used spectrally sloped templates to classify CV stimuli. They found that by using templates to classify the spectral slope of the consonant onset they were able to achieve 85% accuracy. The fact that they were able to achieve this accuracy (15% error) implies that they were going in the right direction. However, note that the human error is more than an order of magnitude smaller (French and Steinberg, 1947; Fletcher and Galt, 1950; Allen, 2005; Phatak and Allen, 2007; Phatak *et al.*, 2008). According to Singh and Allen (2011), an error of 15% error is large.

Remez *et al.* (1981) decomposed traditional theories of speech perception by investigating sine wave speech. They used the short time spectra of the phrase "Where were you a year ago?" to derive three time-varying tones (T1+T2+T3) and presented these tones to listeners in different combinations and under three different instructional conditions (A,B,C). These sine wave stimuli lacked the transitional and/or onset cues that are commonly understood to be perceptually important. Once listeners were not told that the stimulus was speech, and when they were asked of their impression of the stimulus, only 5/31 listeners thought that the complete three-tone stimuli resembled human speech and only two of those five were able to identify the target phrase. When listeners were told that the stimulus was computer-generated speech, and were asked to transcribe the phrase, 9 listeners were able to transcribe the entire phrase, 10 recognized no sentence at all, and the others were able to transcribe only some of the syllables. When the target phrase was given to the listeners, in the three-tone case most listeners

felt they were confident that they actually heard the target phrase, on average they were able to recognize most of the words in the sentence, and most listeners found the stimulus to be unnatural sounding. Based on the results of these experiments, the authors concluded that listeners do not need the traditional transition and/or onset features in order to identify speech sounds.

Remez *et al.* (2008) went on to find the importance of the *synchrony* of the three tones from Remez *et al.* (1981) to the perception of speech. Fifteen sentences were converted to three-tone, sine wave speech, as in Remez *et al.* (1981, 1994), the tonal analog of the F2 transition region was temporally shifted, and the stimuli were presented to listeners for transcription. Results showed that recognition scores dropped from 72% to less than 10% as the feature was shifted in time, which allowed the authors to conclude that speech perception is quite intolerant of asynchrony. Remez *et al.* continue to see the strength in their method of reducing natural speech to sine wave speech as conceived in Remez *et al.* (1981), implying that they continue to see the lack of the traditional acoustic features (burst, F2 transition) as unimportant to the intelligibility of natural speech.

Dubno and Levitt (1981) attempted to determine the value of 11 acoustic features in both the quiet and noisy (speech-weighted) conditions for 91 naturally spoken utterances. Stimuli from the Nonsense Syllable Test (NST) (Dubno, 1978) were used that contained CV and VC utterances spoken by a male talker. They systematically varied a CV's level from 20 to 54 dB SPL, measured the acoustic difference of the variables over that range, and correlated it with the confusion rate. They found that in the quiet condition, the consonant energy, consonant duration, and the origin frequency of the second formant transition were most important. In the noisy condition, the consonant-to-noise ratio, consonant spectral peak frequency, and the consonant duration had the highest correlation with recognition scores. The importance of each of the acoustic variables varied from consonant to consonant, i.e., some features were important for voiced consonants but not for unvoiced sounds. The authors suggested that a useful future study would be to systematically remove each of the acoustic features and observe the resultant confusions.

Dubno *et al.* (1987) investigated the role of the consonant onset spectra in response to Stevens and Blumstein (1978) and Blumstein and Stevens

(1979). Using similar synthetic stimuli as those in Stevens and Blumstein (1978), they discovered that the duration of the onset spectra/voicing needed to be 20 [ms] or longer to have greater than 87% recognition rate with normal hearing listeners.

Turner *et al.* (1992) performed both detection and recognition experiments on synthetic sounds in a variable intensity of white noise signal-to-noise ratio (**SNR**). They used a Klatt Synthesizer to create burst + five formant transitions + 5 formant vowels stimuli. Part of their experiment was to truncate the stimuli to include only the first 40 [ms] of the sound. They presented both the short and long versions of the stimuli to listeners over varying **SNR** and found that there was no significant difference in consonant recognition between the two versions. Listeners were able to identify consonant based on the first 40 [ms] just as well as they could with the entire stimuli. This result supports Stevens and Blumstein (1978) and Blumstein and Stevens (1979) since it emphasizes that invariant cues are within the consonant onset, while it provides evidence against Liberman *et al.* (1967) since the short stimuli did *not* include vowel information.

Shannon *et al.* (1995) offered significant evidence *against* the role of the F2 transition feature as a perceptual cue by reducing the spectral content of the speech to 1, 2, 3, and 4 bands of noise. They used the envelope of the signal within designated frequency bands to modulate white noise and measured listener receognition scores. This operation significantly removed much of the spectral content of the signal; the F2 transitions were greatly diminished or even totally removed. Yet listeners responded with $> 90\%$ accuracy when the spectra was reduced to only 3 noise bands. This result provides a strong argument against the F2 transition as a cue.

Hazan and Simpson (1998) used models of speech perception to build a speech-enhancement algorithm for communication channels. Rather than removing the environmental noise that reduces intelligibility, they enhanced natural speech by making changes to the short time spectra of CV sounds. They made three types of changes to the sounds: they applied gain to the wide-band consonantal release region (C), they applied a pass band filter to the region of highest energy in the consonantal release (F), and they applied gain to the wide-band formant transitional region (T). This study showed that the consonantal release gain (C), alone or in combination with other enhancements, provided a 4–12% increase in intelligibility depending

on the noise condition and which other enhancements were combined with the (C) enhancement. On the other hand, boosting the (T) region offered no significant increase in the intelligibility of the speech. This is concordant to the idea that the consonantal release is the feature that cues consonant perception, rather than the transitional information.

Diehl, Lotto, and Holt (2004) pitted themselves against Motor Theory by reviewing evidence for and against the concept that speech requires a special coupling between productive and perceptive neural circuitry that is both speech-specific and human-specific. They argued that there is no such coupling since speech perception is not human-specific, considering that it was possible to train animals to respond to specific speech tokens, and that it is not speech-specific as many of the processes function similarly for non-speech sounds. They concluded that Motor Theory was challenged by having to find more compelling evidence for its claims. The authors also combated Direct Realism by noting that acoustic signals do not imply a unique vocal tract formation, a core assumption of the theory.

Li, Menon, and Allen (2010) used three independent experiments to develop the 3D Deep Search (3DDS) method of finding perceptual cues in natural speech. They investigated the time and frequency properties of speech cues for stop consonants by systematically truncating, high-pass/low-pass filtering, and changing the **SNR** of the signal and observing the resultant listener errors. Often these systematic changes to the signal resulted in a large increase in listener error over a small change in the variable. They used these time, frequency, and **SNR** thresholds to triangulate the perceptual cue over the relevant dimensions. For the stop consonants tested, the cues were determined to be one of the spectral peaks in the consonantal release, depending on the gesture with which the sound was articulated. Labial front sounds had cues near the first formant frequencies, glottal back sounds had cues near second formant resonance, and alveolar mid consonants had cues at higher (third or greater) formant resonances. The truncation experiment is significant to the discussion of the second formant transition since listener error dramatically increases when the consonantal release is removed from the signal.

Li and Allen (2011) used the information of speech cues from Li, Menon, and Allen (2010) to manipulate nonsense syllables in order to demonstrate that it is possible to change the perceived sound by manipulating small time-

frequency regions of the signal, while removing other, sometimes larger, time-frequency regions *reduced* the number of listener confusions. Both Li, Menon, and Allen (2010) and Li and Allen (2011) emphasize the importance of the relevant spectral peaks in the consonant release rather than the consonant vowel transitional information.

1.2 Discussion and Criticism

The strength of a model is judged by its ability to predict experimental outcomes and to verify and enlighten understanding. Theories and models must be built on experimental data to succinctly summarize what we know.

We propose that a *viable* model of speech perception must be able to make predictions about natural speech since, in the end, our goal is to be able to predict what we hear in everyday conversation; our goal is *not* to predict how a listener will respond to synthetic stimuli. Furthermore, perceptual results need to coincide with neural processing models in order to better understand *how* humans perceive speech. Stevens and Blumstein exemplify this since they developed their onset template theory with synthetic sounds in 1978 and verified it with natural speech in 1979.

Often poor-quality speech sounds are used in experiments that become the basis of theory (Liberman, 1996; Remez *et al.*, 1981). Using poor-quality stimuli is not an appropriate way to evaluate the perceptual significance of acoustic features in natural speech since the quality of natural speech is extremely high (0% lister error) (Singh and Allen, 2011). Other studies attempt to reevaluate old data in order build or critique competing theories (Fowler, 1986, 1996; Diehl *et al.*, 2004; Galantucci *et al.*, 2006) rather than employing new experiments. Without any insight from new data to support their points, these arguments seem ad hoc.

The Motor Theory model was created based on past experimental data using synthetic speech. Motor Theory and Direct Realism state that, with synthetic sounds, the F2 transition codes for place of articulation information; much of these theories have been built upon this single observation. Motor Theory predicts the significance of this feature within natural speech; but for the theory to remain viable, this prediction must be verified. Liberman et. al. did not give priority to such a verification in their research; the

vast majority of their evidence investigated synthetic stimuli.

To our knowledge, the results of Liberman *et al.* (1954) and Liberman (1957) have never been experimentally verified using natural speech. This verification is especially important since their original presentation methods could have been significantly refined with time. For example, the synthetic stimuli were of low quality, which was explicitly recognized by Liberman (1996, 12-14). Listeners were able to respond with only three consonants, either /b,d,g/ or /p,t,k/. One wonders if they would have achieved the same result if a greater number of consonants was allowed in the response set.

The large variety of articulated theories and results shows that important unaddressed questions remain open. It is the goal of this study to quantify the role of the F2 transition in natural speech by removing it from acoustic signal and observing the resultant change in perceptual scores of the speech token. This experiment was not done by Liberman *et al.* but was suggested by Dubno (1978).

CHAPTER 2

EXPERIMENT

2.1 Methods

The goal of this study is to quantify the role of the F2 transition feature as a speech cue. Our experimental method is to remove this feature and analyze the listener responses.

It is necessary to define our terms. *Acoustic feature* is used to describe any part of the physical domain acoustic signal that is identifiable in a spectrogram. An acoustic feature is a *perceptual cue* if it is essential for a listener to recognize a given speech sound.

The *F2 trajectory* is the first 5-10 [cs] of the second formant energy. If the F2 trajectory is associated with a rapid change in the frequency of this energy, the feature is an *F2 transition*. Within the context of this document, however, this feature will always be referred to as the *F2 trajectory* since only 15 of the 25 sounds studied contain rapid frequency changes during the first 5-10 [cs] of the second formant energy.

2.1.1 Stimuli

A set of 25 consonant-vowel (CV) utterances was taken from the Linguistic Data Consortium (LDC) database to test plosive consonants /d,g,k,p/ in the context of the vowels /a,ɪ,u,æ/. Two utterances for each CV combination of /d,g,k/ and /a,ɪ,u,æ/ were used; however, only one utterance of the CV /pɪ/ was used for the consonant /p/ due to a lack of easily distinguishable features for this consonant. Thirteen utterances were spoken by eight female talkers and twelve stimuli were spoken by seven male talkers, however this distribution was not uniform across CVs. Instead, the utterances that were chosen were shown to be the most robust in noise given data from Phatak and

Allen (2007). Figure 2.1 shows the probability of listener error across **SNR** in white noise. Note that all stimuli have little (1 error in 34 trials, 1 error in 36 trials for m112 /dæ/ and f101 /kæ/, respectively) or no errors in more than 30 trials for **SNR** conditions above 0 dB. 150 seed sounds were added to limit listener response bias toward the tested plosives. The experiment then included 25 utterances x 16 conditions + 150 seed sounds = 550 total stimuli. All stimuli were tested in 12 dB **SNR** of white noise. The RMS of the signal and noise level was determined by taking the max of the RMS vector after a $\tau = 20[cs]$ exponential mean filter as described in Appendix D.

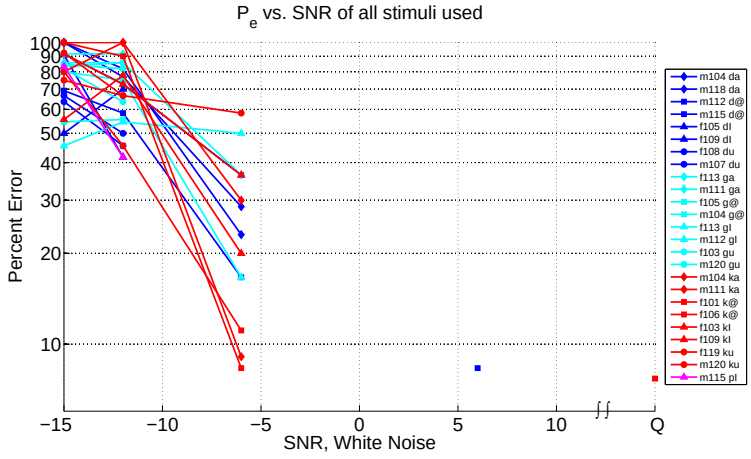

Figure 2.1: The probability of listener error across **SNR** in white noise. Note that all stimuli have little or no error at and above 0 dB **SNR**.

AI Gram

All signals are quantified with the AI gram (Regnier and Allen, 2008). Unlike the short time Fourier transform (STFT), which uses the the Fourier Transform's linearly spaced bands, the AI gram uses critical band filters to evenly sample the basilar membrane of the cochlea. The pixel value in the image is the critical band **SNR** of the signal.

Short Time Fourier Transform Modifications

The STFT (Allen, 1977; Allen and Rabiner, 1977) was used to make modifications to the speech waveforms. The method is similar to that in Li and Allen (2011).

A moving Kaiser window ($\beta = 12.12174$)) $w(t)$ was used to isolate the signal $s(t)$, followed by a Fourier Transform to derive the STFT.

$$y(t) = s(t) \cdot w((t))_D \tag{2.1}$$
$$y(t) = s(t) \cdot (w(t) * \delta((t))_D) \tag{2.2}$$

where $x((\cdot))_D$ denotes that the signal is periodic with $period = D$ seconds. D is the step size of the moving window. This equation has equivalent representation in the frequency domain

$$Y(\omega) = S(\omega) * \left(W(\omega) \cdot \delta((\omega))_{\frac{2\pi}{D}} \right) \tag{2.3}$$

If $W(\omega)$ is a low-pass function with bandwidth B and the cutoff frequency $\frac{B}{2} < \frac{2\pi}{D}$, then only the DC component of $\delta((\omega))_{\frac{2\pi}{D}}$ will be passed. The equation then simplifies to

$$Y(\omega) = S(\omega) * \hat{\delta}(\omega) \tag{2.4}$$
$$\hat{Y}(\omega) = \hat{S}(\omega) \tag{2.5}$$

There is some residual error in $Y(\omega)$ corresponding to the stop band attenuation of $W(\omega)$; however, for this Kaiser window the magnitude of this error is $-91[dB]$ below the signal and can be considered negligible.

First a portion of the signal is multiplied by the window, and then the product is overlap-added (OLA) with the rest of the windowed frames. After the signal has been windowed, the FFT is taken of each short time frame to form the STFT/spectrogram image. The original signal can then be resynthesized with negligible error ($-91dB$) by performing an inverse FFT on each frame of the STFT image and using the OLA operation.

Short time spectral modifications can be made to the signal by multiplying the STFT with a 2-dimensional gain. Multiplying the spectra of a signal with a frequency-dependent gain performs a circular convolution in the time domain, which is artifact-free since the signal has been zero padded. The

OLA resynthesized waveform contains these modifications to the STFT. This method has proven to be highly accurate and artifact-free.

Masking Regions

As shown in Figure 2.2, four feature regions were chosen to test the hypotheses: the non-burst consonant wide-band onset (N), the primary consonant burst (B), the F2 onset (O), and the F2 trajectory (T). These regions were identified in the AI gram of the sound, and 2-dimensional masks were created for each feature. These masks were combined with other masks of the test condition, and then the combination was multiplied with the STFT followed by an OLA resynthesis. Feature masks were handpicked to capture these features in the acoustic signal. Natural acoustic boundaries were used to separate feature regions when proximate. Example feature masks for each of the consonants /d,g,k,p/ are shown in Figure 2.2.

The word "burst" in the context of this experiment is *not* the articulated plosive burst/release as in some experiments that use natural speech; instead the word "burst" corresponds to a region of high energy in the consonant onset, similar to its usage by Cooper *et al.* (1952); Li *et al.* (2010). Combining feature regions N and B are necessary to capture the articulated wideband plosive release within the context of this experiment. A consonantal release is physically represented by a wide-band signal with significant spectral peaks at the formant frequencies, resulting from the resonance with the vocal tract.

A particular vocal tract gesture for a plosive sound can excite the resonant frequencies corresponding to the formant frequencies of the subsequent vowel. These formant resonances are excited in different ways depending on which consonant is being articulated. Labial front releases often have greater excitation near the first formant frequencies, glottal back releases tend to excite the second formant resonance, and alveolar mid consonants excite higher (third or greater) formant resonances. The tendancies toward certain resonant frequencies does not exclude the possibility of energy at other frequencies. However consonant energy spectral peaks tend toward the formant frequencies associated with the place of release. The Burst feature (B) corresponds to the excited resonance F_1, F_2, or F_3 of the resonance that is producing the sound. For example, the consonants /g,k/ excites the second formant resonance F_2, while F_1, F_3 are only weakly excited. In the context

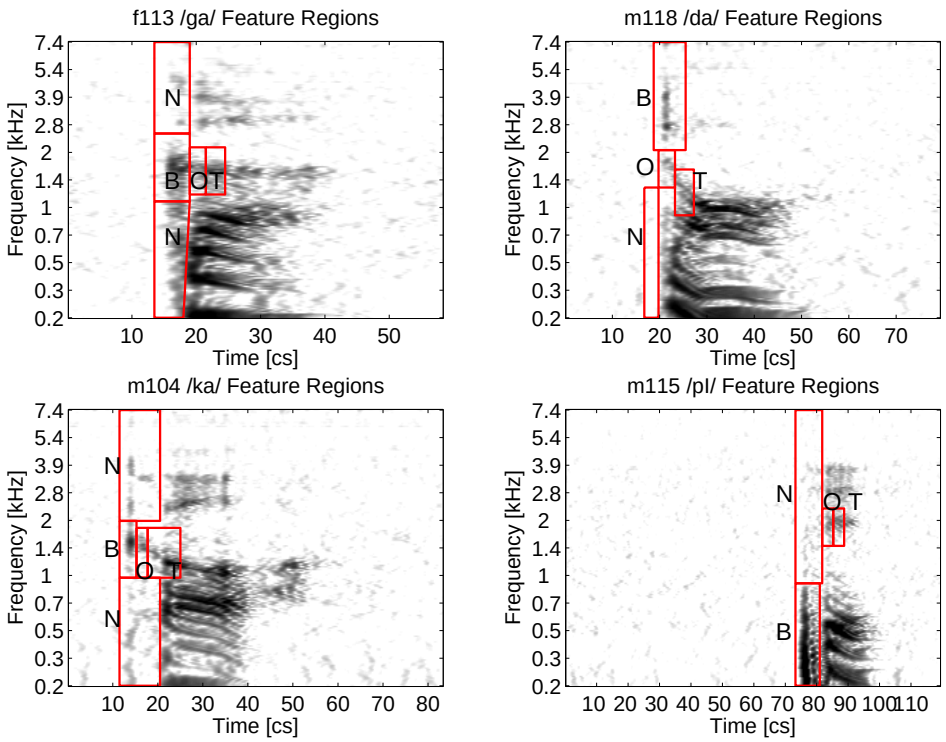

Figure 2.2: Example feature masks for f113 /ga/ (far left), m118 /da/ (center left), m104 /ka/ (center right), and m115 /pɪ/ (far right). Label B denotes the primary burst feature, label N denotes the non-burst consonant onset, O denotes the F2 onset, and T denotes the F2 trajectory.

of this document, the Burst feature (B) refers to F_2 and the non-burst consonant onset (N) refers to both F_1, F_3. Likewise the consonant /d/ excites F_3 and leaves F_1, F_2 weakly excited, and the consonant /p/ excites F_1 but weakly excites F_2, F_3. Li and Allen (2011) refer to the (N) feature as the *Conflicting Cues*.

Both the primary burst and the non-burst onset occur during the consonantal release 2-4 [cs] before the vowel onset for voiced plosives and 10-15 [cs] for unvoiced plosives. The F2 trajectory is the first 5-10 [cs] of the second formant energy. In this experiment this feature is separated into two different feature masks; the F2 trajectory onset (O), which includes the first 2-5 [cs] of the F2 trajectory, and the greater F2 trajectory (T), which is the following 3-6 [cs] of the second formant energy. The combined OT mask represents the entire initial 5-10 [cs] duration of the second formant energy.

Using the four masking regions for each sound, all 16 feature combinations

were generated and presented to the listeners. The condition in which no features were removed from the sound is the negative control, for which the zero-error condition has been confirmed. The condition in which all four features were removed can be considered a positive control for which high errors are expected. All stimuli wav files and AI grams are available publicly online at `http://hear.ai.uiuc.edu/public/F2Test_supplemental.zip`, as well as in Appendix A.

2.1.2 Listeners

In all, 24 normal-hearing listeners volunteered to take the test. All listeners were L1=English and, based on their speech scores at 12 dB **SNR**, had no observable hearing impairment. Since the speech stimuli were chosen based on their robustness and zero error in low noise, listeners were disqualified from the results if they reported any errors to the negative control (unmodified token) stimuli.

2.1.3 Presentation

For each presentation, listeners were asked to respond with one of the 18 consonants /b, d, f, g, h, k, n, m, p, s, t, v, w, y, z, θ (that), ð (think), ʃ (shoe)/, "vowel" if only a vowel was heard, and "other" if what they heard was not represented in the set of choices. If the listener chose the "other" option, they were prompted to enter the sound that they heard. Stimuli were presented in a quasi-random order; a random order was selected and then an iterative method was used to modify the order so that stimuli derived from the same utterance were greater than or equal to 10 presentations apart. Prior to the graded test, listeners were presented with 25 practice sounds drawn from the seeded sounds to familiarize them with the experimental proceedure. Listeners were tested in a soundproof booth and stimuli were presented binaurally through Sennheiser HD 280 Pro 64 Ω headphones.

Listeners were able to replay the stimuli twice, if needed. The user's Most Comfortable Level (MCL) was calibrated prior to taking the test to a level between 65 and 85 dB-SPL. Listeners were allowed to change the level at any point during the test and to take as much time as necessary to respond to a

given stimuli.

2.1.4 Data Analysis

The distributions of listener error over all utterances for each condition will be compared using a ANalysis Of VAriance (ANOVA) (Gaito, 1973) to find the variance accounted for by each of the masking conditions. Each of the distributions is compared to the unmodified token listener error, which is controlled to be zero error for all 25 utterances. Results of the ANOVA are given by the F statistic. Each comparison is made with two groups that contain 25 elements each, yielding $k - 1 = 2 - 1 = 1$ and $k(n - 1) = 2(25 - 1) = 48$ degrees of freedom. Thus, level of significance, p, will be taken from the F(1,48) distribution. The null hypothesis H_0 for the ANOVA is that the two groups poll the same distribution. An insignificant p value for a masking condition is interpreted as the feature having no significant perceptual value. A significant p value is interpreted as the feature having high perceptual value.

2.2 Results

Of the 24 listeners tested, 4 did not fulfill the control conditions. Thus these listeners were removed from the results so that each stimulus had 20 corresponding listener responses. On occasion, listeners chose the "other" button and included the semi-vowel /l/ with the standard consonant response. For example, in some situations some listeners chose to respond using /kl/ rather than the offered option /k/. Listeners were not instructed to respond in this manner so this type of response was inconsistant across listeners, therefore it is impossible to use the data in our analysis. As a result, when this type of response was given, the response was scored based on the first consonant (i.e., /k/ in our example).

The distribution response error across utterances for when each of the masked features N, B, O, and T is given in histograpm plots in Figure 2.3. The mean and standard deviation of each of these distributions is shown in Table 2.1, as well as the F(1,48) and p statistics based on the ANOVA.

The listener average error across all utterances for the O, T, N, and B

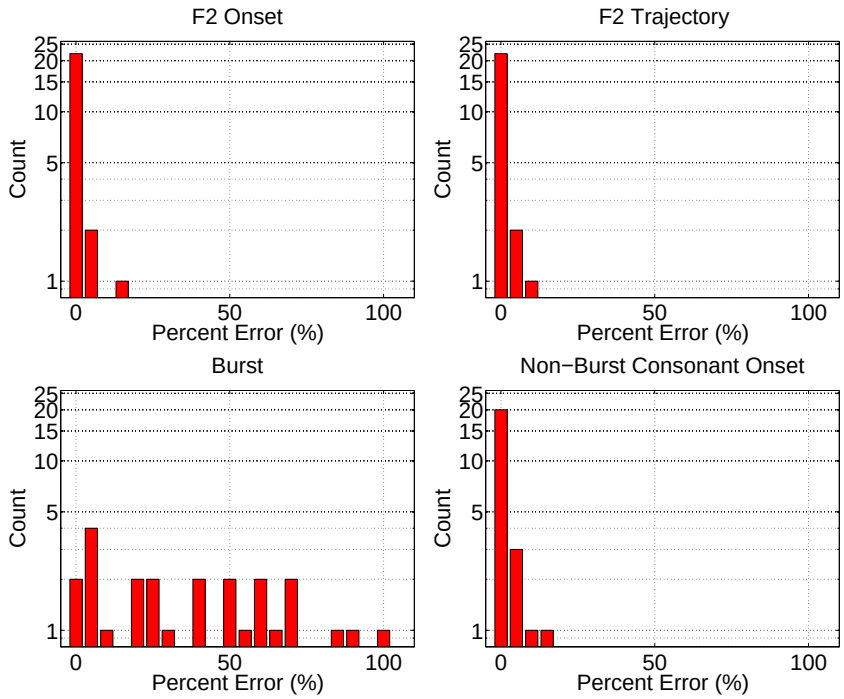

Figure 2.3: Histograms of errors over all 25 utterances when O (top left), T (top right), B (bottom left), and N (bottom right) were removed individually. For example, when O is removed, 22 utterances maintained zero error, two sounds had 5% error, and one sound had 10% error.

conditions were 1.0% ($F(2, 48) = 1.3$, $p = .269$), 0.8% ($F(1, 48) = 1.5$, $p = .227$), 1.6% ($F(1, 48) = 2.4$, $p = .129$), and 39.4% ($F(1, 48) = 21.9$, $p = 2.3 \times 10^{-5}$), respectively. Removing *both* O and T introduced mildly significant listener error of 3.4% on average ($F(1, 48) = 4.3$, $p = .040$), as compared to when no features were removed. The positive control mask in which all features were masked (NBOT) introduced highly significant 65.6% listener error ($F(1, 48) = 60.5$, $p = 4.7 \times 10^{-10}$) on average.

Response error for when each of the features was masked *for each individual utterance* are given in Figures 2.4–2.6 with the example AI grams for the utterance f113 /ga/. A more extensive analysis of the utterances shows that only utterances f104 /ka/ and f101 /kæ/ had greater than 15% error in the OT mask condition. Both utterances were with the consonant /k/. Of the 4 m104 /ka/ errors for this condition, 3 were the consonant /p/ and 1 was the consonant /t/. Of the 4 f101 /kæ/ errors, 2 were /t/, 1 was /p/, and one listener responded with "other-pk", which we interpret as either p or k.

Table 2.1: The mean and standard deviation of listener errors for each single-variable masking condition, as well as the F(1,48), p values for each distribution as compared to the unmodified signal listener errors using a one-way ANOVA (Gaito, 1973).

Condition	Mean Error (%)	StDev (%)	F	p
Unmodified	0	0	0	1
O	1.0	3.2	1.3	0.269
T	0.8	2.3	1.5	0.227
OT	3.4	5.9	4.3	0.04
N	1.6	3.7	2.4	0.129
B	39.4	30.4	21.9	2.3×10^{-05}
NBOT	65.6	30.4	60.5	4.7×10^{-10}

The result of removing B is given in Figure 2.5. 18 of the 25 utterances showed greater than 15% error when this feature was removed. Both of the utterances from the CVs /gæ/ and /gɪ/ showed less than 15% error, as well as f108 /du/, f103 /kɪ/, and m120 /ku/. The result of removing N is given in Figure 2.6. All utterances had less than or equal to 15% error in this condition.

The data *prove* that the B feature is the only significant contributor to the perception of natural speech of the four features that were tested. When each of the other features N, O, and T is removed, the listeners were still able to recognize the tokens with zero or near-zero error. Both conclusions seem highly significant.

2.2.1 Extended Feature Analysis

If the primary feature, the Burst, is removed from the speech signal, it is possible to evaluate whether any of the remaining features *becomes* relevant to perception. It is assumed that when the primary feature has been removed, the perceptual task relies on the remaining features within the signal to decide the speech token that was presented. Table 2.2 shows the statistics of the listener error distributions during the conditions where the burst has already been removed. Note that none of the features becomes significant for all utterances after the burst has already been removed. As a result, the

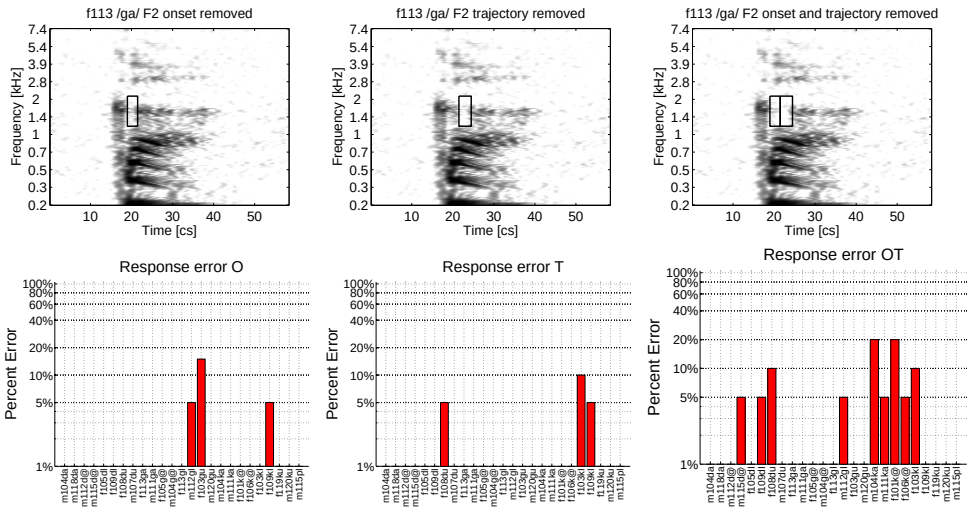

Figure 2.4: The AI grams of f113 /ga/ and bar plots of response errors across utterances when the O (left), T (center), and *both* O and T (right) were removed. For most sounds there was no significant response error. There was mildly significant error when the O and T were both removed for the consonant /k/.

response error, confusions, and entropy are reviewed for each utterance in Table C.1 in Appendix C when features are further removed from the signal.

Table 2.3 the confusion matrix for the conditions B, NB, BO, and BT summed across utterances for each consonant /d, g, k, p/. For /d/, removing B places the listeners in a state of confusion between /d/ and /g/. When either N or O is removed in addition to B, the sound becomes less ambiguous, while removing T makes the sound more often confused. Removing B from /g/ creates listener confusions between /g,d/. When N is removed in addition to B, listener error drops by 10% and the confusion group becomes more diverse by including /y,θ/. Removing both O and T in addition to B increases the listener confusions with /d/. Removing B from /k/ creates a confusion group with /t,p/. Removing either O or T in addition to removing B causes no significant change in listener errors or confusions. Removing N in addition to B causes listeners to no longer confuse the sound with /t/ but instead with /h,p/. For /p/, listener responses are similarly confused with /f/ for B, BO, and BT. NB increases listener error by 65% and causes listeners to respond with /d,ð,h/.

This analysis shows that each of the features affects listener responses

Table 2.2: The distribution of errors for the conditions for which the burst was already removed. The F(1,48) and p values indicate the ANOVA statistics for each of the conditions compared to the burst-removed condition. Note that none of the features becomes significant for all utterances after the burst has already been removed.

Condition	Mean Error (%)	StDev (%)	F	p
B	39.40	30.36	1.61×10^{-30}	1.00
BO	40.00	27.73	0.00	0.96
BT	42.60	28.87	0.08	0.78
NB	41.40	30.33	0.03	0.87

Table 2.3: The confusion matrix showing the listener confusions after B has already been removed. All of the utterances for each consonant /g,d,k,p/ were summed for each condition.

	b	d	f	g	h	k	m	n	p	s	t	v	w	y	z	θ	ð	ʃ	other	vowel	Total
d_B	4	87	0	42	0	0	0	0	0	0	2	4	0	0	0	17	4	0	0	0	160
d_{NB}	4	106	0	22	0	0	0	1	0	0	0	5	0	0	0	15	5	0	0	2	160
d_{BO}	3	103	0	34	2	0	0	1	0	0	0	1	0	1	0	9	1	0	4	1	160
d_{BT}	4	68	0	60	0	0	0	0	0	0	0	8	0	1	0	16	3	0	0	0	160
g_B	0	40	0	112	0	0	0	0	0	0	1	0	0	0	0	7	0	0	0	0	160
g_{NB}	2	10	0	127	0	0	0	0	1	0	0	0	0	8	0	8	0	0	0	4	160
g_{BO}	0	51	0	103	0	0	0	0	0	0	0	0	0	0	0	3	3	0	0	0	160
g_{BT}	1	36	0	116	0	0	0	0	0	0	0	0	0	0	0	5	2	0	0	0	160
k_B	0	0	0	0	0	89	0	0	16	0	54	0	0	0	0	0	1	0	0	0	160
k_{NB}	0	0	1	0	68	58	0	0	27	0	0	0	5	0	0	0	0	0	1	0	160
k_{BO}	0	0	0	0	1	79	0	0	13	0	64	0	0	0	0	0	0	0	3	0	160
k_{BT}	0	0	0	0	1	89	1	0	19	0	49	0	0	0	0	0	0	0	0	0	160
p_B	0	0	4	0	0	0	0	0	15	0	1	0	0	0	0	0	0	0	0	0	20
p_{NB}	0	9	0	0	3	0	0	0	2	0	0	0	0	0	0	1	5	0	0	0	20
p_{BO}	0	0	1	0	1	1	0	0	15	0	1	0	0	0	0	0	1	0	0	0	20
p_{BT}	0	0	2	0	2	1	0	0	14	0	0	0	0	0	0	0	1	0	0	0	20

in a consonant-dependent manner. It is important that no one feature now completely directs listener perception; it reinforces the fact that the B feature is the necessary feature for listeners to accurately and unambiguously be able to identify the target sound.

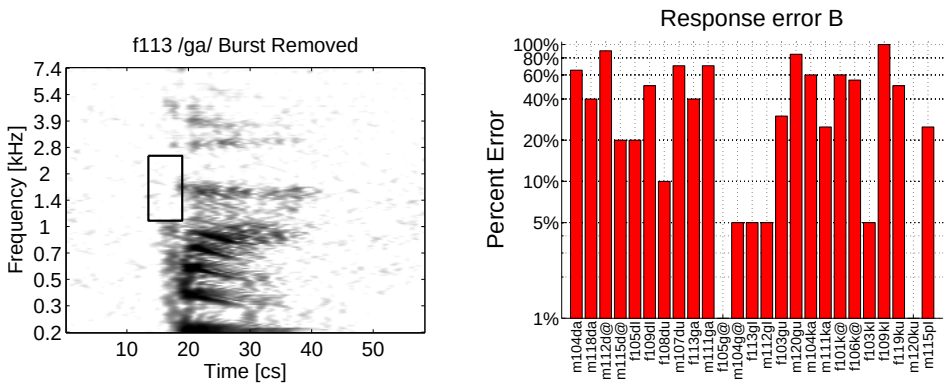

Figure 2.5: The AI gram of f113 /ga/ (left) and the bar plot of listener errors across utterances when B was removed (right). 18 of the 25 sounds had a significant difference in response error when the burst was removed. Mean error for this condition was 39% with a standard deviation of 30% and entropy of 3.76 bits.

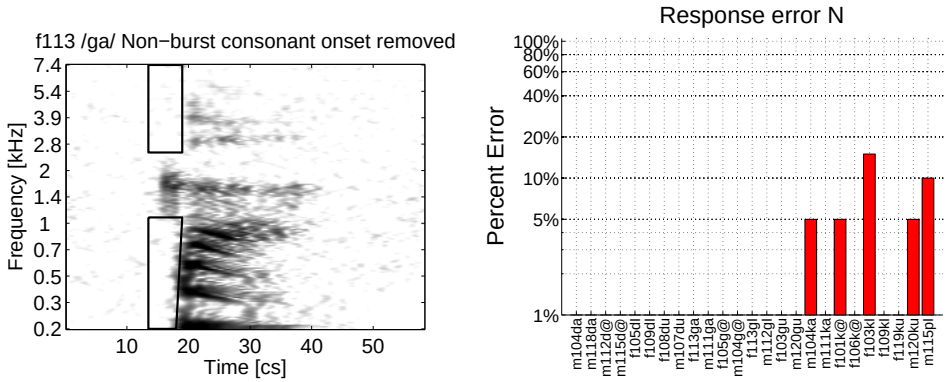

Figure 2.6: Bar plot of errors across utterances when N was removed. For all utterances, there was no significant response error. Of the sounds that had error, even if not significant, 4 of the 5 were utterances with the consonant /k/.

CHAPTER 3

CONCLUSION

Based on the results, the significance of the F2 trajectory feature as a perceptual cue in natural speech is not confirmed, as was first suggested by Stevens and Blumstein (1978); Blumstein and Stevens (1979, 1980). There is no significant change in consonant error when this feature is removed; listeners are able to identify a consonant correctly dispite there being *no* F2 trajectory.

Past studies have suggested that speech cues are redundant and that, when the F2 trajectory is removed, other acoustic features may be used to trigger a perceptual event in a listener. Such a theory of redundant cues does not explain why significant error was observed only when the burst feature is removed. If speech information is truly redundant, perceptual scores would have been unaffected when removing the burst.

Singh and Allen (2011) and Kapoor and Allen (2011) provide further evidence for *one* acoustic feature that is perceptually necessary and sufficient for perception. Singh and Allen (2011) analyzed consonant errors across **SNR** in speech-weighted noise. They found that for naturally spoken CV utterances there is zero error in quiet, and that error approaches chance performance over a 6 dB change in **SNR**. This threshold varied per utterance; the ensemble average of all the utterances is consistent with the Articulation Index model of speech perception. Their result implies that error increased when just one critical acoustic feature was masked by noise. Kapoor and Allen (2011) were able to shift the threshold by adding gain to very specific parts of the acoustic wave form. By boosting and attenuating the acoustic burst (whose definition is similar to that of the current study) by 6 dB, they were able to shift the threshold by a similar amount. Given the steep score transition (6 dB), this result strongly suggests, and fully supported by the results present here, that there is just one critical feature, the burst.

Further, given that there is only *one* perceptual cue, we observe the results of the ANOVA analysis to evaluate the candidates. Table 2.1 shows that the

only features that provided a significant change in listener errors were OT and B. Since the significance level for the feature masks B and OT were $p = 2.3 \times 10^{-5}$ and $p = 0.04$, respectively, and there is only one relevant acoustic feature, we conclude that the burst is the only feature that is relevent to the perception of the consonant. These results of Singh and Allen (2011) and Kapoor and Allen (2011), combined with the current observations, prove that speech cues are not redundant and that the F2 trajectory does not directly cue speech perception as either a primary or secondary mode of information. Instead, the consonant resonant burst frequency acts as the sole primary cue.

Results from others, such as Fogerty and Kewly-Port (2009), are interpreted as providing evidence for the F2 transition as a perceptual cue in natural speech. They used their result as evidence for the traditional cues from Liberman *et al.* (1967); however, there is a significant difference between the two methods. Liberman *et al.* (1967) used data from Cooper *et al.* (1952); Liberman *et al.* (1954); Liberman (1957) to outline how speech works as a code rather than a cipher. Liberman *et al.* (1967) developed the theory based on *consonant recognition scores in CVs* rather than *word recognition scores in sentences*. The added context within the word and the sentence lowers the entropy of the task, making it easier for listeners to identify individual phonemes. If the goal is to make a statement about the *consonant* cues, as is their implied goal, then recognition scores should be based on correct consonant identification. As a result, we believe that Fogerty and Kewley-Port (2009) is an inadequate verification of the transition information in natural speech.

The result is coherent with evidence from Stevens and Blumstein (1978); Blumstein and Stevens (1979, 1980); Dubno *et al.* (1987) that states that the cue for place of articulation information is the onset spectra of the consonant release rather than the transition energy. Further, this result confirms the work of Hazan and Simpson (1998); Li and Allen (2011); Li *et al.* (2010); Kapoor and Allen (2011) that place of articulation information is located in the resonant burst of the consonantal release.

These results presents serious challenges on all theories of speech perception that place the F2 trajectory in a prominant position, namely those that are based on the concept of coarticulation and the gestural transmission of cues, such as Motor Theory and Direct Realism. The concept of coarticula-

tion was developed from evidence from Cooper *et al.* (1952) and Liberman *et al.* (1954). The result of this experiment calls into question the results from Liberman *et al.* (1954), where it was observed that two different-sloped F2 trajectories cued one sound, since, within the context of natural speech, the F2 trajectory *does not function as a cue at all.* Although coarticulation is largely based on the evidence from Liberman *et al.* (1954), the results from this paper also raise the need to verify the results of Cooper *et al.* (1952), using natural speech.

APPENDIX A

STIMULI MASKING REGIONS

Figure A.1 shows the labeled feature regions for each of the 25 utterances used in the experiment.

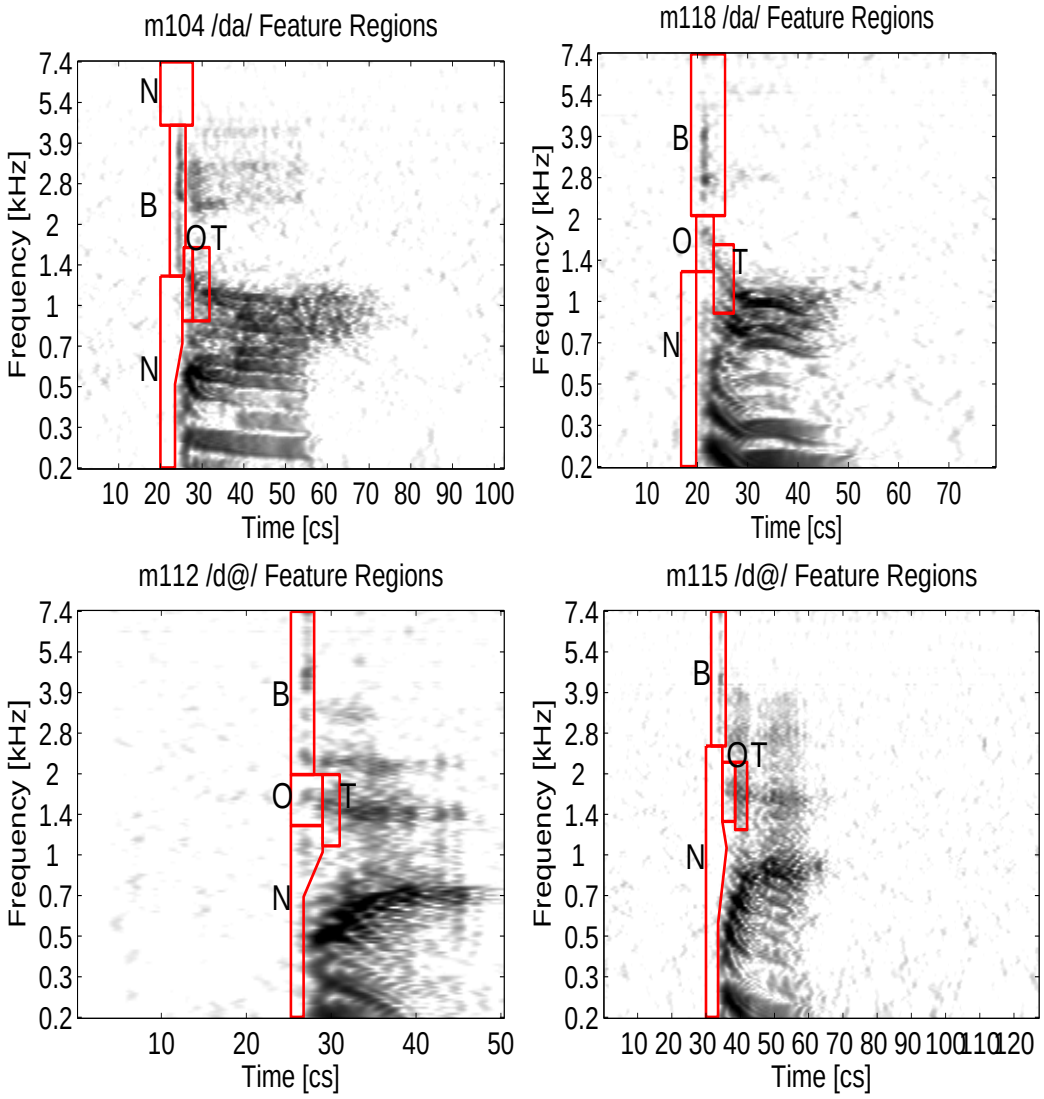

Figure A.1: The labeled feature regions for each of the 25 utterances used in the experiment.

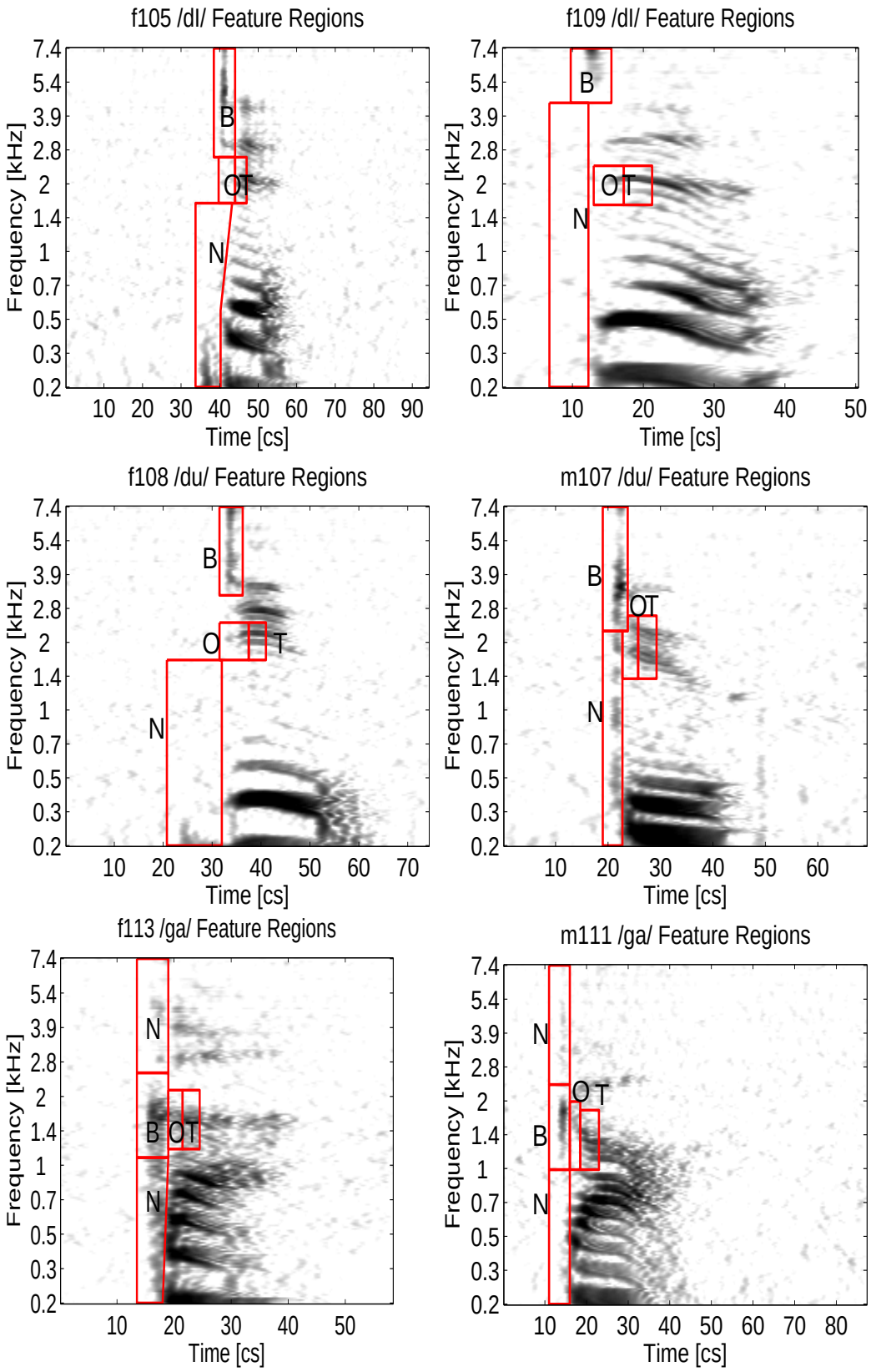

Figure A.1: (continued) The labeled feature regions for each of the 25 utterances used in the experiment.

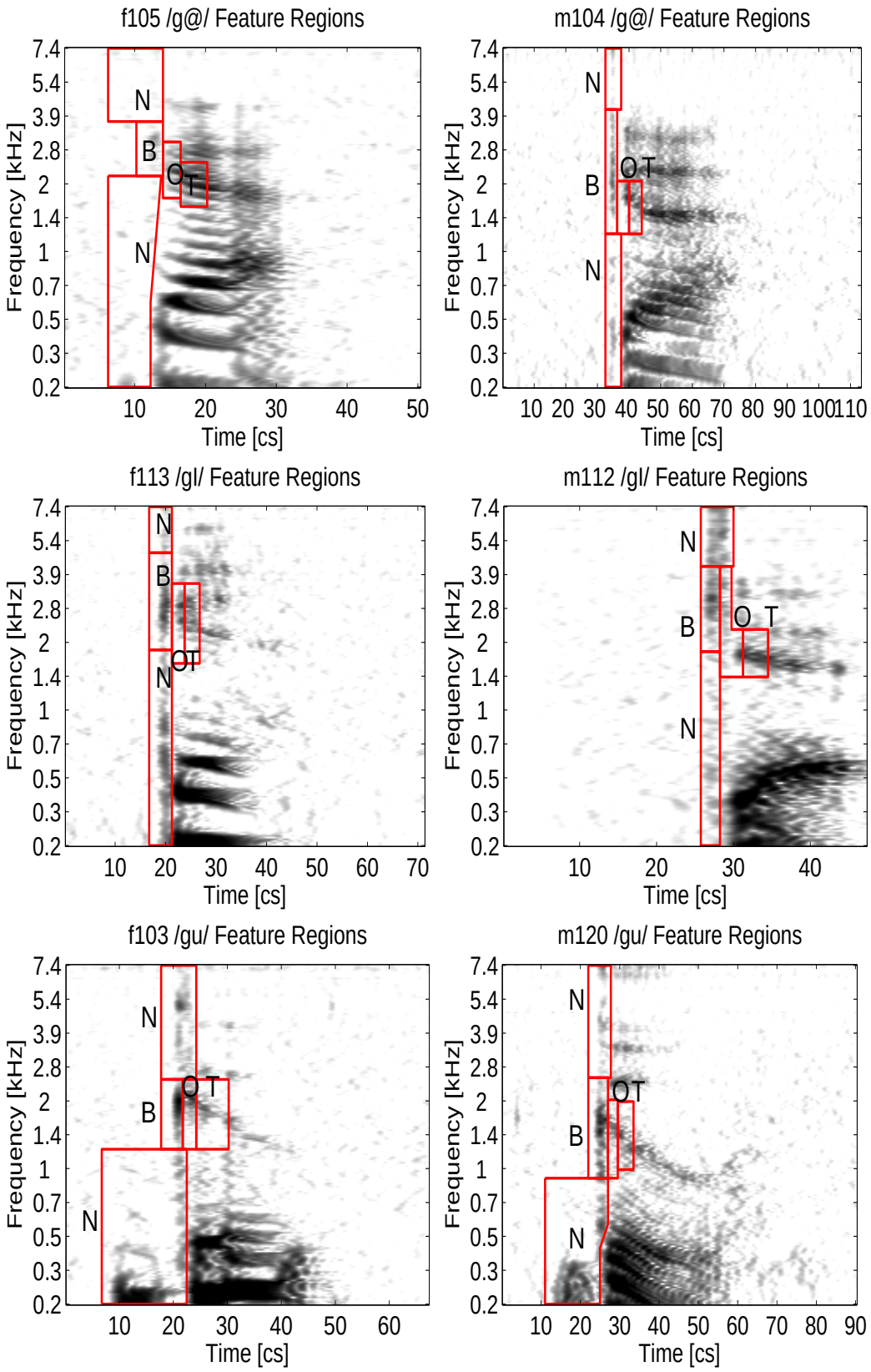

Figure A.1: (continued) The labeled feature regions for each of the 25 utterances used in the experiment.

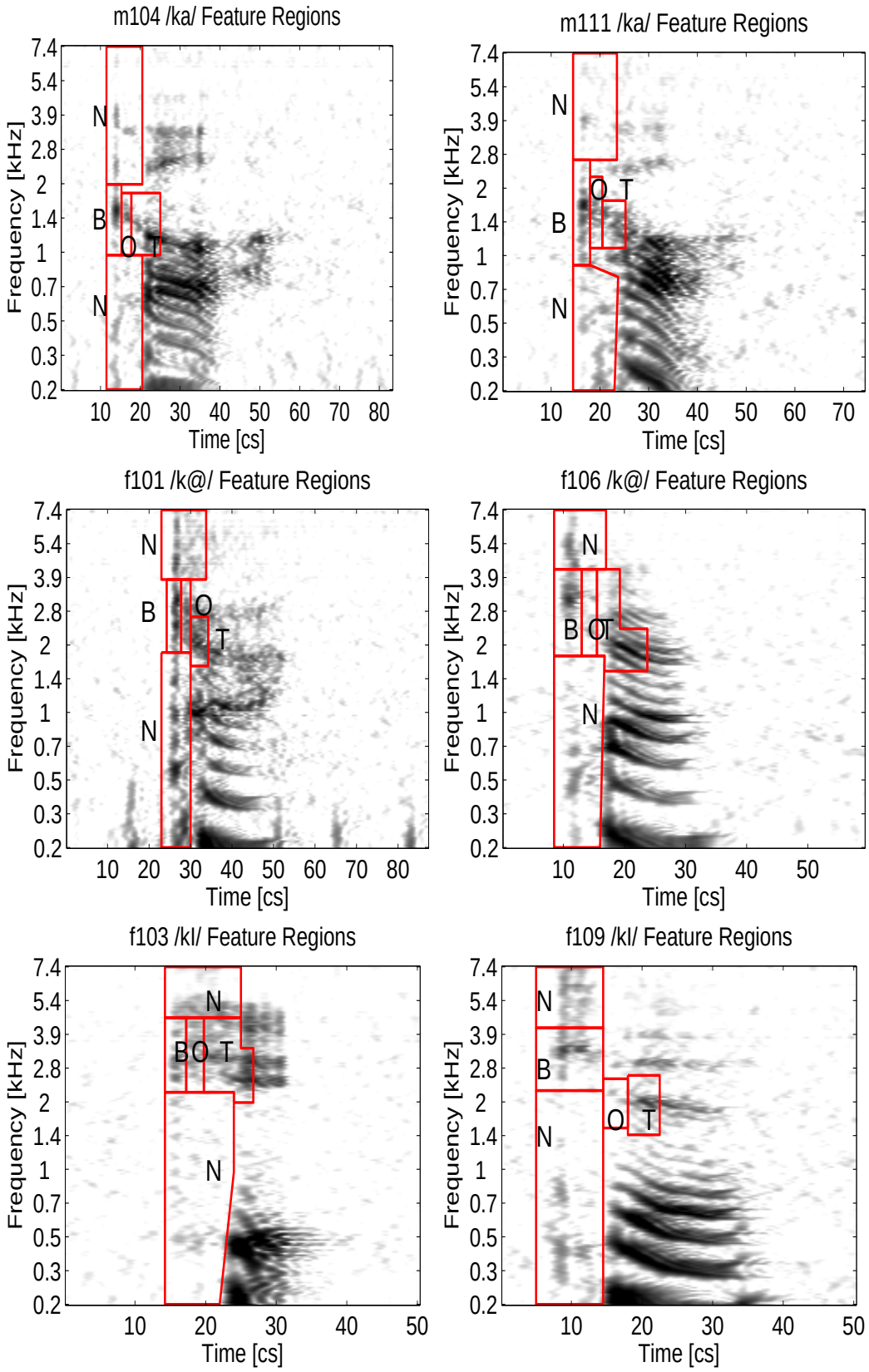

Figure A.1: (continued) The labeled feature regions for each of the 25 utterances used in the experiment.

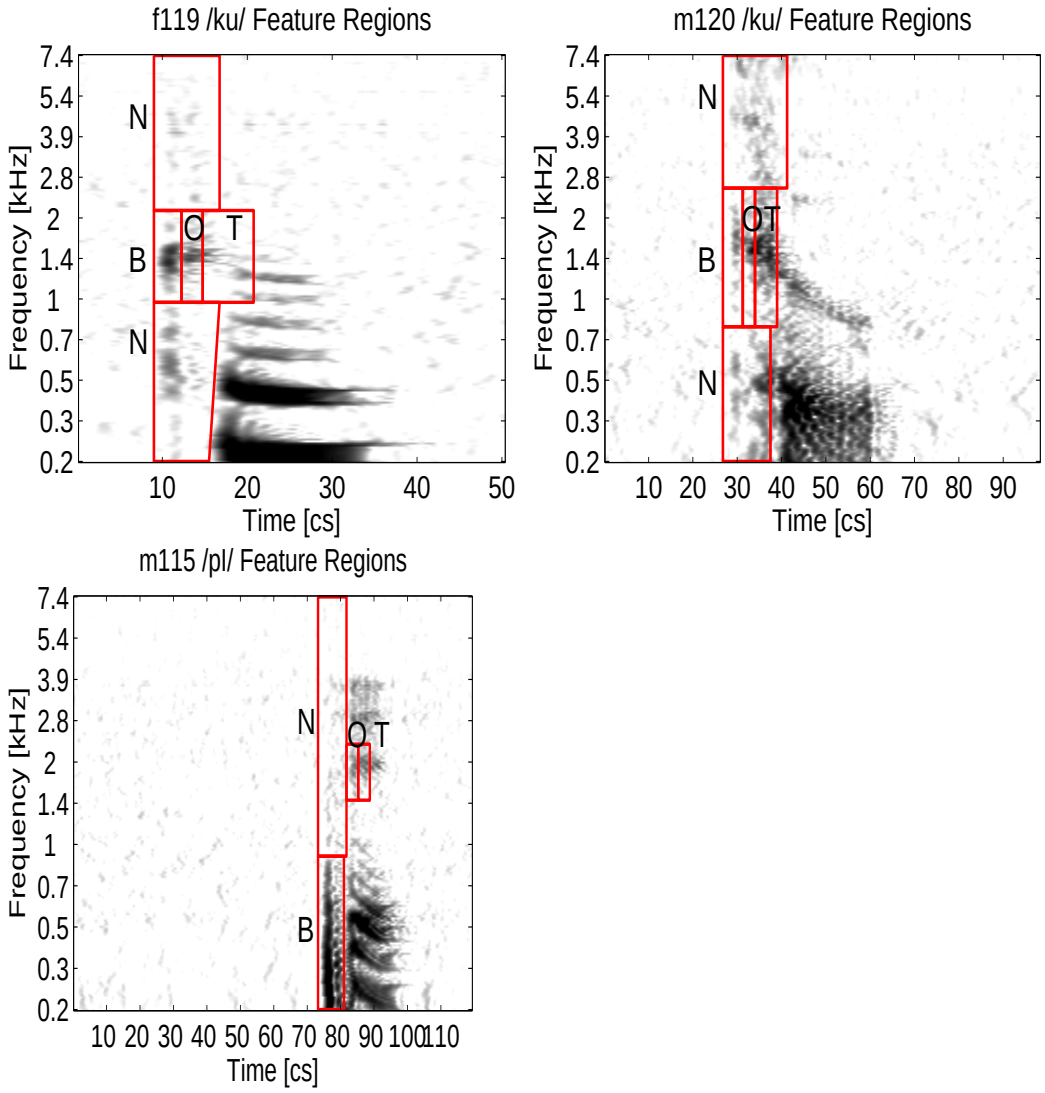

Figure A.1: (continued) The labeled feature regions for each of the 25 utterances used in the experiment.

APPENDIX B

EXPERIMENT DATA: SINGLE-FEATURE MASKS

Figure B.1 shows the results of each individual utterance for each masking condition.

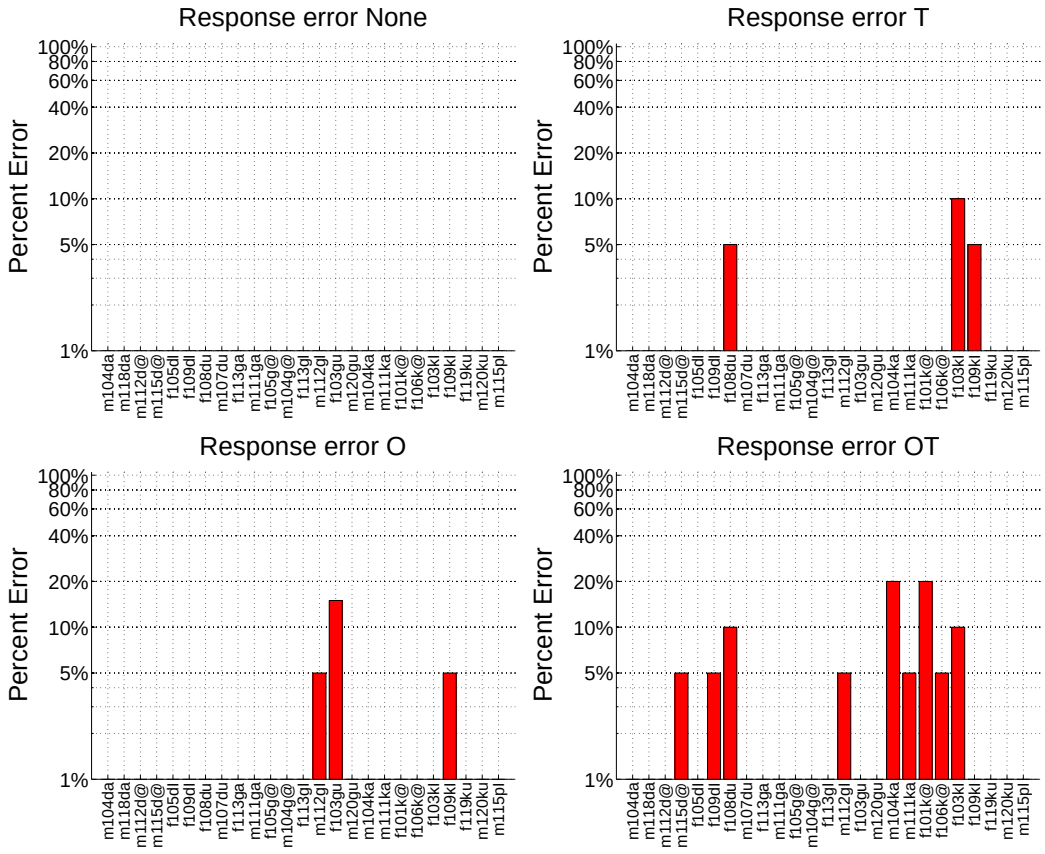

Figure B.1: The results of each individual utterance for each masking condition.

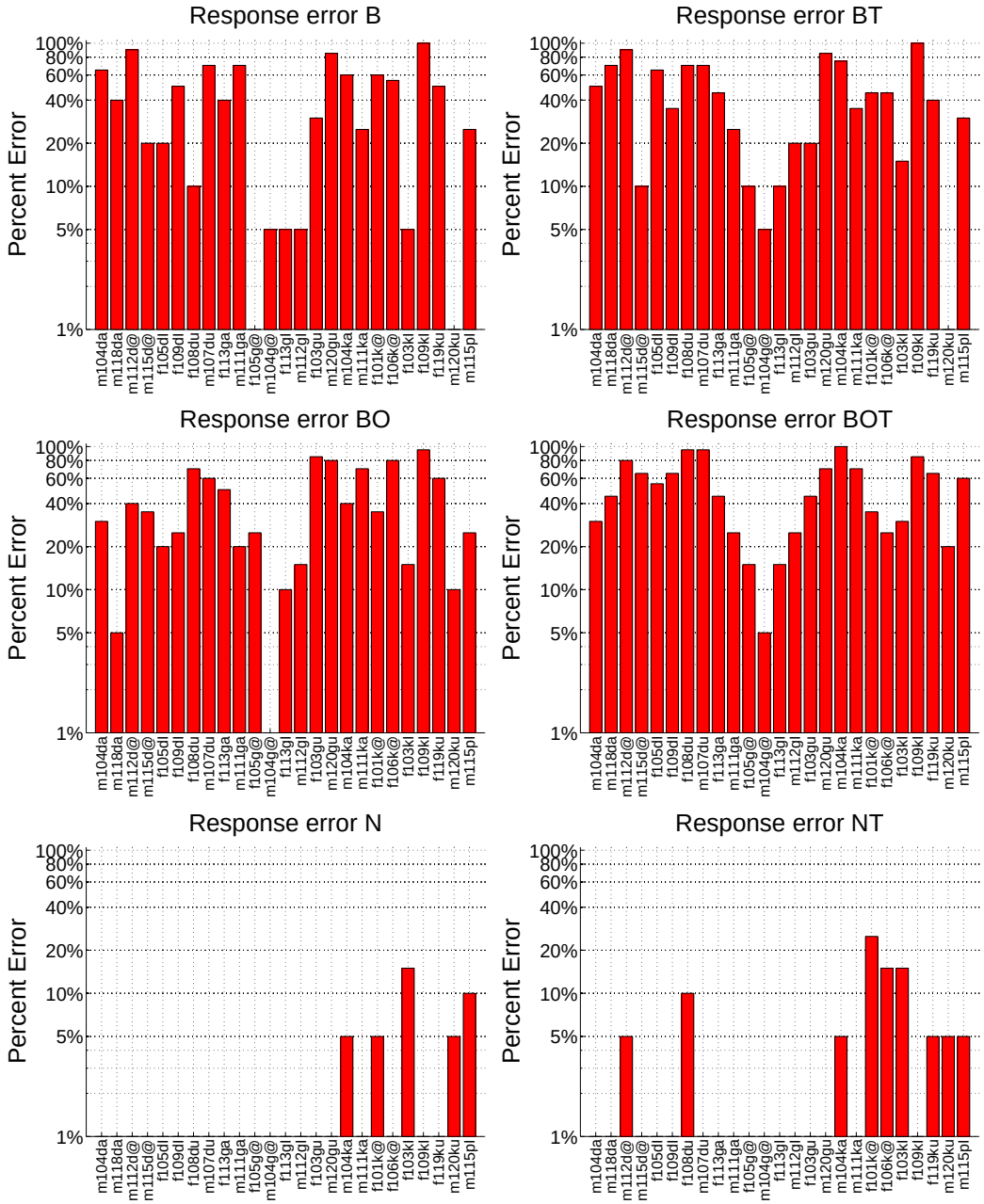

Figure B.1: (continued) The results of each individual utterance for each masking condition.

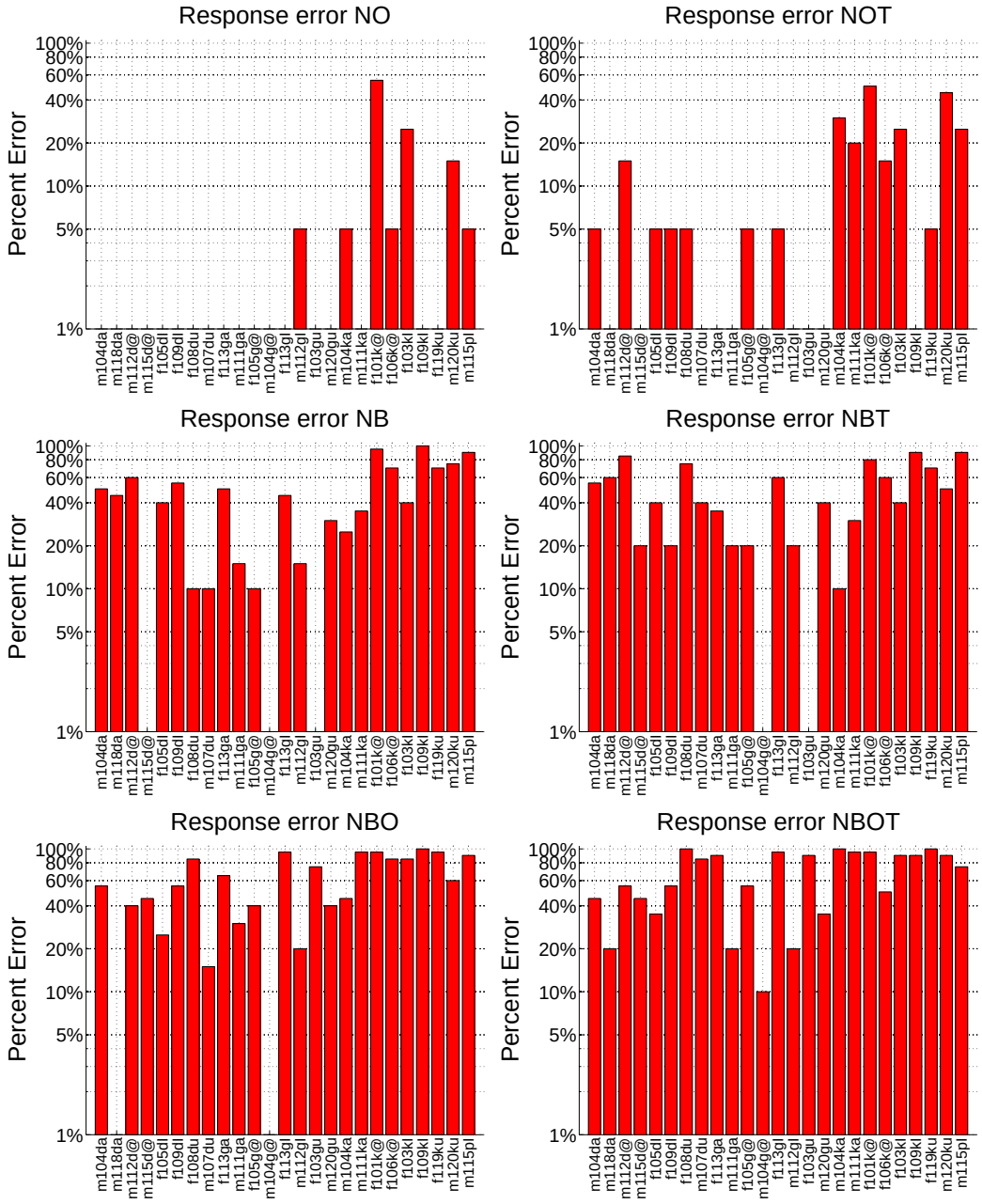

Figure B.1: (continued) The results of each individual utterance for each masking condition.

APPENDIX C

EXTENDED DATA ANALYSIS: BURST-REMOVED CONDITIONS

Table C.1 shows the response error, entropy, and confusions for each of the utterances when the Burst (B) was removed, when the consonantal release was removed (NB), when the Burst and F2 onset (BO) were removed, and when the Burst and the F2 Trajectory (BT) were removed. Each of the metrics can be analyzed for each the NB, BO, and BT conditions against the B reference. Significant metric changes from the B condition are highlighted by a column dependent color, with more significant changes being represented by darker colors.

Table C.1: The response error, entropy, and confusions for each of the utterances when the Burst (B) was removed, when the consonantal release was removed (NB), when the Burst and F2 onset (BO) were removed, and when the Burst and the F2 Trajectory (BT) were removed.

Mask Condition	B			NB			BO			BT		
Utterance	P_e	H	Confusions	P_e	H	Confusions	P_e	H	Confusions	P_e	H	Confusions
m104 /da/	0.650	1.578	/th,TH,g/	0.500	1.578	/th,TH,g/	0.300	1.319	/th,g,TH/	0.500	1.578	/th,g,TH/
m118 /da/	0.400	0.971	/g/	0.450	0.993	/g/	0.050	0.286	/th/	0.700	0.881	/g/
m112 /d@/	0.900	1.022	/g,b,t/	0.600	1.361	/g,th/	0.400	1.353	/g,th/	0.900	0.922	/g,b/
m115 /d@/	0.200	1.022	/g,b,t/	0.000	0	–	0.350	0.934	/g/	0.100	0.469	/th/
f105 /dI/	0.200	1.022	/th,g,TH/	0.400	1.871	/b,g,TH,v,th/	0.200	1.022	/g,b,th/	0.650	1.739	/g,th,TH/
f109 /dI/	0.500	1.880	/th,b,v,TH/	0.550	1.977	/th,v,b,TH/	0.250	1.154	/g,b,other/	0.350	1.579	/v,g,th,TH/
f108 /du/	0.100	0.469	/v/	0.100	0.569	/n,v/	0.700	2.809	/g,other,h,n,y,v,th,vowel/	0.700	2.285	/g,v,th,b,y/
m107 /du/	0.700	0.881	/g/	0.100	0.469	/vowel/	0.600	1.219	/g,b/	0.700	1.141	/g,b/
f113 /ga/	0.400	1.188	/th,d/	0.500	1.578	/th,d,b/	0.500	1.743	/d,th,TH/	0.450	1.601	/th,d,TH/
m111 /ga/	0.700	0.881	/d/	0.150	0.610	/d/	0.200	0.722	/d/	0.250	0.811	/d/
f105 /g@/	0.000	0	–	0.100	0.569	/d,y/	0.250	0.811	/d/	0.100	0.469	/d/
m104 /g@/	0.050	0.286	/d/	0.000	0	–	0.000	0	–	0.050	0.286	/d/
f113 /gI/	0.050	0.286	/d/	0.450	1.437	/y,b,vowel/	0.100	0.469	/d/	0.100	0.569	/b,d/
m112 /gI/	0.050	0.286	/d/	0.150	0.610	/d/	0.150	0.610	/d/	0.200	0.722	/d/
f103 /gu/	0.300	0.881	/d/	0.000	0	–	0.850	0.610	/d/	0.200	0.884	/d,TH/
m120 /gu/	0.850	0.884	/d,t/	0.300	1.419	/vowel,d,p,th/	0.800	0.992	/d,TH/	0.850	0.610	/d/
m104 /ka/	0.600	1.559	/t,p/	0.250	0.811	/h/	0.400	1.353	/t,p/	0.750	1.559	/t,p/
m111 /ka/	0.250	0.811	/t/	0.350	0.934	/h/	0.700	1.141	/t,other/	0.350	1.141	/t,p/
f101 /k@/	0.600	1.219	/t,p/	0.950	1.188	/h,p/	0.350	0.934	/t/	0.450	1.437	/t,m,p/
f106 /k@/	0.550	0.993	/t/	0.700	1.295	/h,p/	0.800	0.722	/t/	0.450	0.982	/t/
f103 /kI/	0.050	0.286	/TH/	0.400	1.188	/h,other/	0.150	0.848	/h,t,other/	0.150	0.748	/p,h/
f109 /kI/	1.000	0	/t/	1.000	0.971	/h,p/	0.950	0.286	/t/	1.000	0	/t/
f119 /ku/	0.500	1.000	/p/	0.700	1.895	/h,p,w/	0.600	1.361	/p,t/	0.400	0.971	/p/
m120 /ku/	0.000	0	–	0.750	2.121	/h,p,w,f/	0.100	0.569	/t,other/	0.000	0	–
m115 /pI/	0.250	0.992	/f,t/	0.900	1.977	/d,TH,h,th/	0.250	1.392	/f,h,k,t,TH/	0.300	1.457	/f,h,k,TH/
Mean	0.394	0.816		0.414	1.097		0.400	0.986		0.426	0.994	

APPENDIX D

SIGNAL-TO-NOISE RATIO CALCULATIONS

All **SNR** calculations are in dB using the equation

$$\mathbf{SNR}|_{dB} \quad = \quad 20\log_{10}\left(\frac{\sigma_S}{\sigma_N}\right) \tag{D.1}$$

where σ_S and σ_N are the signal strength of the speech signal and the noise, respectively. Several methods were considered to determine the signal and/or noise level. The peak value is a poor metric for signal strength since one outlying peak might cause a mischaracterization of the strength of the rest of the signal.

Instead the root mean square (RMS) can be used to characterize the signal strength. One problem with using the RMS of an entire duration of a signal is that it is not robust to zero padding; adding zeros to the signal will lower the RMS considerably. Instead, a moving window can be used to find a local RMS throughout the signal. Conceptually, a mean is a rectangular window that is normalized to unity energy, which has the frequency response equivalent to a low-pass filter. If, instead of applying the crude rectangular window to implement the low-pass filter, we design a filter with better low-pass frequency response, we can achieve a more accurate RMS calculation.

An exponential filter has significally better frequency response properties (lower side lobes, sharper cutoff, etc.) and can offer a better mean calculation. In designing a filter to determine the RMS, it is important that its impulse response is physiologically based rather than arbitrarily chosen. Munson (1947) loudness vs. time data note that it takes roughly 20 [cs] for loudness to reach its full value for pure tones in humans. Thus, using that duration as a time constant for the exponential mean filter is a reliable way to represent human sensation. Dunn and White (1940) used an exponential filter with $\tau = 12.5$ $[cs]$ to match the average syllable length.

Another option when determining signal strength is to use the ANSI spec-

ification for the VU, which has been widely used in the past. According to Lobdell and Allen (2007), the ANSI specification for the VU corresponds to a full-wave rectification followed by a low-pass filter with an impulse response that yields a step response with a 30 [cs] rise time and less than 1.5% overshoot. The result is then converted from [vu] units to decibels, which is designated in [VU] units.

The goal is to use a method that is physiologically based and can be consistently translated into measurements made by other signal strength algorithms. To investigate the effect of using various signal strength determination methods in the context of natural speech, several of these methods were applied to the entire syllable corpus of the LDC database (phrases were not used in this analysis). The signal strength of each CV was first normalized to 1 using the 20 [cs] exponential RMS metric, and then the signal strength was measured by each of the other methods. The methods used to measure the signal strength are the moving rms using exponential filters with $\tau = 20$ $[cs]$ and $\tau = 12.5$ $[cs]$, a moving rms using rectangular filters of length 20 $[cs]$ and 12.5 $[cs]$, the full signal-peak, and the full-signal RMS. All filters were normalized to have an energy of 1.

The exponential filters hold the form

$$H(z) = \frac{\alpha}{s + \alpha - 1} m \tag{D.2}$$

where α is chosen to acheive the appropriate time constants. In this case, $\alpha = 3.126e-4$ and $\alpha = 5.0e-4$ were used to set the filter to $\tau = 20$ $[cs]$ and $\tau = 12.5$ $[cs]$, respectively. The signal was squared, the filter was applied, and the square root of the result was taken. The max over time of the resultant vector was used as the metric for signal strength.

40292 tokens were analyzed. The average length of the sounds was 91.4 [cs]. The distributions for each of the level calculations is shown in Figure D.1. Note that, given the 20 [cs] exponential RMS as reference, there is a mean level shift from unity and non-zero variance introduced in each measurement. The statistics of each of the methods is given in Table D.1. Note that the large variance using the Peak method verifies the poor quality of this method.

It is important to understand how these various level calculation methods differ in a quantitative way. For example, if it is assumed that each method will output the same level value for stationary noise, the distributions can

Table D.1: The statistics for each of the level calculation methods.

Method	Mean Level	StDev	Min	Max
exp20 RMS	1.0000	0.0000	1.0000	1.0000
exp12.5 RMS	1.1566	0.0271	1.0482	1.2564
rect20 RMS	1.4124	0.0678	1.0576	1.5754
rect12.5 RMS	1.2109	0.0542	0.9490	1.3596
vu	1.1604	0.1151	0.5194	1.4983
Full Length RMS	0.6279	0.0984	0.3369	1.0654
Peak	5.8220	1.7749	1.6639	22.3990

be used to estimate changes in **SNR** given a conversion from one method to another. Table D.2 needed to convert from one level **SNR** calculation method to another in decibels. If **SNR** calculations were originally made using a 20 [cs] exponential filter, to estimate the **SNR** using the full-length RMS method you could simply subtract 4 dB. Note that the quality of these conversion factors is dependent on the variance of the distributions in Figure D.1.

Table D.2: The factors needed to convert from one level method to another. Each factor is the ratio of the means of two distributions and represented it in decibels. For example, if one wanted to convert from 12 dB **SNR** based on the 20 [cs] exponential filter to an **SNR** based on the full length RMS, subtract 4 dB from the input **SNR**, i.e. the output **SNR** would be 8 dB. The reliability of these conversion factors depends on the variance of each of these distributions.

	exp20 RMS	exp12.5 RMS	rect20 RMS	rect12.5 RMS	vu	Full Length RMS	Peak
exp20 RMS	0.0000	1.2635	2.9991	1.6624	1.2919	-4.0421	15.3014
exp12.5 RMS	-1.2635	0.0000	1.7356	0.3988	0.0283	-5.3056	14.0378
rect20 RMS	-2.9991	-1.7356	0.0000	-1.3367	-1.7072	-7.0412	12.3023
rect12.5 RMS	-1.6624	-0.3988	1.3367	0.0000	-0.3705	-5.7044	13.6390
vu	-1.2919	-0.0283	1.7072	0.3705	0.0000	-5.3339	14.0095
Full Length RMS	4.0421	5.3056	7.0412	5.7044	5.3339	0.0000	19.3435
Peak	-15.3014	-14.0378	-12.3023	-13.6390	-14.0095	-19.3435	0.0000

Overall, the RMS methods that use a moving window are most consistent with the other methods; these methods have the smallest variance in the observed distributions, and as a result it is resonable to estimate the level given other methods by just multiplying the signal by a scalar. The full-length RMS can also be estimated from a moving window method, however

the estimate will be less accurate.

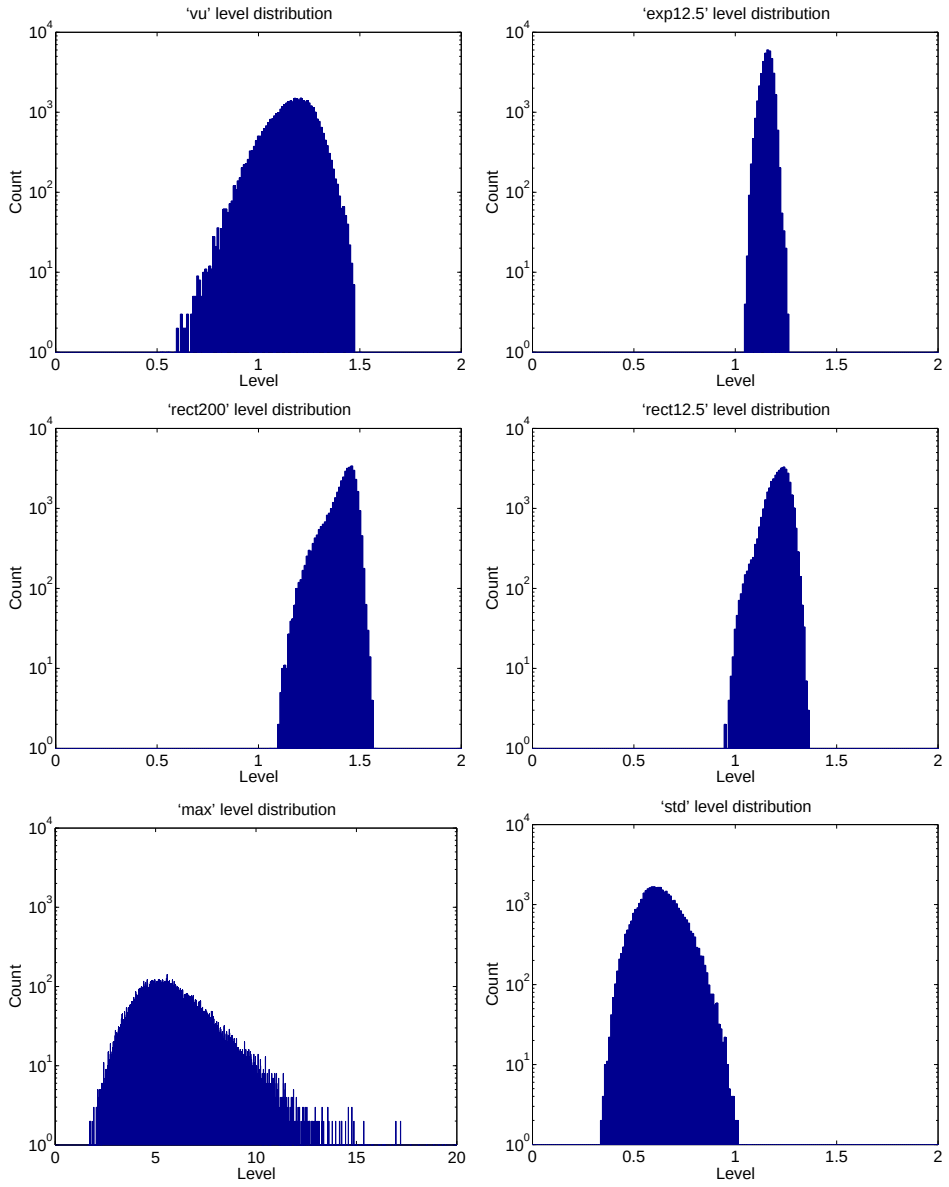

Figure D.1: The distributions of each of the signal-level calculation methods with reference to the level set by the RMS using a 20 [cs] exponential mean. Top left: vu, Top right: RMS exponential mean 12.5 [cs], Center left: RMS rectangular mean 20 [cs], Center right: RMS rectangular mean 12.5 [cs] Bottom left: Peak, Bottom right: Full-duration RMS. The VU is represented in lowercase [vu] units rather than decibel [VU] units. Also, the Peak method has such a large mean and variance that the horizontal axis needed to be rescaled.

REFERENCES

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

Allen, J. B. (**2005**). "Consonant recognition and the articulation index," J. Acoust. Soc. Am. **117**, 2212–2223.

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

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

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

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

Cole, R. and Scott, B. (**1974**). "Toward a Theory of Speech Perception," Psychol. Review **81**, 348–374.

Cooke, M. A. (**2003**). "Glimpsing speech," Journal of Phonetics **31**, 579–584.

Cooper, F., Delattre, P., Liberman, A., Borst, J., and Gerstman, L. (**1952**). "Some experiments on the perception of synthetic speech sounds," J. Acoust. Soc. Am. **24**, 597–606, haskins work on painted speech.

Diehl, R., Lotto, A., and Holt, L. (**2004**). "Speech perception," Annu. Rev. Psychol. **55**, 149–179.

Dubno, J. R. (**1978**). "Predicting consonant confusions in noise on the basis of acoustic analysis," Ph.D. thesis, City University of New York, doctoral dissertation.

Dubno, J. R., Dirks, D., and Schaefer, A. (**1987**). "Effects of hearing loss on utilization of short–duration spectral cues in stop consonant recognition," J. Acoust. Soc. Am. **81**, 1940–1947.

Dubno, J. R. and Levitt, H. (**1981**). "Predicting consonant confusions from acoustic analysis," J. Acoust. Soc. Am. **69**, 249–261.

Dunn, H. K. and White, S. D. (**1940**). "Statistical measurements on conversational speech," J. Acoust. Soc. Am. **11**, 278–288.

Eimas, P. D. and Corbit, J. D. (**1973**). "Selective adaptation of linguistic feature detectors," Cognitive Psychology **4**, 99–109.

Fletcher, H. and Galt, R. (**1950**). "Perception of speech and its relation to telephony," J. Acoust. Soc. Am. **22**, 89–151.

Fogerty, D. and Kewley-Port, D. (**2009**). "Perceptual contributions of the consonant-vowel boundary to sentence intelligibility," J. Acoust. Soc. Am. **126**, 847–857.

Fowler, C. A. (**1986**). "An event approach to the study of speech perception from a direct-realist perspective," Journal of Phonetics **14**, 3–28.

Fowler, C. A. (**1996**). "Listeners do hear sounds, not tongues," J. Acoust. Soc. Am. **99**, 1730–1741.

French, N. R. and Steinberg, J. C. (**1947**). "Factors governing the intelligibility of speech sounds," J. Acoust. Soc. Am. **19**, 90–119.

Gaito, J. (**1973**). *Introduction to analysis of variance procedures* (MSS Information Corporation, 655 Madison Avenue, New York, N.Y. 10021).

Galantucci, B., Fowler, C., and Turvey, M. (**2006**). "The motor theory of speech perception reviewed," Psychonomic Bulletin and Review **13**, 361–377.

Hazan, V. and Simpson, A. (**1998**). "The effect of cue-enhancement on the intelligibility of nonsense word and sentence materials presented in noise," Speech Communication **24**, 211–226.

Kapoor, A. and Allen, J. B. (**2011**). "Perceptual effects of plosive feature modification," J. Acoust. Soc. Am. Under review.

Li, F. and Allen, J. (**2011**). "Manipulation of consonants in natural speech," IEEE Trans. Aud. Speech and Lang. Processing **19**, 496–504.

Li, F., Menon, A., and Allen, J. (**2010**). "A psychoacoustic method to find the perceptual cues of stop consonants in natural speech," J. Acoust. Soc. Am. **127**, 2599–2610.

Liberman, A. (**1957**). "Some results of research on speech perception," J. Acoust. Soc. Am. **29**, 117–123.

Liberman, A. (**1996**). *Speech: A Special Code* (The MIT Press, Cambridge, MA).

Liberman, A., Cooper, F., and Delattre, P. (**1955**). "Acoustic loci and transitional cues for consonants," J. Acoust. Soc. Am. **27**.

Liberman, A., Cooper, F., Shankweiler, D., and Studdert-Kennedy, M. (**1967**). "Perception of the speech code," Psychol. Review **74**, 431–61.

Liberman, A., Delattre, P., Cooper, F., and Gerstman, L. (**1954**). "The role of consonant-vowel transitions in the perception of the stop and nasal consonants," Psychol. Mono. **68**.

Liberman, A. and Mattingly, A. (**1985**). "The Motor Theory of Speech Perception Revised," Cognition **21**, 1–36.

Liberman, A. and Mattingly, A. (**1989**). "A specialization of speech perception," Science **243**, 489–494.

Lobdell, B. and Allen, J. (**2007**). "A model of the VU (volume-unit) meter, with speech applications," J. Acoust. Soc. Am. **121**, 279–285.

McGurk, H. and MacDonald, J. (**1976**). "Hearing lips and seeing voices," Nature **264**, 746–748.

Munson, W. (**1947**). "The growth of auditory sensation," J. Acoust. Soc. Am. **19**, 584–591.

Phatak, S. and Allen, J. B. (**2007**). "Consonant and vowel confusions in speech-weighted noise," J. Acoust. Soc. Am. **121**, 2312–26.

Phatak, S., Lovitt, A., and Allen, J. B. (**2008**). "Consonant confusions in white noise," J. Acoust. Soc. Am. **124**, 1220–33.

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

Remez, R., Ferro, D., Wissig, S., and Landau, C. (**2008**). "Asynchrony tolerance in the perceptual organization of speech," Psychonomic Bulletin & Review **15**, 861–865.

Remez, R., Rubin, P., Berns, S., Pardo, J., and Lang, J. (**1994**). "On the perceptual organization of speech," Psychol. Review **101**, 129–156.

Remez, R., Rubin, P., Pisoni, D., and Carrell, T. (**1981**). "Speech perception without traditional speech cues," Science **212**, 947–949.

Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., and Ekelid, M. (**1995**). "Speech recognition with primarily temporal cues," Science **270**, 303–304.

Singh, R. and Allen, J. B. (**2011**). "Sources of stop consonant errors in low-noise environments," J. Acoust. Soc. Am. Under Review.

Stevens, K. N. and Blumstein, S. E. (**1978**). "Invariant cues for place of articulation in stop consonants," J. Acoust. Soc. Am. **64**, 1358–1368.

Turner, C., Fabry, D., Barrett, S., and Horwitz, A. (**1992**). "Detection and Recognition of Stop Consonants by Normal-Hearing and Hearing-Impaired Listeners," J. Speech and Hearing Research **35**, 942–949.