ABSTRACT

Title of Document: THE USE OF ACOUSTIC CUES IN PHONETIC PERCEPTION: EFFECTS OF SPECTRAL DEGRADATION, LIMITED BANDWIDTH AND BACKGROUND NOISE.

Matthew Brandon Winn, Ph.D., 2011

Directed By: Dr. Monita Chatterjee,
Department of Hearing and Speech Sciences

Hearing impairment, cochlear implantation, background noise and other auditory degradations result in the loss or distortion of sound information thought to be critical to speech perception. In many cases, listeners can still identify speech sounds despite degradations, but understanding of how this is accomplished is incomplete. Experiments presented here tested the hypothesis that listeners would utilize acoustic-phonetic cues differently if one or more cues were degraded by hearing impairment or simulated hearing impairment. Results supported this hypothesis for various listening conditions that are directly relevant for clinical populations. Analysis included mixed-effects logistic modeling of contributions of individual acoustic cues for various contrasts. Listeners with cochlear implants (CIs) or normal-hearing (NH) listeners in CI simulations showed increased use of acoustic cues in the temporal domain and decreased use of cues in the spectral domain for the tense/lax vowel contrast and the word-final fricative voicing contrast. For the word-initial stop voicing contrast, NH listeners made less use of voice-onset time and greater use of voice pitch in conditions that simulated high-frequency hearing impairment and/or masking noise; influence of these cues was further modulated by consonant place of articulation. A pair of experiments measured phonetic context effects for the /s/-/ʃ/ (“s/sh”) contrast, replicating previously observed effects for NH listeners and generalizing them to CI listeners as well, despite known deficiencies in spectral resolution for CI listeners. For NH listeners in CI simulations, these context effects were absent or negligible. Audio-visual delivery of this experiment revealed enhanced influence of visual lip-rounding cues for CI listeners and NH listeners in CI simulations. Additionally, CI listeners demonstrated that visual cues to gender influence phonetic perception in a manner consistent with gender-related voice acoustics. All of these results suggest that listeners are able to accommodate challenging listening situations by capitalizing on the natural (multimodal) covariance in speech signals. Additionally, these results imply that there are potential differences in speech perception by NH listeners and listeners with hearing impairment that would be overlooked by traditional word recognition or consonant confusion matrix analysis.

THE USE OF ACOUSTIC CUES IN PHONETIC PERCEPTION: EFFECTS OF SPECTRAL DEGRADATION, LIMIED BANDWIDTH AND BACKGROUND NOISE.

By

Matthew Brandon Winn

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2011

Advisory Committee:
Dr. Monita Chatterjee, Chair
Dr. Sandra Gordon-Salant
Dr. Norbert Hornstein
Dr. William Idsardi
Dr. Rochelle Newman
Dr. Shu-Chen Peng

Acknowledgements

This research has been supported by NIH funding (R01 DC 004786; Monita Chatterjee) and a University of Maryland Center for Comparative and Evolutionary Biology of Hearing Training Grant (NIH T32 DC000046-17; Arthur N. Popper).

The following people played a special role in my work as a scientist and as an audiologist, and deserve special recognition:

My parents, Amy and John Winn, and my grandmother, Irene Buchalter, who continued to support me even when we all lost count of how many years I've been in school

My committee, especially Monita Chatterjee and Bill Idsardi, who have provided guidance and support that extend far beyond this dissertation.

Shu-Chen Peng, Nelson Lu, and Ewan Dunbar
for sharing incredibly helpful ideas and technical skills

Ariane Rhone
who helped develop the ideas and stimuli for Experiment 5

Nan Bernstein Ratner, Allison Blodgett and Jessica Bauman
for valuable support outside the classroom

Margaret McCabe, Judy Schafer, and the Maryland audiology staff
for helping me to hone my skills in the clinic

Friends that provided a meaningful, colorful and rewarding experience:
Cherish Giusto, Kara Schvartz, Viral Tejani, Justine Cannavo Krista Follmer, Nick Follmer, Julian Jenkins, and Phil Monahan, who pulled me back into the lab when this document seemed unattainable.

Finally, this document would have never gotten past page one without the support and kindness of Allison Golden.

Table of Contents

List of Tables

X.Y. indicates Chapter X, Table Y.

List of Figures

X.Y. indicates Chapter X, Figure Y.

List of Abbreviations and Acronyms

AIC: Akaike information criterion

AV: Audio-visual

AVI: audio-video interleave

BW: Bandwidth

CI: Cochlear implant

Est.: Estimate (coefficient of factor in logistic model)

F0: Fundamental frequency

F1, F2, F3: First formant, second formant, third formant

F1T: First formant transition

HI: Hearing-impaired

HINT: Hearing in noise test

Int.: Intercept (of logistic model)

LPF: low-pass filter

NBV: noise-band vocoder/vocoding

NH: Normal-hearing

PSOLA: Pitch-synchronous overlap-add

SNR: Signal to noise ratio

SPP: Spectral peak prominence

VISC: Vowel-inherent spectral change

VOT: Voice onset time

Chapter 1: Introduction to phonetic perception and hearing impairment

Phonetics and Acoustics

Speech perception is arguably the most critical function of the auditory system. The auditory cognitive processes that underlie this ability have been studied for many decades, yet our understanding remains incomplete. In particular, there are a number of unanswered questions regarding the perception of speech sounds by people with hearing impairment or people in adverse listening conditions like background noise. Thorough understanding is elusive for many reasons. Among them are the complex acoustic nature of speech (which includes multiple concurrent changes in multiple dimensions), and the complexity of the auditory system, especially as it relates to cognitive functions like categorization of speech sounds. A principal idea in this dissertation is that the same speech sound can be recognized in more than one way; exploration of strategies taken to identify speech sounds can reveal the nature of auditory processing in a way that is overlooked by commonly used methods like basic word recognition and information transfer analysis. In other words, this investigation diverges from the question of *whether* a listener correctly perceives a speech sound; the goal here is to explain *how* or *why* a speech sound was correctly (or incorrectly) perceived in terms of the acoustic cues that gave rise to a perception.

Phonetic features are linguistic categories such as consonant place, manner, voicing, etc. These are thought by many to participate in the conceptual structure of a language at the phonological level, and may or may not be realized in physical form (i.e. in the speech signal). These categories can be signaled by objectively measurable acoustic cues like duration, frequency, spectral peaks, etc. The correspondence of cues to features is redundant because multiple cues convey the same feature information. The cues can be described along multiple dimensions, including frequency, timing, amplitude, or some dynamic combination of all three. Additionally, some of these characteristics can vary even for the same sound, depending on the phonetic context in which they appear. Since multiple acoustic attributes are typically co-varying, it is not always clear which are actually most essential to perception. Many years have passed without consensus on what acoustic cues are critical to the identification to any particular phoneme. This has been coined the "lack of invariance problem" because seemingly invariant phonetic percepts do not map to invariant acoustic attributes. Although this has been a source of inquiry (and perhaps frustration) for numerous speech and hearing scientists, the perspective taken here is that this multi-dimensional aspect of speech is a potential resource to people with hearing impairment, for whom one or more cues are rendered unclear or unavailable.

The presence of multiple phonetic cues is observed for many different phonetic contrasts. A classic example is the contrast between voiced and voiceless stops in word-medial position, which has been claimed to contain at least 16 different acoustic cues (Lisker, 1978). A wealth of literature has revealed that changes in one acoustic dimension can be compensated by conflicting changes in another dimension (for multiple examples, see Repp, 1982). For example, trading relations can be observed in the integration of cues for syllable-initial stop consonant voicing; changes in voice-onset-time that signal voicing can be somewhat offset by changes in the pitch domain that signal voicelessness (Whalen, Abramson, Lisker & Mody, 1993). As these and other cues co-vary in natural speech, the listener must integrate them in a way that yields reliable and accurate identification of the incoming information. It has been shown that the use of acoustic cues for phonetic contrasts is affected by the developmental age (Nittrouer, 2004; 2005) as well as language background (Morrison, 2005) of a listener. Perhaps it is also affected by spectral resolution in a way that is useful for understanding the experience of listeners with hearing loss or who use cochlear implants.

There are several models of phonetic cue perception present in the literature, and the current studies were not designed to explicitly test or challenge any of them. Of particular interest, however, are those accounts which specifically acknowledge the reliability with which the signal is represented. The cue weighting-by-reliability model by Toscano and McMurray (2010) suggests that the weighting of acoustic cues in phonetic perception can be predicted by their distributional properties in the input; the basic theme of this research permeates other work on speech perception (Holt & Idemaru, 2011) and basic auditory categorization (Holt & Lotto, 2006). Specifically, a cue is more reliable (and hence should be more heavily-weighted) if the contrastive level means are far apart and have low variance. In the current study, it could be argued that signal degradation (whether via real or simulated hearing impairment) would diminish the reliability of some cues since their acoustic representation is less clear. The underlying theme of all the experiments here is that although some cues are thus degraded, listeners can capitalize on the perception of residual cues that remain unaffected (since they are implemented in a different domain).

The current investigation exploits the existence of multiple co-varying cues in phonetic perception by positing that the cues that contribute most to perception by normal-hearing listeners may be different than those that contribute most to perception by listeners with hearing impairment. While this idea is not surprising in the face of the auditory constraints of hearing loss, it has been largely unexplored in the literature. Some common kinds of analysis might even draw attention *away* from this issue. For example, both researchers and audiologists frequently measure the number of words correctly heard/repeated in a listening task. When a phonetic distinction (e.g. voicing) is perceived correctly, it might be dismissed as unproblematic or unworthy of further exploration. While this may represent a kind of "bottom line" of assessment for a listener, it does not probe thorough understanding of the perceptual processes that give rise to speech perception. Even analyses that specifically identify the phonetic features that are perceived (e.g. Information transfer analysis; Miller & Nicely, 1955) do not identify acoustic cues relevant for phonetic feature perception. This is even true of studies designed specifically to explore the use of phonetic cues by such populations as cochlear implant users (Iverson, Smith & Evans, 2006). This problem is akin to exploring the journey of two travelers who arrive at the same destination; it is not clear whether they each traveled with the same amount of comfort, efficiency or ease, despite both ending up at the same place. Thus the purpose of the investigations presented here is to explore whether the strategies used by those with hearing impairment differ notably from those used by those with normal hearing.

Overview of Chapters

Chapters 2 and 4 address the issue of spectral degradation, which is essential to understanding the experience of individuals with cochlear implants (CIs). Although CIs have been immeasurably successful as a neural prosthetic device (Zeng, Rebscher, Harrison, Sun & Feng, 2008), they suffer from a number of limitations that result in a distorted and unclear representation of sound in the frequency (spectral) domain (Henry, Turner & Behrens, 2005). These limitations are discussed in Chapter 2. In the experiments presented here, only a small number of CI listeners were available to test, so the comparison between them and normal-hearing (NH) listeners is supplemented by a group of NH listeners presented with speech signals processed simulations of a CI. Although these simulations differ from actual electric hearing (the use of a CI) in a number of ways, they have become a customary means of predicting CI listener performance in speech perception tasks.

It was hypothesized that this spectral degradation would result in a decreased ability to utilize acoustic cues in the spectral domain, and also result in an increased utilization of acoustic cues in other domains, such as the temporal domain. Chapter 2 addresses this hypothesis by exploring the perception of two phonetic contrasts in American English: the tense/lax vowel contrast and the final consonant voicing contrast. These contrasts were chosen because they have been shown to be perceptually driven primarily by cues in the spectral domain, and only minimally affected by temporal phonetic cues when spectral complexity of speech is faithfully maintained in stimuli (Hillenbrand, Clark & Houde, 2000; Hillenbrand, Ingrisano, Smith & Flege, 1984; Wardrip-Fruin, 1982). Thus, examining these contrasts is a good avenue to explore how listeners behave in a situation where the usual strategy is compromised, but when there are alternative strategies available. Results of such experiments can shed light on the robustness of speech perception under listening constraints, and can help to contribute a more complete understanding of the impaired auditory system.

Chapter 3 addresses the problem of high-frequency hearing loss and/or background noise, which are experienced by many listeners. The issue explored here is the perception of syllable-initial stop consonant voicing contrasts. This is one of the easiest contrasts for a listener, regardless of hearing impairment or the presence of background noise. Chapter 3 addresses the question of why this voicing contrast can be accomplished so accurately and reliably in such adverse conditions. As in Chapter 2, the hypothesis was that when some acoustic cues for this contrast are compromised by an adverse listening condition, listeners shift reliance to other cues that remain intact in the signal (even if those residual cues were not very important in optimal listening conditions). The primary issue was whether listeners can shift from reliance on voice onset time to reliance on voice pitch. Voice pitch has previously been implicated as playing an important role when listening in noise (Brokx & Noteboom, 1982; Laures & Wiesmer, 1999), and the current investigation helps to unpack the contributions of this cue at the segmental level. Some additional considerations within this chapter include the differences in phonetic cue perception for two contrasts typically regarded as the same at the feature level (i.e. p/b and t/d), and the interaction of masking noise with the inaudibility of high-frequency energy. Experiments in Chapter 3 did not utilize individuals with hearing impairment. Instead, normal-hearing listeners were presented with sounds that roughly simulated hearing impairment (in terms of the loss of high-frequency energy). It should be noted that hearing impairment is a much more complex phenomenon than this simulation, as listeners with hearing impairment suffer from poorer-than-normal spectral resolution (although not to the same extent as CI users), and other distortions of sounds that are loud enough to be heard (i.e. suprathreshold distortion). Still, the signal processing used in the experiments in Chapter 3 offers insight into the role of one critical aspect of hearing impairment, and does so with rigorous control over the frequency range of the signal.

Chapter 4 addresses the effect of various auditory and visual contexts in phonetic perception, and how this varies with spectral degradation encountered by CI listeners. It has been shown that when listeners identify an acoustic segment, they are sensitive to the context in which that segment is presented. For example, when presented with a sound intermediate between “s” and “sh,” they tend to hear “s” when the talker is perceived to be male, but hear “sh” when the talker is perceived to be female (Mann & Repp, 1980). Various other similar phenomena have been observed, which have grown to become known collectively as phonetic context effects or spectral contrast effects. For this kind of perception to occur, the listener must attend not only to the phonetic cues that comprise a segment, but also to those that comprise neighboring segments, and be sensitive to the regular coarticulatory influences that segments exert on each other. For example, when producing the word “Sue,” the lips are round for the /u/ vowel during the production of the /s/ sound, thereby creating a vocal tract configuration that lowers the frequency composition of the /s/. Listeners are normally able to accommodate this variation in /s/ (as well as that for other phonemes) by accordingly lowering the frequency range they consider to be /s/. It is not clear whether a CI listener can show this accommodation because the acoustic cues for the various contexts are themselves degraded. This degradation results in less distinction between contexts, and therefore potentially lower likelihood that listeners would accommodate those contexts with differences in phonetic perceptions.

A growing number of researchers attribute phonetic contexts effects largely to the presence of spectral contrast between neighboring segments (Holt, Stephens & Lotto, 2005). It has been observed that speech perception and basic auditory science could be integrated more closely (Holt & Lotto, 2008), and this type of experiment is one avenue through which each branch of science could be mutually beneficial. Specifically, context effects reveal how listeners adjust perception of some discrete segment according to how it changes over time; this is one of the basic understandings of how information would be coded by neurons (i.e. that they encode stimulus *change*), so the translation of behavioral results to a biological understanding could potentially be easier (Kluender, Coady & Kiefte, 2003). Additionally, the task of showing sensitivity to phonetic context can be interpreted as a more subtle kind of auditory categorization ability than that tested in Chapter 2.

The results in this dissertation help to shed light on just how robust the human auditory system is. While critical issues and problems are still abundant for people with hearing impairment, the results presented here offer evidence that some major auditory signal degradations can be overcome. Specifically, listeners are able to utilize multiple sound components when perceiving speech sounds, and can compensate for the loss of one cue with increased use of other cues.

Chapter 5 summarizes the conclusions of all the experiments presented here, and provides commentary on the implications for auditory/cognitive science, audiology, and the manufacturing of assistive devices for listeners with hearing impairment.

General methodological approach

The issue of phonetic cue-weighting or cue use has been rarely applied to listeners with hearing impairment, and there is a dearth of understanding of this crucial topic. Unlike traditional methods of phonetic feature perception analysis, the results in this dissertation offer a window into the perceptual processes (i.e. acoustic contributions) that underlie feature perception. Importantly, signal processing and statistical methods have improved markedly since some of the well-known work on this general topic. The present study has benefited considerably from the availability of superior tools.

Most previous studies (especially those that incorporate information transfer analysis; Miller & Nicely, 1955) measured the amount of success that listeners have in perceiving phonetic features. In fact, at least one study (Wang & Bilger, 1973) was conducted in order to evaluate which sets of features are optimal for this exercise. The rationale behind this analysis is that we know some of the acoustic cues that correspond to each feature, and therefore can determine which cues are perceived correctly by measuring success in perception of those features. It must be acknowledged however, that information transfer analysis only provides an indirect assessment of the use of acoustic cues. This dissertation takes this analysis one step further by decomposing features into their constituent acoustic cues. The advantage of this rigorous approach is that it minimizes assumptions about the correspondence of features to acoustic cues.

Some previous assessments of cue use (Wardrip-Fruin, 1985; Hillenbrand & Nearey, 1999; Hillenbrand et al., 2000; Iverson et al., 2006; Li & Allen, 2011) measured performance when that cue was either present in the signal or artificially removed (using various means such as neutralization, noise masking or filtering). If there was negligible change in perceptual identification, the cue was presumed to not contribute. If there was a dramatic perceptual change, then the cue was assumed to contribute. In this dissertation, the question of whether cues contribute is only part of the story. Of interest in this dissertation is whether the cues are used with the same facility by different listeners who experience different impairments or signal degradations. In other words, there are many subtleties that exist between presence or absence of cue reliance, and some of those subtleties are explored here.

The experiments in this dissertation used entirely novel stimulus materials created by the author (with the exception of the visual component of the last experiment, which was developed by Ariane E. Rhone specifically for this project). With the exception of stimuli in the first experiment and synthetic fricatives in the last experiment, all materials were created from natural speech, which preserves the rich signal complexity previously shown to be compromised by speech synthesizers.

The materials were designed to change systematically various acoustic cues (such as voice onset time, fundamental frequency (F0), formants, vowel and consonant duration, etc.) in such a way that they could be independently evaluated for perceptual consequence. Thus, the known problem of cue redundancy in speech was addressed to the best extent possible.

Following the cogent argument by Morrison and Kondaurova (2009) in favor of logistic regression as a modeling technique, the current analyses used generalized linear (logistic) mixed-effects modeling (GLMM). This modeled the prediction of the likelihood of perceptual response not only at the tested levels, but also at levels intermediate and beyond those tested. Importantly, it provided a quantitative measure of the rate of perceptual change that results from a change in a stimulus attribute. The GLMM offers a quantitative tool for evaluating the hypotheses presented here.

Some results presented in this dissertation were expected, while others were surprising. Expected results include the decreased use of spectral cues and increased use of durational cues by listeners with cochlear implants, as well as the increased use of voice pitch as a cue for voicing in stop consonants that are band-pass filtered or masked by noise. More surprising results are found in Experiment 5, which suggests that listeners with cochlear implants not only use visual phonetic cues like lip-rounding to perceive speech sounds, they also are sensitive to visual cues to a talker gender in a way that is consistent with gender-related differences in vocal properties. In other words, listeners with hearing impairment can recruit information not only from different domains in auditory space, but also recruit information from another sensory modality entirely. Undoubtedly, this is the highlight of the dissertation. Another somewhat surprising result in Experiment 2 was that listeners with cochlear implants did not show significant use of voicing within a fricative to perceive the /s/-/z/ contrast. This voicing cue within a fricative is akin to amplitude modulation in a broadband noise. Since cochlear implant users have been shown previously to be sensitive to amplitude modulations (Shannon, 1992), it is surprising that they did not use it effectively in this experiment.

Experiment 3 yielded results that were not entirely surprising, but still novel in the sense of exploring an existing issue more thoroughly. The effects of masking noise and signal bandwidth on the perception of stop consonants was found to depend on place of articulation in a manner consistent with acoustic differences between consonant places. For /t/ and /d/, the audibility of high-frequency energy (above 4 kHz) was especially important: without it, listeners' cue-weighting strategies changed if there was even just a minimal amount of noise in the signal. For /p/ and /b/, this was not the case. Instead, the /p-b/ contrast was driven more heavily by the amount of masking noise applied to the signal, because bandwidth was less important. Only when the signal-to-noise ratio was very unfavorable did listeners switch cue strategies for /p/ and /b/. Thus, although traditional analyses treat /t-d/ and /p-b/ as a phonetically identical contrast (i.e. the results for both pairs are averaged), there is useful insight gained by evaluating them separately.

It is hoped that the care taken to evaluate the contributions of independent acoustic cues will shed new light on the robustness of the auditory system, as well as help to explain older results from which conclusions cannot be easily drawn without making multiple assumptions.

Chapter 2: Effects of spectral degradation on phonetic cue perception

Introduction and Motivation

Cochlear implants and spectral resolution.

In view of the remarkable success of the cochlear implant (CI) as a prosthetic device (Zeng et al., 2008), and in the context of a continually growing body of research on cochlear implants, literature on phonetic cue perception must be expanded to acknowledge the abilities of individuals fitted with these devices. It is well known that a major obstacle to accurate speech understanding with electric hearing (the use of a CI) is the poor spectral resolution offered by these devices, owing to the limited number of independent spectral processing channels (Fishman, Shannon & Slattery, 1997; Friesen, Shannon, Başkent & Wang, 2001), interactions between the electrodes which carry information from those channels (Chatterjee & Shannon, 1998), as well as a distorted tonotopic map (Fu & Shannon, 1999). Thus, the subtle fine-grained spectral differences perceptible to those with normal hearing are not reliably distinguished by those who use CIs (Kewley-Port & Zheng, 1998; Loizou & Poroy, 2001; Henry et al., 2005). In view of some of these studies, it is presumed that phonetic cues driven by spectral contrasts would be most challenging for CI listeners. Although numerous studies have explored word, phoneme and feature recognition with various kinds of signal degradations (such as band-pass filtering, masking noise and vocoding), few have explored the use of acoustic cues that contribute to these perceptions.

Not all sound components are compromised in electric hearing; temporal processing can be as good or better than that of normal-hearing (NH) listeners, as evidenced by temporal modulation transfer functions (Shannon, 1992) and gap detection tasks (Shannon, 1989). Thus, although some phonetic cues are obscured by spectral degradation, it is expected that CI listeners should be able to use cues in the temporal domain, including segment duration and time-varying amplitude changes. Fittingly, experiments using CI listeners have revealed a large number of errors on place-of-articulation perception (which relies primarily upon spectral cues in the signal, such as spectral peak frequencies and formant transitions), while the manner-of-articulation and voicing features are rarely mis-perceived, because they can be transmitted via temporal cues, which are well-maintained in electric hearing (Dorman, Dankowski, McCandless, Parkin & Smith, 1991). Similar results of poor place perception and excellent voicing perception have been shown continually for NH listeners listening to simulations of cochlear implants (i.e. Shannon, Zeng, Kamath, Wygonski & Ekelid, 1995). A major assumption inherent in these explanations is that there are specific acoustic cues that carry phonetic features. On the contrary, the current state of speech science overwhelmingly supports the notion that there are multiple acoustic cues that collectively signal phonetic contrasts. Therefore, some questions remain as to what cues underlie perception in difficult listening conditions.

Stimulus fidelity and phonetic cue perception

Perception of acoustic cues such as duration, formant frequencies, or the time-varying amplitude envelope all should depend on the fidelity of the stimulus. Perception of vowels and consonants by CI users is genderally poorer than that by NH listeners.

Dorman and Loizou (1997; 1998) and Friesen et al. (2001) suggest that 80% correct performance is achieved by the upper echelon of CI listeners, with mistakes spread across a wide range of speech sounds. Specifically, sounds thought to depend heavily on specific frequency information (vowel formants, formant transitions, or other spectral peaks) are most susceptible to error. This pattern persists for NH listeners who are subjected to speech signals that are degraded in the spectral domain (Shannon et al., 1995; Friesen et al., 2001; Xu, Thompson & Pfingst, 2005). Even with poor resolution, performance can be quite high, but perhaps not as high as that observed under optimal conditions.

Trading relations between temporal and spectral resolution have been observed for the perception of English consonants and vowels (Xu et al., 2005) as well as for Mandarin lexical tones (Xu & Pfingst, 2003). In those studies, the degree of spectral resolution was progressively decreased for NH listeners using noise-band vocoding (NBV), a process where spectral detail is replaced by a fixed level of coarse resolution (to be described in detail later in the Method section of Experiment 1). As the spectral resolution grew poorer, temporal resolution (which was itself altered) appeared to play a larger role in listeners' perceptual accuracy (Xu et al., 2005). Although Xu et al. leveled claims about the contributions of spectral and temporal "cues" in this study, they actually measured the perception of phonetic features, which contain multiple cues. Essentially, they measured the contributions of spectral and temporal *resolution* (clarity of the signal) in the perception of *features* (linguistic categories). Although those measurements are extremely valuable, they are different than the contributions of spectral and temporal *cues* (acoustic characteristics of the phonetic segments) at different levels of resolution. Since no cues were independently adjusted or measured in the study, questions remain as to which cues were used by the listeners to perceive the features. The current study adjusts the experimental approach to address this issue – beyond showing correct and incorrect performance for feature recognition, what can we learn about the avenue that listeners take to perceive those features?

There is reason to believe that listeners will adapt to an altered stimulus input by changing the relative perceptual importance of signal components. For example, listeners can show adjustment in acoustic cue-weighting as a result of targeted training (Francis, Baldwin & Nusbaum, 2000; Francis, Kaganovich & Driscoll-Huber, 2008, among many examples) or passive exposure to the distributional properties of the cues (Holt & Idemaru, 2011). Perhaps cochlear implant listeners and normal-hearing listeners in degraded conditions can adopt new strategies that would suit the challenges and residual abilities available to them.

Experiment 1: Spectral degradation and the tense/lax contrast

Acoustics and perception of the tense/lax contrast

The first experiment explored the high-front lax / tense vowel contrast (/I/ and /i/) in English, which distinguishes word pairs such as hit/heat, fill/feel, hid/heed and bin/bean. The cues that contribute to this distinction include the spectral dimensions of formant structure and vowel-inherent spectral change (VISC), as well as vowel duration. Formant structure has long been known to correspond to vowel categorization, albeit with a considerable amount of overlap between categories (Hillenbrand, Getty, Clark &

Wheeler, 1995). Vowels can be classified according to the first, second and (sometimes) third formants (vocal tract resonances), commonly labeled F1, F2 and F3, respectively. These formant cues are extremely powerful; using only steady-state formants synthesized from measurements taken at one timepoint in a vowel (i.e. a spectral snapshot), human listeners identify vowels with roughly 75% accuracy (in tasks where chance levels vary between 6% – 10%) (Hillenbrand & Gayvert, 1993). Automatic pattern classifiers show similarly good performance with just one sample of formant structure (Hillenbrand et al., 1995).

VISC refers to the “relatively slowly varying changes in formant frequencies associated with vowels themselves, even in the absence of consonantal context” (Nearey & Assmann, 1986). Throughout production of the lax vowel /I/, F1 increases and F2 decreases; the tense vowel /i/ is relatively steady-state by comparison, with only a negligible amount of formant movement, if any (Hillenbrand et al., 1995). VISC plays a role in vowel classification, as indicated by at least 4 kinds of data: 1) production data indicates changes in formant measures taken at different times during vowel pronunciation (Nearey & Assmann, 1986; Hillenbrand et al., 1995), 2) results of pattern classifiers show better performance when spectral change is recognized (Zahorian & Jagharghi, 1993; Hillenbrand et al., 1995), 3) listeners reliably identify vowels with only snapshots of the onset and offset (with silent or masked center portions) (Jenkins, Strange & Edman, 1983; Parker & Diehl, 1984; Nearey & Assmann, 1986), and 4) human listeners show improved identification results when vowels include natural patterns of spectral change; there is generally a 23-26% decline in accuracy for vowels whose formant structure lacks time-varying change (Hillenbrand & Nearey, 1999; Assmann &

Katz, 2005). Specifically, when formants are artificially made to be un-changing via signal processing, there is a significant decline in /I/ recognition, while the vowel /i/ is identified virtually perfectly (Assmann & Katz, 2005), consistent with the acoustics of these vowels.

The duration of tense vowels tend to be longer than that of lax vowels by roughly 33 – 80%, depending on the particular contrast and context (House, 1961; Hillenbrand et al., 1995). However, the role of duration in vowel perception has not always been clear; it appears to be driven at least in part by stimulus fidelity (the degree to which the synthesized signal reproduces the spectral and temporal properties of the natural speech on which it is based). Ainsworth (1972) showed that duration can modulate identification of vowels synthesized with two steady-state formants. Bohn and Flege (1990) and Bohn (1995) revealed a small effect of duration for the /i/ v. /I/ contrast when using three steady-state formants. However, these results are challenged by other studies that preserved relatively richer spectral detail, including time-varying spectral information (Hillenbrand et al., 1995; 2000; Zahorian & Jagharghi, 1993). Using modified natural speech, Hillenbrand et al. (2000) reported that duration-based misidentifications of the /i/-/I/ contrast were especially rare (with an error rate of less than 1%). An emergent theme from Hillenbrand et al. (2000), Nittrouer (2004) and Assmann and Katz (2005) is that the use of acoustic cues in vowels is affected by signal fidelity, to the extent that commonly used formant synthesizers are likely to underestimate the role of time-varying spectral cues, and to overestimate the role of durational cues. That is, listeners use phonetic cues differently depending on the quality with which the sound is presented.

Vowel perception, cochlear implants and signal degradation

Although considerable improvements in speech synthesis and manipulation have improved the quality of signals in perceptual experiments, signal degradation is inescapable for individuals with cochlear implants. Iverson et al. (2006) remarked, “It would be surprising if exactly the same cues were used when recognizing vowels via cochlear implants and normal hearing, because the sensory information provided by acoustic and electric hearing differ substantially.” Despite the aforementioned trend observed in spectral and temporal signal fidelity, Iverson et al. (2006) did not find evidence to suggest that duration was more heavily used by CI listeners or NH listeners in degraded conditions. In fact, as spectral resolution was degraded from 8 to 4 to 2 channels (each representing progressively worse resolution, to be explained further in the Method), NH listeners showed *less* recovery of duration information in the signal. This counterintuitive result may have arisen because of the methods by which duration cue use was assessed. The experimenters used Information Transfer Analysis (ITA) (Miller & Nicely, 1955) to track phonetic features that were recovered or mistaken in the identification tasks. Although these features are commonly thought to correspond regularly to acoustic dimensions (i.e. vowel height as variation in F1 frequency, vowel advancement as variation in F2 frequency, lax/tense as duration), ITA by itself does not reveal the mechanisms (cues) by which the features are recovered. This is particularly important for the duration cue; most dialects of English do not contain vowel pairs that contrast exclusively by duration. Thus, for any vowel that ITA recognizes as “long” or “short,” there are accompanying co-varying spectral cues. If a listener relies on these spectral cues (as would be predicted on the basis of aforementioned work), then it is not surprising that “duration” information transmission declined as spectral resolution decreased. In the ITA sort of analysis, “duration” could be merely a different name for spectral information, unless the latter has been specifically controlled. The question remains then, as to whether changes in vowel duration play a greater role in vowel identification when spectral resolution is degraded.

Despite the limitations of the ITA-based analysis, the work by Iverson et al. (2006) laid the groundwork for studying the role of specific acoustic cues with varying degrees of temporal and spectral resolution. This approach has been only sparingly applied to the problem of speech perception by CI listeners (Dorman et al., 1991 is a rare example), and it is the aim of the present chapter to explore it further using two contrasts that have been shown to involve both spectral and temporal cues. Many previous experiments (Hillenbrand & Nearey, 1999; Hillenbrand et al., 2000; Iverson et al., 2006) have assessed the role of multiple cues by retaining them or neutralizing them in an all-or-none fashion. The current experiment seeks to expand upon this work by manipulating acoustic cues gradually and orthogonally, so as to assess their effects in a fine-grained way that is not feasible in experiments that test for many vowels and consonants concurrently.

Hypotheses

The general hypothesis was that listeners with cochlear implants or listening with cochlear implant simulations would demonstrate less use of formant and VISC cues, and more use of the duration cue for the tense/lax contrast, relative to listeners with normal hearing. Some prior work indicates that listeners with implants do exhibit altered use of acoustic cues in speech perception. In a place-of-articulation identification task, Dorman et al. (1991) showed that CI listeners identified stop consonant place of articulation based on gross spectral tilt cues while NH listeners relied instead on vocalic formant transitions. Kirk, Tye-Murray and Hurtig (1992) found that CI listeners were able to make use of static formant cues in vowels, but did not take advantage of a slow/fast formant transition contrast used by NH listeners. This would suggest that the slow-changing dynamic formant cue for lax vowels may be compromised in degraded conditions, relegating the listener to rely only on the gross spectral pattern or a non-spectral cue like vowel duration. Fittingly, Dorman and Loizou (1997) indicated that CI listeners identified the lax vowel /I/ with accuracy similar to that of NH listeners in conditions where VISC is artificially removed in order to maintain spectral shape across the entire vowel (Hillenbrand & Gayvert, 1993). The perception of speech sounds by CI listeners was therefore expected to fall in line with aforementioned work that implicates signal degradation as an influential force on the use of durational cues. Thus, it was predicted that as spectral resolution became poorer, use of formant cues would decline, the use of VISC cues would decline (if at all present), and the use of temporal cues would increase.

Method

Participants.

Participants included 15 adult (14 between the ages of 19-26; mean age 22.7 years, and one 63 year-old) listeners with normal hearing, defined as having pure-tone thresholds <20 dB HL from 250–8000 Hz in both ears (ANSI, 2010). A second group of participants included 7 adult (age 50-73; mean age 63.5 years) recipients of cochlear implants. CI listeners were all post-lingually deafened. Six were users of the Cochlear Freedom or N24 devices, and one used the Med-El device. See Table 2.1 for demographic information and speech processor parameters for each CI user. All participants were native speakers of American English and were screened for self-reported familiarity with languages that use vowel duration as a phonemic feature (for example, words in Finnish, Hungarian, Arabic, Vietnamese, etc. can reportedly be distinguished solely on the basis of vowel duration with negligible co-variation in other acoustic dimensions), to ensure that no participant entered with a priori bias towards durational feature sensitivity. Normal-hearing participants 01 (the author) and 02 were highly familiar with the stimuli, having been involved in pilot testing and the construction of the materials. It should be noted that the age difference between the normal-hearing and cochlear implant listener group is substantial, and can influence auditory processing in a way that is relevant to this study. Specifically, auditory temporal processing is known to be deficient in older listeners (Gordon-Salant & Fitzgibbons, 1999). The current study explores whether auditory cues in the temporal domain can overcome those that are compromised in the spectral domain. Older listeners may or may not experience deficiencies in the temporal domain that could complicate this matter. Aside from this possible problem, there also exists variability in the durations and etiologies of deafness among the group of listeners with hearing impairment (as is the case in virtually all studies that use CI listeners). For these reasons, direct statistical comparisons between the NH listeners and CI listeners are limited in their utility and thus omitted from this chapter.

Table 2.1.

Demographic information about the CI participants.

ID #	Gender	Etiology of HL	Duration of HL	Age at testing	Age at impl.	Device	Pulse rate
C12	F	Unknown	Unknown	66	63	Freedom	900
C18	F	Genetic	10 years	66	63	Freedom	1800
C20	M	Unknown	22 years	64	57	N 24	900
C25	M	Labyrinthitis	11 years	50	40	N 24	720
C30	M	Unknown	Unknown	56	54	Med-El	1515
C36	F	Measles	59 years	71	66	Freedom	1800
C42	F	Unknown	4 years	73	69	Freedom	2400

Note: All CI listeners used the ACE processing strategy and the MP1+2 stimulation mode except for C30, who used the CIS strategy.

Stimuli.

Speech synthesis. Words were synthesized to resemble /hIt/ (“hit”) and /hit/ (“heat”). The vowels in these words varied by formant structure (in 7 unequal steps, with the first four formants all simultaneously varying), vowel-inherent spectral change (in 5 steps, with the first three formants all simultaneously varying throughout the vowel) and vowel duration (in 7 unequal steps). See Table 2.2 for a detailed presentation of the levels for each parameter. This 7x7x5 continuum of words (245 total) was synthesized using HLSYN (Hanson, Stevens & Beaudoin, 1997; Hanson & Stevens, 2002). Formant structure was based on values reported in the online database corresponding to the publication of Hillenbrand et al. (1995). Formant continuum steps were interpolated using the Bark frequency scale (Zwicker & Terhardt, 1980) to reflect the non-linear frequency spacing in the human auditory system. Levels in Bark frequency were converted to Hz in this paper to facilitate ease of interpretation. A second dimension of stimulus construction varied the amount and direction of vowel-inherent spectral change (VISC). Although there are various ways of modeling this cue (Morrison & Nearey, 2007), it is represented here in terms of the difference in the F1, F2 or F3 frequency (in Hz) from the 20% to the 80% timepoints in the vowel; positive numbers indicate upward change in frequency. All three formants were changed in accordance with data from Hillenbrand et al. (1995), except the fourth formant, which was kept constant. The penultimate items in this VISC continuum were modeled after typical lax and tense vowels, and the continuum endpoints were expanded along this parameter, again to account for productions outside the means reported by Hillenbrand et al. See Figure 2.1 for a schematic illustration of VISC, and Table 2.2 for a detailed presentation of this parameter. Vowel durations were modeled from characteristic durations of /i/ and /I/ (before voiceless stop sounds) reported by House (1961), and linearly interpolated (see Table 2.2). Word-initial [h] was 60ms of steady voiceless/aperiodic formant structure that matched that at the onset of the vowel; necessarily, the initial consonant was also varied as a result of the formant continuum. Word-final [-t] transitions targets for F1, F2, F3 and F4 were 300, 2000, 2900, and 3500 Hz, respectively, as used by Bohn and Flege (1990). These transitions all began at the 80% timepoint in the vowel (although this decision resulted in slightly different transition speeds depending on overall duration, it was necessary to ensure that the entire 20% - 80% VISC trajectory could be realized). The formant transition was followed by 65 ms of silent stop closure, followed by 65 ms of diffuse high-frequency [t] burst. Vowel pitch began at 120Hz, rose to 125Hz at the 33% timepoint of the vowel, and fell to 100 Hz by vowel offset.

Figure 2.1 Illustration of vowel-inherent spectral change (VISC)

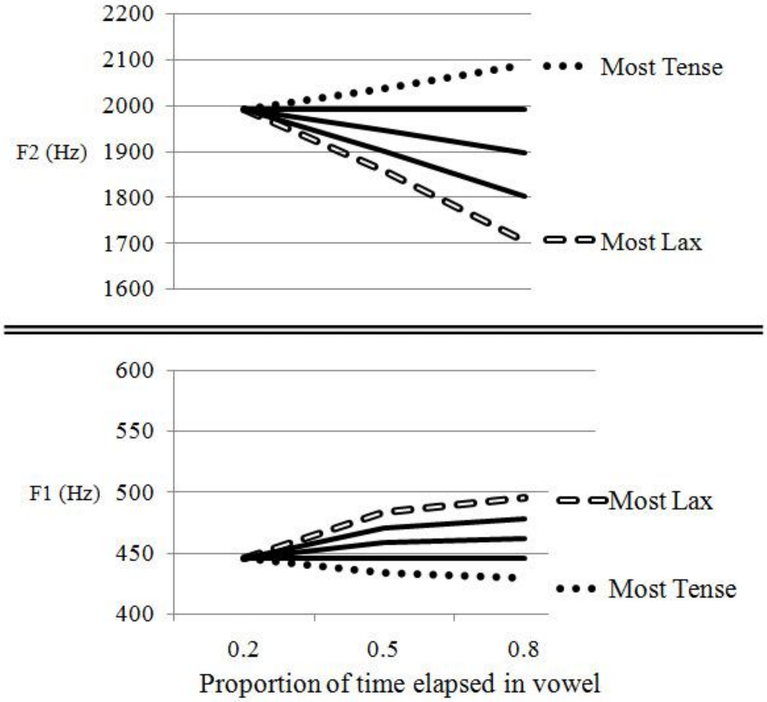

Figure 2.1. Stylized representation of different levels of VISC applied to the same initial formant structure.

Table 2.2.

Acoustic parameter levels defining the continua of formants, vowel-inherent spectral change and vowel duration.

		Continuum step number						
		1	2	3	4	5	6	7
Formants (Hz)	F1	446	418	403	389	375	362	335
	F2	1993	2078	2122	2167	2213	2260	2357
	F3	2657	2717	2747	2778	2809	2841	2905
	F4	3599	3618	3628	3637	3647	3657	3677
VISC (change in Hz)	F1	+49	+33	+16	0	-16		
	F2	-287	-191	-96	0	+96		
	F3	-33	-22	-11	0	+11		
	F4	0	0	0	0	0		
Duration (ms)		85	100	108	115	122	130	145

Note: VISC is represented by change in formant frequencies [Hz] from the 20% to 80% timepoints in the vowel. Each parameter was varied orthogonally.

Spectral degradation: Noise-band vocoding.

Spectral resolution was degraded using noise-band vocoding (NBV), which has become a common way to simulate a cochlear implant (see Shannon et al., 1995). This was accomplished using online signal processing within the iCAST stimulus delivery software (version 5.04.02; Fu, 2006). Stimuli were bandpass filtered into 4 or 8 frequency bands using sixth-order Butterworth filters (24 dB/octave). This number of bands was chosen to best approximate the performance of CI listeners (Friesen et al., 2001). The temporal envelope in each band was extracted by half-wave rectification and low-pass filtering with a 200-Hz cutoff frequency, which is sufficient for good speech understanding (Shannon et al., 1995). The envelope of each band was used to modulate the corresponding bandpass-filtered noise. Specific band frequency cutoff values were determined assuming a 35 mm cochlear length (Greenwood, 1990) and are listed in Table 2.3 below. The lowest frequency of all analysis bands (141 Hz, 31 mm from the base, approximately) was selected to approximate those commonly used in modern CI speech processors. The highest frequency used (6000 Hz, approximately 9 mm from the base) was selected to be within the normal limits of hearing for all listeners. No spectral energy above this frequency was available to listeners in the unprocessed condition. Spectrograms of the word “hit” in the unprocessed (regularly synthesized), 8-channel NBV and 4-channel NBV versions are illustrated in Figure 2.2. The images show that specific formant frequency bands are no longer easily recoverable; the spectral fine structure is replaced by coarse / blurred sampling. Formant differences that remain unresolved within the same spectral channel are coded by the relative level of the noise band carrying that channel, as well as the time-varying amplitude (i.e. beating) owing to the interaction of multiple frequencies added together.

Table 2.3.

Specification of analysis & carrier filter bands for the noise-band vocoding scheme for Experiment 1.

	Channel number							
4-channel	1		2		3		4	
8-channel	1	2	3	4	5	6	7	8
High-pass (Hz)	141	275	471	759	1181	1801	2710	4044
Low-pass (Hz)	275	471	759	1181	1801	2710	4044	6000

Figure 2.2 Spectrograms of "hit" and "heat" with different levels of spectral resolution

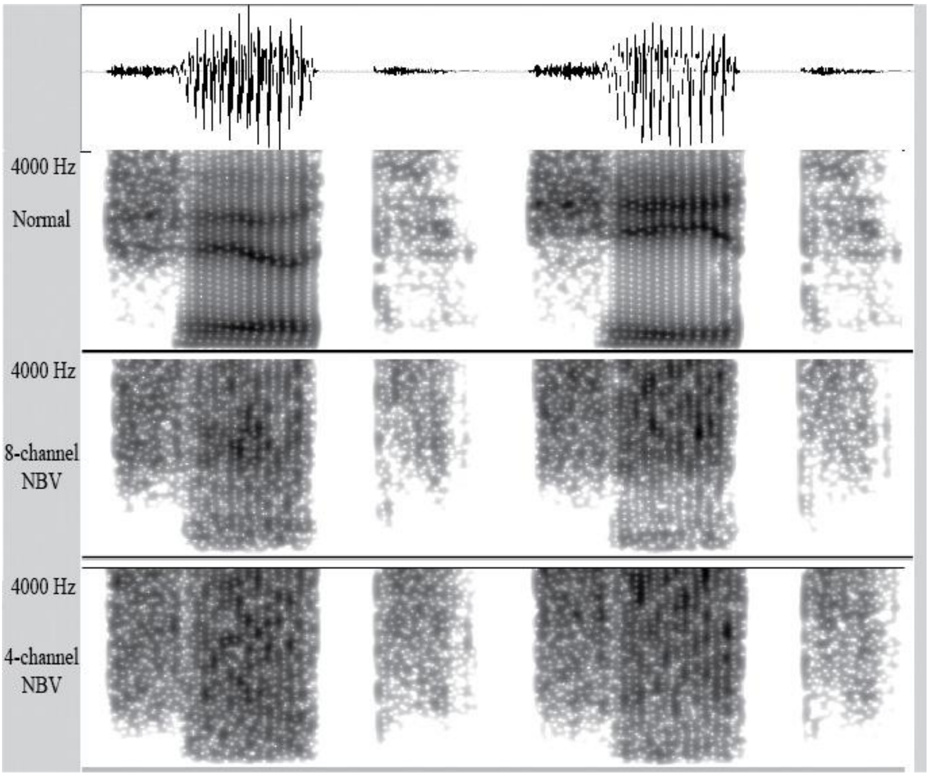

Figure. 2.2. Spectrograms illustrating synthesized words "hit" (left) and "heat" (right) in the normal/unprocessed condition (top), 8-channel noise-band vocoder (middle) and 4-channel noise-band vocoder (bottom) conditions.

Procedure.

All speech recognition testing was conducted in a double-walled sound-treated booth. Volume level was calibrated at the position of the listener’s head using a Radio Shack sound level meter that referenced a 1 kHz tone that was equated in RMS amplitude to the speech stimuli. Stimuli were presented at 65 dBA in the free field through a single Tannoy Reveal studio monitor loudspeaker (frequency response: 65 Hz – 20 kHz) at a distance of 1 – 2 feet placed in front of the listener at eye level. Each token was presented once, and listeners subsequently used a computer mouse to select one of two word choices (“heat” or “hit”) on the screen to indicate their perception. There was no time limit on their response, and they were permitted to enable stimulus repetitions up to three times; stimulus repetitions were very rare. Stimuli were presented in blocks organized by degree of spectral resolution (unprocessed, 8-channel or 4-channel). Ordering of blocks was randomized except for the first block, which was always unprocessed. Presentation of tokens within each block was randomized. In this self-paced task, the 245 stimuli were each heard 5 times in each condition of spectral resolution.

Analysis.

Categorical responses were fit using logistic regression, in accordance with recent trends in perceptual analysis (Morrison & Kondaurova, 2009). Listeners’ binary responses (tense or lax) were fit using a generalized linear (logistic) mixed-effects model (GLMM). This was done in the R software interface (R Development Core Team, 2010), using the lme4 package (Bates & Maechler, 2010). A random effect of participant was used, and the fixed-effects were the stimulus factors described above. The binomial family call function was used because the possibility of a “tense” response could not logically exceed 100% or fall below 0%. This resulted in the use of the logit link function, and an assumption that variance increased with the mean according to the binomial distribution. Parameter levels were centered around 0, since the R GLM call function sets “0” as the default level for a parameter while estimating other parameters. For example, since the median duration was 115 ms, a stimulus with duration of 85 ms was coded as -30 ms, and one with duration of 122 ms was coded as +7 ms. All factors and interactions were added via a forward-selection hill-climbing process that began with an intercept-only model. Factors and interactions competed for greatest improvement to the model and were retained if they improved the model’s predictive value without unnecessarily over-fitting. Model improvement was judged according to the Akaike information criterion (AIC) (Akaike, 1974), as it has become a popular method for evaluating mixed effects models (Vaida & Blanchard, 2005; Fang, 2011). This criterion measures relative goodness of fit of competing models by balancing accuracy and complexity of the model. Analysis was similar to that used by Peng, Lu & Chatterjee, 2009); it tested whether the coefficient of the resulting estimating equation for an acoustic cue was different from 0 and, crucially, whether the coefficient for the same cue changed across different conditions of spectral resolution.

Previous literature suggested that 4 or 8 is a suitable number of channels in a noise-band vocoder as a simulation of a cochlear implant. Both of these were tested in this experiment, not for a regression of cue usage against spectral degradation, but instead to find the best proxy value to simulate electric hearing for the problem at hand. Inspection of the psychometric functions of the NH listeners and CI listeners in preliminary data revealed that the 8-channel simulation was the best model of electric hearing, in accordance with previous assessment of better-performing CI listeners (Dorman & Loizou, 1998; Friesen et al., 2001). Additionally, the high amount of variability in the 4-channel condition in this experiment made it difficult to draw firm conclusions about how listeners perceived the signals. A small number of listeners demonstrated non-monotonic effects of spectral degradation on the use of the phonetic cues (i.e. they showed greater use of formant cues in 4-channel compared to 8-channel conditions, but sometimes reported hearing neither the /i/ nor the /I/ vowel), suggesting that reducing the number channels below 8 did not necessarily change the resolution in a meaningful way *vis a vis* this experimental task. In the 4-channel case, the reduced spectral degradation was likely accompanied by increased availability of temporal envelope cues in periodic (voiced) portions (because of increased numbers of harmonics falling into the broader filters), which may have been accessed/utilized differentially by different participants, depending on the precision of their temporal resolution. Some were able to capitalize on this, while some were not. Although (variations in) this ability is an interesting consideration in the use of noise-band vocoded signals, it is outside the scope of this investigation. Subsequent analysis of the data discarded the 4-channel condition, yielding two sets of data models: 1) NH listeners in both listening conditions (unprocessed and degraded using an 8-channel NBV) and 2) CI listeners hearing the unprocessed stimuli.

Results

Identification functions along the three parameter continua are shown in figures 2.3, 2.4 and 2.5. The following models were found to describe the data optimally:

1) *Perception by NH listeners in different conditions:*

Tense ~ Formant + Duration + VISC + SR + Formant:SR + VISC:SR + Duration:SR + (1 | Participant)

2) *Perception by CI listeners:*

Tense ~ Formant + VISC + Duration + (1 | Participant)

For these two models, the interaction between two factors A and B is indicated by A:B. Independent factors are indicated by "+." "SR" refers to spectral resolution (normal or degraded/NBV), and (1|Participant) is a random effect of participant.

For both models, all three main cues were significant (all $p < 0.001$), and interactions between each cue and spectral resolution were also significant for the normal-hearing listeners (all $p < 0.001$). The parameter estimates all went in the predicted direction, and are listed in Table 2.4. Results suggest that when spectral resolution was degraded, normal-hearing listeners' responses were affected less by formants, less by VISC, and more by duration, compared to when spectral resolution was intact. The predicted trend was seen for the cochlear implant listeners (smaller effect of formants and VISC, greater effect of duration, compared to NH listeners hearing the same signals), although direct statistical comparison between these groups was not conducted (to be discussed further in the summary and discussion). Surprisingly, there were no significant interactions between cues for either group. Typically, one would expect the effects of

VISC and duration to be strongest in an ambiguous range of formant values; raw data suggested this, but the interaction did not reach significance in the model.

Although error bars were omitted from the group psychometric functions (Figures 2.3, 2.4 and 2.5), variability in the use of acoustic cues is presented in Table 2.4. This table reveals that participants n01 (the author) and n02, who were both highly familiar with these stimuli, showed the greatest use of durational cues in the unprocessed condition. Naïve listeners might have thus shown a larger interaction between spectral resolution and the vowel duration cue. It is not clear whether their contributions have any consequence on the group interaction, because these listeners also showed high use (3rd and 2nd highest, respectively) of the duration cue in the degraded condition. Additionally, listener n01 showed the second-highest use of the VISC cue in the unprocessed condition. These data suggest that variability exists even within the normal-hearing population, and that familiarity with the stimuli may play a role in the use of acoustic cues for categorization.

Figure 2.3 Effect of vowel formants on tense/lax perception

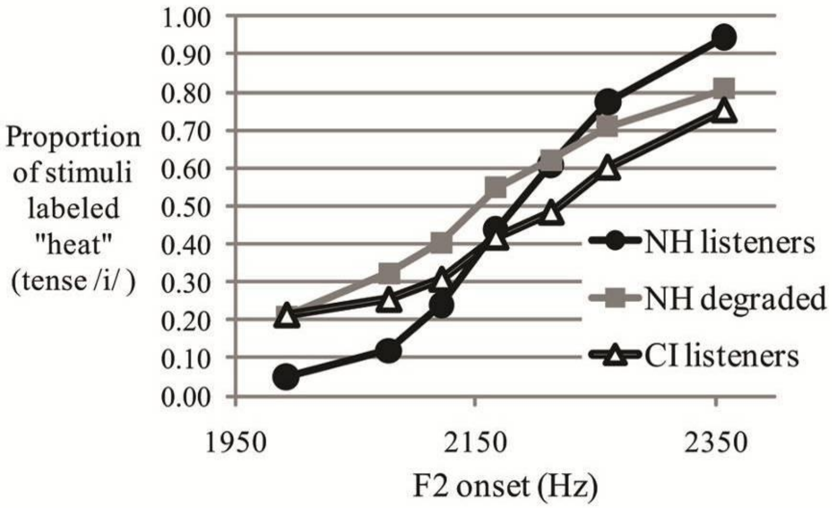

Figure 2.3. Group mean response functions from 15 normal-hearing listeners and 7 cochlear implant listeners along the continuum of vowel formant structure. Although these results are plotted by F2, the other formants were co-varying (see Table 2.2). Error bars were omitted to maintain clarity.

Figure 2.4 Effect of vowel-inherent spectral change (VISC) on tense/lax perception

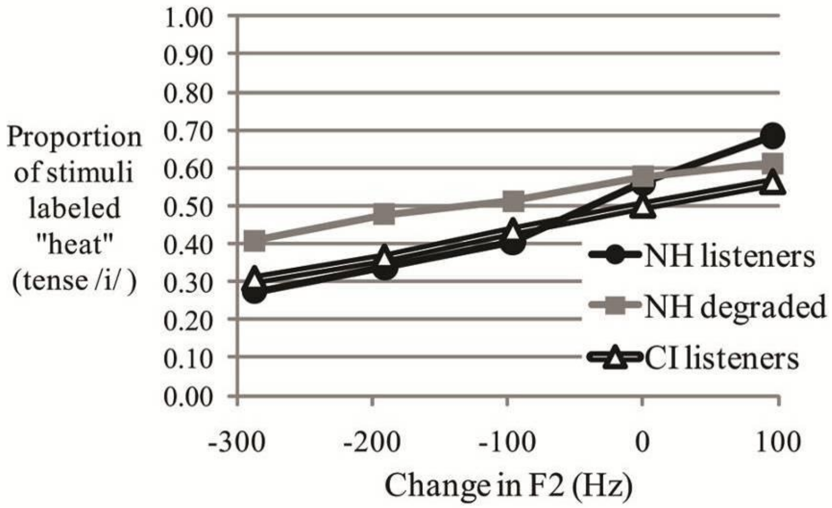

Figure 2.4. Group mean response functions from 15 normal-hearing listeners and 7 cochlear implant listeners along the continuum of vowel-inherent spectral change. Although these results are plotted by change in F2, the other formants were co-varying (see Table 2.2). Error bars were omitted to maintain clarity.

Figure 2.5 Effect of vowel duration on tense/lax perception

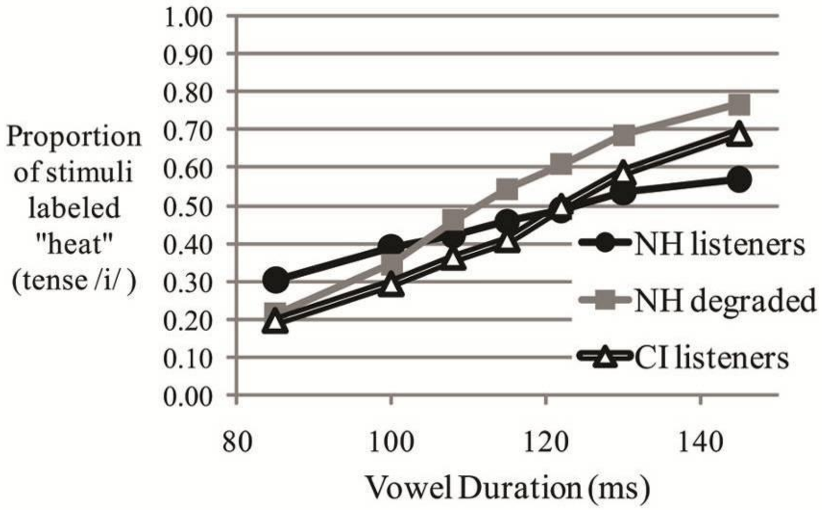

Figure 2.5. Group mean response functions from 15 normal-hearing listeners and 7 cochlear implant listeners along the continuum of vowel duration. Error bars were omitted to maintain clarity.

Although direct statistical comparison was not carried out, the CI listener data are encouraging, as they fall along the same general trend as those from the NH listeners in the simulated conditions. The individual variability is apparently not limited to one group or the other; just as NH listeners have variations in listening strategies, so do the CI listeners, and both groups fall within similar ranges.

Table 2.4.

Parameter estimates and intercepts for listeners in Experiment 1.

	Formants			VISC			Duration		
Group	NH	NBV	CI	NH	NBV	CI	NH	NBV	CI
Est.	0.026	0.011	0.010	0.011	0.004	0.004	0.046	0.061	0.052
Int.	-1.368	-0.22	-0.773	-1.368	-0.22	-0.773	-1.368	-0.22	-0.773
01	0.034	0.014	0.008	0.016	0.007	0.003	0.074	0.091	0.047
02	0.024	0.020	0.015	0.013	0.007	0.007	0.099	0.096	0.040
03	0.028	0.019	0.003	0.012	0.006	0.001	0.041	0.042	0.036
04	0.027	0.015	0.009	0.010	0.005	0.006	0.033	0.045	0.067
05	0.034	0.016	0.015	0.019	0.006	0.006	0.039	0.031	0.056
06	0.025	0.014	0.011	0.012	0.004	0.004	0.038	0.038	0.045
07	0.027	0.013	0.007	0.015	0.004	0.004	0.058	0.049	0.073
08	0.037	0.009		0.012	0.003		0.057	0.067	
09	0.024	0.003		0.008	0.001		0.052	0.049	
10	0.016	0.006		0.010	0.002		0.045	0.040	
11	0.019	0.006		0.013	0.002		0.029	0.050	
12	0.023	0.010		0.006	0.002		0.023	0.069	
13	0.024	0.008		0.003	0.001		0.031	0.097	
14	0.025	0.012		0.007	0.002		0.035	0.085	
15	0.018	0.005		0.004	0.002		0.027	0.065	

Note: Est.= Estimate, Int.= Intercept. Values are derived from the optimal logistic models for Experiment 1. The top portion reflects the group model; data models could be reconstructed using these (centered) variables in an inverse logit equation. Rows in the lower portion reflect parameter estimates from individual listeners within each group.

Because the analyses presented thus far do not speak to perceptual *accuracy* per se, a final analysis was conducted to evaluate the identification of stimuli where all the acoustic cues cooperated to confer a typical “lax” or “tense” vowel. Identification of these stimuli could appropriately be evaluated for correctness. Figure 2.6 illustrates performance levels for all listener groups for the stimuli at the continuum endpoints (which are slightly exaggerated relate to average productions) and also for the penultimate items in the continuum (which correspond more directly to the average production values reported by Hillenbrand et al., in 1995). Results suggest that both the tense and lax vowels were identified reliably by listeners in all conditions; performance was always above 80% and was lowest for tense vowels heard by CI listeners. Essentially, this figure implies that the potentially different perceptual strategies taken by listeners in this experimental task did not necessarily result in substantial differences in identification accuracy.

Figure 2.6 Identification accuracy for tense and lax vowels in Experiment 1.

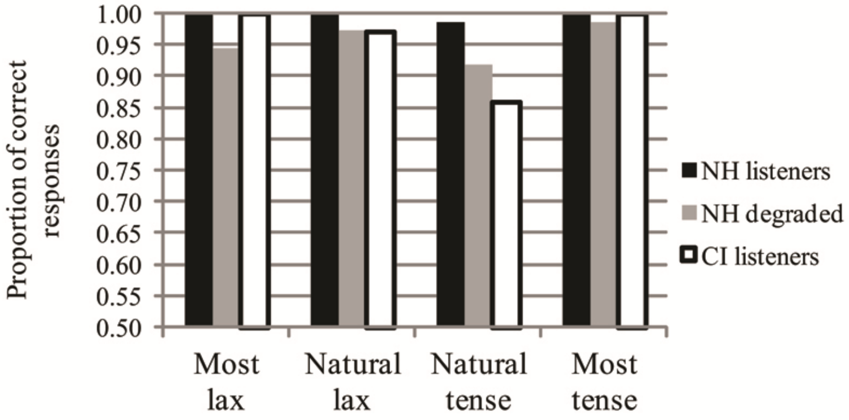

Figure 2.6. Mean accuracy in identification of stimuli at continuum endpoints by different listener groups. "Most" lax items contained cooperating acoustic cues at continuum endpoints, and "Natural" lax/tense contained acoustic cues at penultimate steps in the continua; latter items more closely matched the values identified by Hillenbrand et al. (1995).

Conclusions

In this experiment, listeners were presented with stimuli whose vowels varied along three acoustic dimensions. Normal-hearing listeners heard these stimuli with clear unprocessed spectral resolution and also through 8- and 4-channel noise-band vocoding schemes; the 8-channel condition was a better match to the CI listeners' performance and was thus used for analysis. Cochlear implant listeners heard only the unprocessed stimuli.

Normal-hearing listeners showed decreased use of spectral cues (formant structure and vowel-inherent spectral change), and increased use of vowel duration when spectral resolution was degraded in conditions designed to simulate a cochlear implant. Compared to NH listeners, CI liteners appeared to show less use of spectral cues and greater use of temporal cues; it is possible that the average group differences indicated here (Table 2.4) would reach significance with a larger and more homogenous group of implanted listeners. Although this experiment tests merely one phonetic contrast, it appears to suggest that the NBV simulations hold some predictive value in determining the use of phonetic cues by CI uers.

In view of previous studies using synthesized speech, it is possible, despite the high quality of the speech synthesized by HLSYN, that the role of duration for NH listeners in the unprocessed condition was overestimated. Previous work suggests that duration is largely neglected by NH listeners for this vowel contrast when natural speech quality is preserved (Hillenbrand et al., 2000). Thus, the differences in the use of duration by NH listeners in diferent conditions (and possibly the differences in the use of duration by NH listeners and CI listeners) may be larger than what these data suggest. Another important consideration is the relatively advanced age of the CI user group, which will be discussed later in the Summary and Discussion section.

Experiment 2:
Spectral degradation and the final consonant voicing contrast

Acoustics and perception of the final consonant voicing contrast

A second phonetic contrast was explored to supplement the findings derived from Experiment 1. The second experiment explored the final consonant voicing contrast, which distinguishes /s/ and /z/ in word pairs such as bus-buzz, grace-graze and loss-laws. The cues that contribute to this distinction include (but are not limited to) the offset frequency/transition of the first formant of the preceding vowel, the duration of the preceding vowel, the duration of the consonant, and the amount of voicing (low-frequency energy/amplitude modulation) within that consonant. Vowel duration has received the most consideration in the literature; vowels are longer before voiced sounds than before voiceless ones (House & Fairbanks, 1953; House, 1961). Chen (1970) and Raphael (1972) suggested that this duration difference is an essential perceptual cue for this distinction. However, just as for the aforementioned study by Ainsworth (1972), the limited spectral integrity of Raphael's stimuli (three steady-state synthesized formants) may have caused an over-estimation of the effect of vowel duration. Furthermore, stimuli in Raphael's study that contained vowels of intermediate duration were contrasted reliably by the presence or absence of a vowel-offset F1 transition (F1T). When the F1T appeared at the end of the vowel, listeners reliably heard the following consonant as voiced.

Warren and Marslen-Wilson (1989) also suggested vowel duration to be an essential cue for consonant voicing. Their experiment used a gating paradigm, whereby listeners identify a stimulus (a word) that was truncated in time. This method is problematic for this contrast, however, because it confounds the cues of vowel duration and F1T. When a signal is truncated before the F1T, the duration is shortened *and* the F1T is removed; the independent contributions of each cue are not recoverable in this paradigm. When truncation points fell before the region of the F1T, perception of voicing dramatically declined, but perhaps because of the absence of F1T rather than because of the shortened vocalic duration. Virtually no effect of vowel duration is observed when vowel portions are deleted from the middle (Revoile, 1982) or beginning (Wardrip-Fruin, 1982) of the segment. Only when portions are deleted from the offset (region of F1T) does the perception reliably change from voiced to voiceless (Hogan & Rozsypal, 1980; Wardrip-Fruin, 1982; Hillenbrand et al., 1984; Warren & Marslen-Wilson, 1989). Hillenbrand et al. (1984) noted that compressing the duration of vowels before voiced stops does not significantly alter listeners' perception. Similar findings were reported by Wardrip-Fruin (1982), who showed that a falling F1T signaled voicing across the whole range of vowel durations tested, while syllables without this transition yielded no more than 60% voiced responses even at the longest vowel duration. Summers (1988) suggested that voicing-related F1T differences are not limited to vowel offset; F1 is lower before voiced consonants at earlier-occurring times in the vowel as well. The importance of F1 is also underscored by the results of Hogan and Rozsypal (1980), who observed that excising the vowel offset had a smaller effect on high vowels, for which the F1 is already low and therefore a less useful cue since there is no room for transition.

A meta-analysis by Walsh and Parker (1984) suggests that vowel duration exhibits perceptual consequence only for "artificial or abnormal circumstances." For example, Revoile (1982) showed that vowel duration was used as a voicing cue by individuals with hearing impairment, but not those with normal hearing. Wardrip-Fruin (1985) observed vowel duration effects for words presented in low-pass filtered noise, but not in quiet (1982). In experiments by Nittrouer (2004; 2005), vowel duration served as a voicing cue for synthetic speech, but this effect was strongly reduced and overpowered by the F1T cue when natural speech tokens were used. Thus, just as for previous experiments with vowels, the effect of duration on perceptual judgments appears to be driven at least partly by spectral fidelity of the signal.

Not surprisingly, there are acoustic cues that correspond to the voicing contrast within the fricative consonant itself. Voiceless fricatives are longer than voiced ones (Denes, 1955; Haggard, 1978), further increasing the vowel:consonant duration ratio (VCR) for voiced fricatives. VCR and duration of voicing within the fricative noise were shown by Hogan and Rozsypal (1980) to be reliable cues for perception of voicing in sounds in an experiment where extension of vowel duration by itself did not force a change in voicing perception. Voicing during the consonant is not thought to be essential for perception of the voicing feature, since voiced fricatives are routinely devoiced in natural speech (Klatt, 1976; Haggard, 1978). Listeners reliably perceive voicing despite this apparent omission (Hogan & Rozsypal, 1980). Perception of this cue should be asymmetrical; the presence of voicing is inconsistent with voiceless sounds, while the lack of voicing is consistent with either /s/ or /z/ (Smith, 1996).

There are even more cues to the /s/-/z/ contrast than are discussed here (some are mentioned in the Method section of this experiment), but the aforementioned cues have been given the most consideration in the literature, and are thought to play a crucial role in perception of this contrast. The second experiment in this chapter was designed to assess the use of these acoustic cues in listening conditions similar to those used in Experiment 1.

Hypotheses

It was hypothesized that for listeners with CIs or NH listeners in CI simulations, the F1 transition cue would be used less, and the durational cues (vowel and consonant duration, or a ratio of the vowel and consonant durations) would be used more, compared to listeners with normal hearing. Even though consonant voicing is implemented in the spectral domain (via low-frequency energy), it is also implemented in the temporal domain (as temporal amplitude modulations of varying duration). Therefore, the voicing cue was hypothesized to be used more by CI listeners and by NH listeners in CI simulations.

Method

Participants.

Participants for Experiment 2 were comprised of 11 adult (ages 18-37; average 28.9 years) listeners with normal-hearing, defined as having pure-tone thresholds ≤20 dB HL from 250–8000 Hz in both ears (ANSI, 2010) and 7 CI listeners whose demographics were the same as those for Experiment 1 (see Table 2.1). Four of the NH listeners and all 7 CI listeners also participated in Experiment 1. Normal-hearing participants 01 (the first author) and 02 were highly familiar with the stimuli, having been involved in pilot testing and the construction of the materials.

Stimuli

Natural speech manipulation.

Stimuli for the second experiment were constructed using modified natural recordings of the words “loss” and “laws” spoken by a male native speaker of American English. These words were recorded in a double-walled sound-treated room using an AKG C1000 microphone at 44.1 kHz sampling. The stimulus set consisted of 126 items that varied in four dimensions: presence/absence of vowel-offset falling F1 transition (2 levels), vowel duration (7 levels), duration of fricative (3 levels), and duration of voicing within that fricative (3 levels). See Table 2.5 for a detailed presentation of the levels for each parameter. A single /l/ segment was chosen as the onset of all stimuli in the experiment, to neutralize it as a cue for final voicing (see Hawkins & Nguyen, 2004). The non-high vowel in “laws” was chosen because the F1 transition cue present in low vowels has been hypothesized to be compromised or absent in high vowels (Summers, 1988). The vowel was segmented from a recording of “laws,” and thus contained a "voiced" F1 offset transition from roughly 635 Hz at vowel steady-state to 450 Hz at vowel offset, which is in the range of transitions observed in natural speech by Hillenbrand et al. (1984). A "voiceless" offset transition was created by deleting the final five pitch periods of the vowel in "laws," (maintaining a flat 635 Hz F1 offset) and expanding the duration to the original value using the pitch synchronous overlap-add (PSOLA) function in the Praat software (Boersma & Weenink, 2010). Rather than using recordings from “loss” and “laws” separately, this manipulation was preferable, in order to maintain consistent volume, phonation quality and other cues that may have inadvertently signaled the feature in question. In other words, it permitted the attribution of influence directly to the

F1 offset level, since earlier portions of the vowel were consistent across different levels of this parameter. A uniform decaying amplitude envelope was applied to the final 60 ms of all vowels, as in Flege (1985); it resembled a contour intermediate to those observed in the natural productions, and was used to neutralize offset amplitude decay as a cue for voicing (see Hillenbrand et al., 1984). Vowel durations were manipulated using PSOLA to create a 7-step continuum between 175ms and 325ms, based on values from natural production reported by House (1961) and Stevens et al. (1992), and used by Flege (1985) in perceptual experiments. All vowels were manipulated using PSOLA to contain the same falling pitch contour (which started at 96 Hz and ended at 83 Hz), to neutralize pitch as a cue for final fricative voicing (see Derr & Massaro, 1980; Gruenenfelder & Pisoni, 1980). A 250 ms segment of frication noise was extracted from a natural /s/ segment. An amplitude contour was applied to the fricative offset to create a 50 ms rise time and 30 ms decay-time. Two other durations (100 and 175 ms) of frication noise were created by applying the offset envelope at correspondingly earlier times. The resulting values ranging from 100 - 250 ms frication duration resembled those used by Soli (1982) and Flege and Hillenbrand (1985). Voicing was added to these fricatives by replacing 30 or 50 ms onset portions with equivalently-long onset portions of a naturally-produced voiced /z/ segment. These three levels of voicing thus varied in the range of 0 - 50 ms, which resembles the range used in perceptual experiments by Stevens et al (1992), who reported that phonetically voiced fricatives typically have at least 30ms of voicing. It should be noted that voiced fricatives are often phonetically de-voiced in English, but native-English speaking listeners can reliably distinguish /s/ from devoiced /z/ (Stevens, 1992; Smith, 1996). These fricatives were appended to all 14 of the aforementioned vowel segments with onset /l/. For fricatives with onset voicing, the first pitch period of voiced fricative noise was blended with the last pitch period of the vowel (each at 50% volume) to produce a smooth transition between segments. Although the stimuli were not designed explicitly to vary the vowel-consonant duration ratio, this ratio naturally changed as a function of each independently varied duration factor.

Table 2.5

Acoustic parameter levels defining the four factors in Experiment 2.

First formant offset (Hz)	450	615					
Vowel Duration (ms)	175	200	225	250	275	300	325
Consonant Duration (ms)	100	175	250				
Voicing Duration (ms)	0	30	50				

Note: Each cue was varied orthogonally.

Spectral degradation: Noise-band vocoding.

Noise-band vocoding was accomplished using the same procedure described for Experiment 1 (described earlier), except that the upper-limit of the analysis and filter bands was changed from 6 kHz to 7 kHz, to ensure that a substantial amount of frication noise was represented within the spectrally-degraded output. Analysis/carrier band cutoff frequencies for Experiment 2 are displayed in Table 2.6.

Table 2.6

Specification of analysis and carrier filter bands for the noise-band vocoding scheme for Experiment 2.

Channel:	1	2	3	4	5	6	7	8
High-pass (Hz)	141	283	495	812	1285	1994	3052	4634
Low-pass (Hz)	283	495	812	1285	1994	3052	4634	7000

Procedure.

The procedure for Experiment 2 was the same as that for Experiment 1 (described earlier), except that the word choices were different in label and in number. Visual word choices were "Loss" and "Laws," and the 126-item stimulus set was presented in alternating blocks of unprocessed and 8-channel noise-band vocoder conditions. In view of the results of the first experiment, no 4-channel NBV condition was used for Experiment 2. The 126 stimulus items were heard by NH listeners 5 times in both conditions of spectral resolution. Cochlear implant listeners only heard the natural (unprocessed) items 5 times each.

Analysis.

Listeners' binary responses (voiced or voiceless) were fit using a generalized linear (logistic) mixed-effects model (GLMM), using the same procedure as in Experiment 1. This experiment produced two sets of data: 1) NH listeners in both conditions of spectral resolution, and 2) CI listeners listening to the modified natural sounds with intact spectral resolution.

Results

Identification functions along the four parameter continua are shown in figures 2.7, 2.8 2.9 and 2.10. Although vowel:consonant duration ratio was not explicitly planned in stimulus construction, it was easily calculated and included as a separate factor in the model (since this factor was not fully crossed with the others, listeners' responses were not plotted for this cue). The following models were found to describe the data optimally:

1) *Perception by NH listeners in different conditions:*

Voiced ~ VCRatio + F1T + VDuration + Voicing + SR + F1T:SR + VCRatio:VDuration + Voicing:SR + CDuration + VCRatio:Voicing + VDuration:SR + (1|Participant)

2) *Perception by CI listeners:*

Voiced ~ VDuration + CDuration + Voicing + F1T + VDuration:F1T + F1T:Voicing + VDuration:Cduration + CDuration:F1T + VDuration:CDuration:Voicing + (1|Participant)

For these two models, the interaction between two factors A and B is indicated by A:B. Independent factors are indicated by "+." "VCRatio" refers to the ratio of vowel duration to consonant duration. "SR" refers to spectral resolution (normal or degraded/NBV), and (1|Participant) is a random effect of participant. Predictors are listed in the order in which they were added to the model (this was determined by the AIC metric). Parameter estimates for the groups and for each participant are listed in Tables 2.6 and 2.7.

For the NH listener model, all five main factors (including VCRatio) were significant (all $p < 0.001$), although consonant duration was by far the least powerful main factor in the model, according to the AIC metric. Spectral resolution significantly interacted with F1 Transition (p < 0.001; F1 transition was a weaker cue in the degraded condition), and with voicing duration (p < 0.001; voicing duration was a weaker cue in the degraded condition), but not with VCRatio nor consonant duration. The interaction between spectral resolution and vowel duration did not reach statistical significance (p = 0.11; vowel duration was a slightly stronger cue in the degraded condition), but its inclusion improved the model according to the AIC metric. The effect of VCRatio changed slightly in the expected direction in the degraded condition, but this interaction did not reach statistical significance (p = 0.60), and was not included in the model. Although the raw data suggested that the interaction between consonant duration and spectral resolution would go in the expected direction, the model did not confirm this; it did not reach significance (p = 0.42), and was not included in the model. There were significant interactions between VCRatio and vowel duration (p < 0.001), and between VCRatio and voicing duration (p = 0.005), indicating a complex inter-dependence of multiple cues for this contrast.

CI listeners were able to use the F1 transition cue (p < 0.001), but apparently not to the same extent as NH listeners (the parameter estimate was lower for the CI group). CI listeners showed use of the vowel duration cue (p < 0.001) that appears to be greater than that by NH listeners (the parameter estimate was higher for the CI group). The effect of consonant duration was significant (p < 0.001) and appears to be similar to that observed in the NH group. F1 transition significantly interacted with vowel duration (p < 0.001), with voicing duration (p < 0.001) and with consonant duration (p = 0.017). Although the effect of VCRatio did not reach significance, vowel duration significantly interacted with consonant duration ($p = 0.019$). There was a three-way interaction between vowel duration, consonant duration and voicing duration that did not reach significance ($p = 0.15$), but its inclusion produced a significant improvement in the model, according to the AIC metric. Just as for NH listeners, the hearing-impaired listeners showed complex inter-dependence of cues for this contrast. A large amount of variability was seen in the CI listener group for all cues (Table 2.7), especially for voicing duration and VCRatio, where several individuals' parameter estimates actually went in the reverse direction. Variability also presented in the NH group (Table 2.7), suggesting that there may not be a uniform pattern of perception even for listeners with normal hearing.

Figure 2.7 Effect of F1 transition cue on voicing perception in Experiment 2

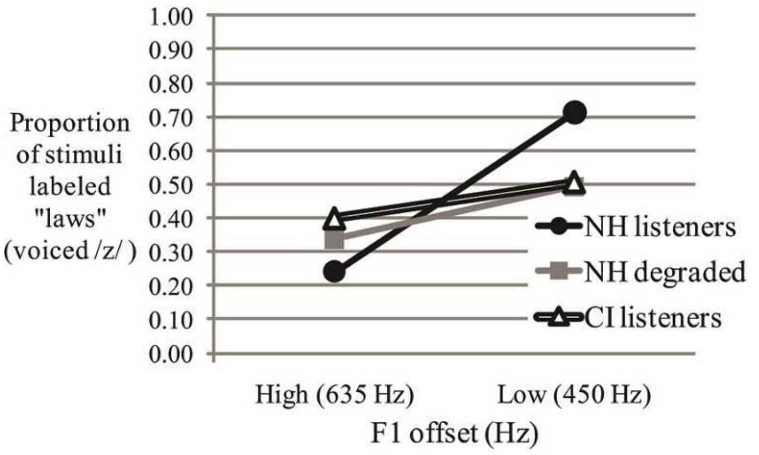

Figure 2.7. Group mean response functions from 11 normal-hearing listeners and 7 cochlear implant listeners for both levels of the F1 transition offset. Greater disparity of responses between levels indicates greater influence of this cue in perceptual responses. Error bars were omitted to maintain clarity.

Figure 2.8 Effect of vowel duration on voicing perception in Experiment 2

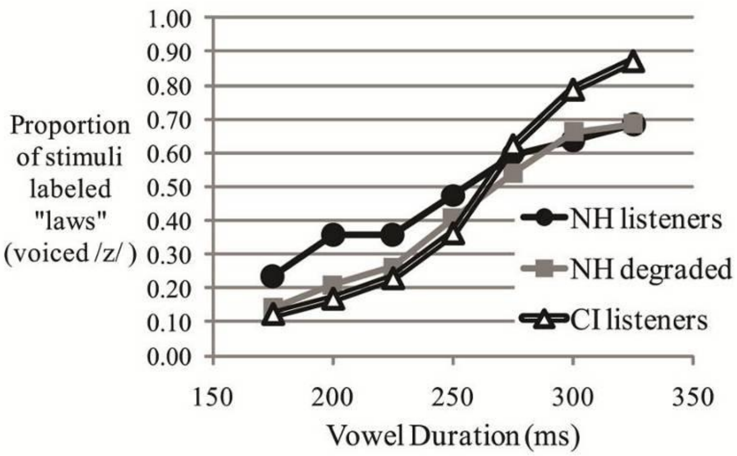

Figure 2.8. Group mean response functions from 11 normal-hearing listeners and 7 cochlear implant listeners along the continuum of vowel duration. Error bars were omitted to maintain clarity.

Figure 2.9 Effect of consonant duration on voicing perception in Experiment 2

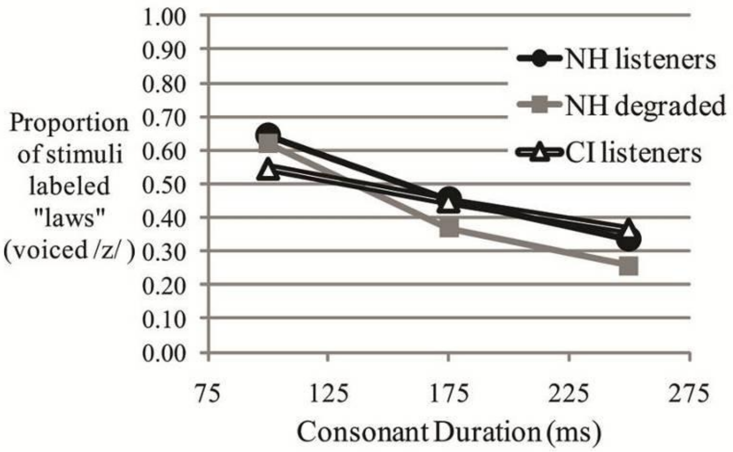

Figure 2.9. Group mean response functions from 11 normal-hearing listeners and 7 cochlear implant listeners along the continuum of consonant duration. Error bars were omitted to maintain clarity.

Figure 2.10 Effect of consonant voicing duration on voicing perception in Experiment 2

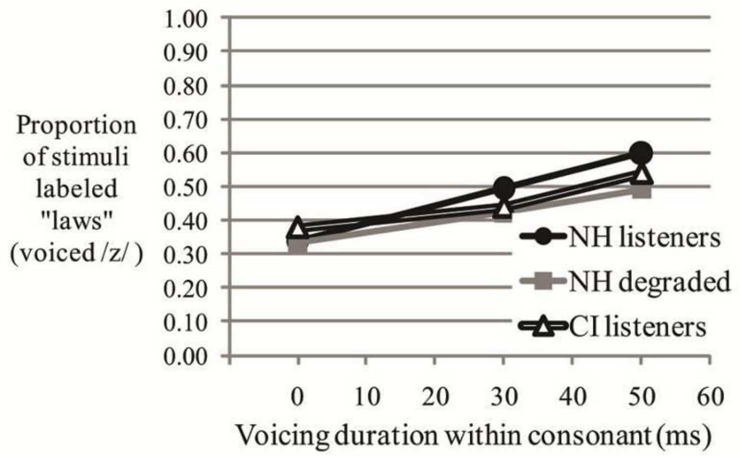

Figure 2.10. Group mean response functions from 11 normal-hearing listeners and 7 cochlear implant listeners along the continuum of consonant voicing duration. Error bars were omitted to maintain clarity.

Table 2.7a.

Intercepts and parameter estimates for the optimal logistic models for Experiment 2.

		Vowel Duration			F1 Transition	
Group	NH	NBV	CI	NH	NBV	CI
Est.	**0.015**	**0.015**	**0.032**	**0.017**	**0.005**	**0.003**
Int.	0.213	-0.163	0.295	0.213	-0.163	0.295
01	0.053	0.066	0.030	0.030	0.005	0.011
02	0.018	0.022	0.075	0.028	0.010	0.030
03	0.025	0.025	0.027	0.022	0.007	-0.001
04	0.057	0.058	0.104	0.020	0.003	0.011
05	0.015	0.018	0.079	0.019	0.006	0.016
06	0.046	0.034	0.101	0.027	0.008	0.026
07	0.032	0.030	0.014	0.027	0.015	0.007
08	0.024	0.023		0.014	0.005	
09	0.040	0.061		0.018	0.005	
10	0.021	0.017		0.007	0.001	
11	0.018	0.025		0.012	0.004	

Table 2.7 (continued)

		Voicing Duration			Consonant Duration	
Group	NH	NBV	CI	NH	NBV	CI
Est.	**0.037**	**0.018**	**0.020**	**-0.008**	**-0.006**	**-0.011**
Int.	0.213	-0.163	0.295	0.190	-0.137	-0.295
01	0.128	0.063	-0.083	-0.055	-0.022	-0.028
02	0.055	0.037	-0.144	-0.007	-0.008	0.019
03	0.081	0.045	-0.049	0.005	-0.003	-0.025
04	0.045	0.014	-0.011	-0.008	0.008	0.044
05	0.063	0.044	0.102	-0.007	-0.002	-0.061
06	0.054	0.047	0.045	-0.014	-0.001	0.027
07	0.072	0.030	0.003	-0.020	-0.004	-0.038
08	0.059	0.045		0.000	-0.017	
09	0.032	0.015		-0.023	-0.004	
10	0.010	0.007		-0.003	-0.010	
11	0.027	0.006		-0.005	0.009	

		V:C Ratio	
Group	NH	NBV	CI
Est.	**0.783**	**0.910**	**-0.036**
Int.	0.190	-0.137	-0.307
01	-1.36	1.30	-0.48
02	1.47	1.45	-0.18
03	2.32	1.08	-0.33
04	0.35	1.82	-0.10
05	0.95	1.09	1.15
06	2.71	4.11	-0.30
07	1.30	2.30	0.17
08	1.31	0.10	
09	-0.14	1.28	
10	0.46	-0.22	
11	0.56	1.19	

Note: In each table, the top portion reflects the group model; data models could be reconstructed using these variables in an inverse logit equation. Rows in the lower portions reflect parameter estimates from individual listeners within each group. NH and NBV refer to normal-hearing listeners in the unprocessed and degraded (8-channel noise-band vocoded) conditions, respectively. CI refers to CI listeners. NH and NBV intercepts for VCRatio and Consonant duration were derived from a separate model where they could be individually computed with an interaction with spectral degradation.

Because the analyses presented thus far do not speak to perceptual *accuracy* per se, a final analysis was conducted to evaluate the identification of stimuli where all the acoustic cues cooperated to confer a typical "voiceless" or "voiced" fricative. Identification of these stimuli could appropriately be evaluated for correctness. Figure 2.11 illustrates performance levels for all listener groups for the stimuli at the continuum endpoints. Results suggest that both the voiceless and voiced fricatives were identified reliably by listeners in all conditions; performance was always above 90% and was lowest for voiceless fricatives heard by CI listeners. Essentially, this figure implies that the potentially different perceptual strategies taken by listeners in this experimental task did not necessarily result in differences in identification accuracy.

Figure 2.11. Identification accuracy for /s/ and /z/ sounds in Experiment 2

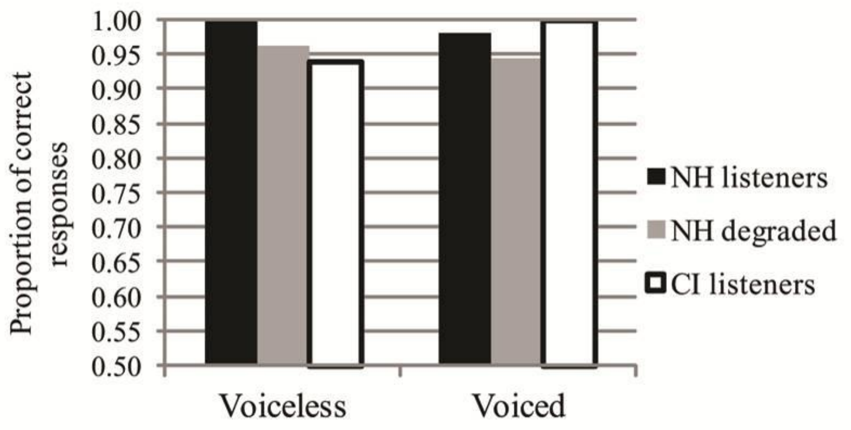

Figure 2.11 Mean accuracy in identification of stimuli at continuum endpoints by different listener groups. Voiceless and voiced items for this analysis were limited to those where all cues cooperated appropriately at continuum endpoints to signal the same feature.

Conclusions

In Experiment 2, listeners were presented with stimuli that varied along four acoustic dimensions. In conditions that are thought to roughly simulate the use of a CI, NH listeners maintained use of all four cues, but showed decreased use of the F1 transition and consonant voicing cues. Reliance upon vowel duration did not change significantly when the resolution was degraded. The effect of vowel duration for NH listeners was larger than what was expected based on previous literature (perhaps because early-occcuring spectral informayion in the vowel was neutralized).

Statistical comparisons were not made between NH and CI listeners, but the data suggest that CI listeners on average made less use of the F1 transition and consonant voicing cues, and made more use of the vowel duration cue. These results are in agreement with Experiment 1 in that listeners altered their use of phonetic cues when spectral resolution is degraded, and that CI listeners may use phonetic cues differently than NH listeners. It should be noted again however, that the use of the F1 transition cue is probably dependent on the vowel environment. The F1 cue would be less useful for consonants following the /i/ or /u/ vowels; the F1 value in these segments is already low, so any F1 movement would be subtle, if at all present (Hogan & Rozsypal, 1980; Hillenbrand et al., 1984). It is thus possible that durational cues are already more dominant in these high vowel contexts, resulting in no need for altered lisstening strategies in degraded conditions.

Summary and Discussion of Experiments 1 and 2

In these experiments, listeners categorized speech tokens that varied in multiple dimensions. The influence of each of those dimensions was modulated by the degree of spectral resolution with which the signal was delivered, or by whether the listener used a cochlear implant. The following general conclusions emerged:

(1) As spectral resolution is degraded, spectral cues (such as formant structure, vowel-inherent spectral change, and a vowel-offset formant transition) played a smaller role, and some temporal cues played a larger role in normal-hearing listeners' phonetic identifications.

(2) Cochlear implant listeners appeared to show less use of spectral cues, and greater use of temporal cues for phonetic identification (Tables 2.7, 2.8), compared to normal-hearing listeners. This effect was more pronounced for the final consonant voicing contrast than for the lax / tense vowel contrast.

(3) There was a high amount of variability in the individual data; some NH listeners showed different use of cues in degraded conditions while others did not. Similarly, some CI listeners showed patterns similar to those in the NH group, while others showed distinctively different patterns. It is not yet known whether either of these patterns can be associated with more general success in speech perception.

(4) Under conditions of normal redundancy of acoustic cues, a NH listener and a CI user can thus achieve the same or very similar performance on speech recognition tasks (Figures 2.6 and 2.11) (word recognition, phoneme recognition, confusion matrix/information transfer analysis), but by use of different cues.

It should be noted that there are various limitations in the generalization of the CI simulations to real CI listeners. Among these are 1) noise-band vocoding is only a crude approximation of the experience of electric hearing, 2) the two-alternative forced-choice task is an atypical listening scenario, lacking top-down influences such as contextual clues and visual information, which could resolve perceptual ambiguity, 3) the NH participants listening to the simulations were generally much younger than the CI listeners, and 4) the simulated conditions are essentially simulating an initial activation rather than an everyday experience; most NH listeners in this experiment had no prior experience with noise-band vocoding, whereas the CI listeners had all been wearing their devices for multiple years. It is thus possible that the degraded conditions did not simulate the eventual everyday performance of a CI listener, who has an opportunity to adapt to the degraded spectral resolution, as well as the opportunity adapt to spectral shifting (Rosen, 1999). Another important consideration is the dramatic age difference between the NH and CI groups. This is known to be associated with poorer temporal processing for basic psychophysical tasks (Gordon-Salant & Fitzgibbons, 1993; 1999), and tasks involving perception of temporal phonetic cues (Gordon-Salant, Yeni-Konshian, Fitzgibbons & Barrett 2006); this will be discussed at greater length in Chapter 5. Notably, the direction of the expected age effect is opposite to that which was observed in the experiments. One would expect that the temporal phonetic cues would be less influential for the older listeners, since they should be perceived less well. On the contrary, the CI listeners made greater use of the temporal cues despite being of more mature age than the NH listener group. The one older NH listener (n04 in Experiment 1) does not provide sufficient basis for age-matched group comparison, but it is reassuring that this listener's data were not markedly different from the NH group mean (Table 2.4).

The decomposition of phonetic perception into weighting of multiple cues represents a level of analysis that has been largely unexplored for hearing-impaired listeners. In Chapter 4, an even finer-grained analysis will explore not only the use of acoustic cues, but the dynamic use of cues that is driven by various auditory and visual contexts. It is hoped that by exploring multiple levels of speech perception analysis, the normal and impaired auditory system can be understood more completely.

Chapter 3: Effects of noise and limited bandwidth on phonetic cue perception

Introduction and motivation

Consonant voicing contrasts are nearly ubiquitous in the world's languages (Ladefoged & Maddieson, 1996) and the perception of acoustic cues underlying these contrasts has been explored thoroughly for normal-hearing listeners. Relatively less is known about how voicing perception is accomplished by individuals with hearing impairment or individuals in adverse conditions like background noise. It is clear, however, that the voicing contrast is very robust to such adverse constraints. Miller and Nicely's (1955) classic study of phonetic confusion patterns in noise has been replicated many times in various conditions, and results consistently suggest that the voicing feature of phonemes is robust to signal degradations such as background noise (Wang & Bilger, 1973; Phatak & Allen, 2007; Phatak, Lovitt & Allen, 2008), hearing impairment (Bilger & Wang, 1976), spectral degradation (Shannon et al., 1995; Xu et al, 2005) or cochlear implantation (Friesen et al, 2001). It is thus often stated that the amount of "information transfer" is high for the voicing feature relative to other consonant features. This finding is so consistent that some studies dispense with potential voicing confusions in the very design of the experiment (Dubno & Levitt, 1981). In view of the constraints facing listeners in adverse conditions or listeners with hearing impairment, it is not apparent that these individuals perceive the voicing contrast using the same perceptual strategies used by normal hearing listeners. The investigation in this chapter was designed to assess the use of acoustic cues that drive the stop consonant voicing contrast in conditions that simulate the experience of hearing loss and/or background noise.

Speech perception in noise

It is well known that word recognition is poorer in the presence of background noise than in quiet. This problem is paramount for manufacturers of hearing aids, cochlear implants, and also designers of automatic speech recognition systems. Confusions in speech perception can be understood at various levels of analysis, including the sentence level, word level, phoneme level, and the phonetic feature level. The phonetic feature level refers to linguistic categories such as place, manner and voicing, which describe components of individual segments. Features are conveyed in the acoustic signal (speech) by multiple co-occurring cues, any of which could potentially be used to differentiate segments. This investigation attempts to explore the underlying acoustic cues that are used by listeners to recover the phonetic feature of voicing.

Miller and Nicely (1955, henceforth MN55) explored masked and filtered consonant confusions using a live female voice masked by white noise. Confusions were analyzed using information transfer analysis, a method to calculate the amount of independent phonetic feature information recovered by the listener. For this method, not all confusions are treated equally; confusion of /d/ (perception) for [b] (actual sound) implies that the features of voicing and manner of articulation were correctly perceived, since only place was incorrect; confusion of /d/ for [p] implies that only manner was correctly perceived, since place and voicing were incorrect. Most errors in MN55 occurred on the consonant place of articulation, consistent with acoustic and perceptual analyses showing that place contrasts rely heavily on spectral information in the high frequencies (which was masked heavily by the noise in this study). Across the various noise conditions in this study, it was found that the voicing feature was rather robust, being detectable above chance levels even at -12 dB SNR, and consistently recovered more accurately than the other features (such as place and manner of articulation).

Information transfer analysis has been used to explore perception of speech in various listening conditions. Wang and Bilger (1973) expanded on MN55 by incorporating consonants in both the initial and final positions in words, and by using multiple vowel environments. In general, normal-hearing (NH) listeners were most successful at identifying consonants followed by /u/, and least successful at identifying consonants followed by /i/. Consonants preceded by /i/ were easiest; those preceded by /*a*/ were most difficult. Again, place contrast errors were numerous, but voicing errors were uncommon. These authors found a similar pattern of errors for listeners with hearing impairment (Bilger & Wang, 1976). Although levels of performance were highly variable for hearing-impaired (HI) listeners, the relative contributions of phonetic features were very stable. Voicing was consistently transmitted at a high rate, second only to sibilance (which for some listeners might be inaudible because of high-frequency loss; it would be inaudible in other studies that use low-pass filtering with NH listeners).

More recent studies (Phatak & Allen, 2007; Phatak et al., 2008) have revealed that the spectrum of the noise masker has implications for the types of phonetic feature errors. In these studies, the series of single-SNR confusion matrices (used in earlier studies) was transformed into phoneme-specific confusion patterns, which reveal not only particular phonetic confusions, but their non-uniform susceptibility to noise, and the rate at which particular error patterns are ameliorated as the SNR becomes more favorable. A comparison of Phatak and Allen (2007) with Phatak et al. (2008) showed that consonants were more heavily masked by white noise than speech-spectrum noise, if SNR was equated. Consistent with earlier studies, place of articulation confusions dominated the errors. Speech-shaped noise led to relatively more confusions of voicing (especially among fricatives), owing to the greater power of energy at lower frequencies, which can be confused for low-frequency voicing energy. The sibilant phonemes /s, z, ʃ, ʒ/ and /t/ all have the greatest advantage in speech-shaped noise, since their strong high-frequency energy remains audible in the bass-heavy long-term average speech spectrum. Thus, there are some asymmetries in phonetic perceptions in steady-state noise that arise predictably from speech acoustics.

Speech can also be masked by one or more competing talkers. Since speech is an amplitude-modulated signal, the SNR of speech against a competing talker is dynamic, replete with short opportunities to “glimpse” the target during brief dips in masker volume. Because the focus of the current study is on acoustic cues that are tightly constrained in time (i.e. the cue is found at a very specific point in the syllable), running speech will not be a focal point of the discussion, since the audibility of that precise timepoint would not be under rigorous control. It could be presumed that success on a perceptual task could be predicted at least partially by the mere placement of an amplitude peak or valley concurrent with the acoustic cue in question. It should be noted, however, that listeners with hearing impairment are less able to capitalize on these short-term “glimpses” of the target signal when the masker is momentarily quieter (Carhart & Tillman, 1970; Festen & Plomp, 1990). To simplify analysis in this experiment, steady-state maskers were used.

The role of bandwidth

In MN55, the effects of low-pass and high-pass filtering were explored at various SNRs. The rarity of voicing errors in noise persisted even when the bandwidth was severely limited. For example, with a favorable SNR of +12 dB, voiceless stops were confused with their voiced cognates less than 3% of the time, even when the signal was filtered into a narrow spectral band between 200 and 300 Hz (this error rate was less than 2% when the low-pass was raised to a mere 400 Hz). Place-of-articulation errors were abundant in low-passed conditions (particularly for /t/), even as the upper limit of this band was increased to 1200 Hz, consistent with confusion patterns by listeners with hearing impairment (Dubno, Dirks & Langhofer, 1982). Only when the low-pass cutoff was extended to 5000 Hz was /t/ reliably identified. When the signal was high-pass filtered, performance for /t/ was still extremely good (even for the narrow 4500-5000 Hz bandpass filter), but performance for /p/ suffered. Thus, although /p/ and /t/ share common phonological/articulatory traits (manner of articulation, voicing), these asymmetries in error patterns highlight potentially different cues necessary for their perception.

Oxenham and Simonson (2009) have shown that perception of words in sentences can be extremely good when low-passed below 1500 or high-passed above 1200 Hz, even with a modest amount of masking noise. Despite these findings, few would concede that a spectral band between 1200 and 1500 Hz is the essential key to speech perception. For example, the Articulation Index (French & Steinberg, 1947), Speech Transmission Index (Houtgast & Steeneken, 1971) and Speech Intelligibility Index (ANSI, 2007) all suggest that acoustic information between 1500 and 3000 Hz is particularly important for speech understanding (although frequency-importance bands for running speech can be notably different than those for individual syllables). In spite of this, the spectral regions below 1500 Hz still contain acoustic cues redundant with those in the higher-frequency regions (contrary to the assumption of many indices that assume frequency-band independence). It is this redundancy that may prove to be invaluable to listeners with hearing loss that compromises the high-frequency acoustic spectrum.

Listeners face an especially challenging problem when encountering both masking noise and limited bandwidth. As Oxenham and Simonson (2009) imply, the signal redundancy necessary to perform with either constraint is counteracted by the inclusion of the second constraint. Stuart, Phillips and Green (1995) compared word recognition in continuous and interrupted noise by NH listeners with and without a simulated hearing loss (created by low-pass filtering). The experimenters measured masking release, which is the improvement in speech perception performance gained when a steady-state masker is modulated in amplitude (or is replaced by single-talker speech, which is also modulated in amplitude). This can be measured in terms of word recognition performance or SNR difference. Listeners in Stuart et al.’s (1995) experiment showed reduced masking release (less benefit of amplitude modulation) when the speech was low-pass filtered, leading the authors to suggest that the residual low-frequency hearing used by these listeners and HI listeners is insufficient for masking release. This effect was explored further by Scott, Green and Stuart (2001), who found that masking release benefit declined as the speech was delivered with progressively lower frequency cutoffs on the low-pass filters. With a LPF of 1000 Hz, listeners showed especially poor performance at all SNRs tested, and failed to reach 50% word recognition even at +10 dB

SNR. Similar results were found by Nilsson, Soli and Sullivan (1994) in their development of the Hearing in Noise Test (HINT); words in sentences were more heavily masked by modest amounts of noise when bandwidth was limited. In particular, when spectral information above 2500 Hz was eliminated, speech recognition thresholds (SRTs) were elevated by about 3 dB; when bandwidth was further limited to 1000 Hz, SRTs rose another 7 – 9 dB.

At the segmental level, the combination of noise and filtering appears to be more detrimental than filtering alone. In a sequence of studies using HI listeners with and without presumed cochlear dead regions, Vickers, Moore and Baer (2001) and Baer, Moore and Kluk (2002) showed that consonant errors in band-limited conditions were more numerous in noise than in quiet. While perception of manner and place of articulation were dramatically poor in both conditions, perception of the voicing feature was relatively good (70%) in both conditions (even with a low-pass filter of 800 Hz), highlighting its robustness to both of these adversities. Thus, either the cues necessary for voicing perception are found in the lower frequency region and/or listeners use some high-frequency cues in more favorable listening conditions, and switch to a different kind of cue in the presence of these adverse constraints. The latter explanation would be consistent with masking release gained by listeners for both low-pass and high-pass filtered speech (Oxenham & Simonson, 2009).

The use of F0 in noise

There are several lines of evidence suggesting that fundamental frequency (F0) is an especially important cue for listening in noise. Laures and Wiesmer (1999) explored the use of F0 contour in noise in a task designed to model dysarthric speech. When target sentence F0 contours were flattened, word recognition subsequently declined, and subjective intelligibility was purported to drop dramatically. As the mean F0 of competing talkers becomes increasingly disparate, they become easier to perceptually segregate (Brokx & Noteboom, 1982). Even at the segmental level, the F0 cue can be useful; McAdams (1989) showed that F0 modulation in one of two simultaneously-presented vowels can facilitate perceptual segregation.

A more thorough test of F0 benefit in noise was conducted by Binns and Culling (2007), who tested listeners in the presence of speech-shaped noise or a competing talker. Sentences were presented with their natural F0 contours, contours inverted around the mean, or with artificially flattened contours (like those used by Laures & Wiesmer). When the masker was wideband noise, the effects of F0 manipulation were negligible, compared to those for signals masked by a competing talker. In the competing talker condition, SRTs for target sentences with unmodified (natural) F0 contours were 2 dB better than those with flattened contours. SRTs with inverted contours were 3.8 dB worse than those for the unmodified sentences, suggesting that that it is not merely F0 variation, but *correct* or *appropriate* F0 variation that drove the F0 contour benefit. This was corroborated by Miller, Schlauch and Watson (2010), who showed that exaggerated F0 contours led to poorer performance than with the natural contours. In that study, poor performance was observed in conditions with F0 inversion, and also when F0 was sinusoidally-modulated. It thus appears that listeners are able to benefit from F0 variation only when it is linguistically appropriate.

Few studies have looked at the role of F0 in noise for signals that are limited in bandwidth. Hillenbrand (2003) showed that the deleterious effect of F0 disturbances is exacerbated when noise-masked speech is low-pass filtered. Oxenham and Simonson (2009) did not manipulate F0, but suggested that masking release (frequently attributed at least partly to F0 encoding) is impaired when bandwidth is limited.

It is important to note that all of the studies discussed in this section have used filtered speech with normal-hearing listeners. Generalization to listeners with hearing impairment is challenging because of the confluence of supra-threshold factors that are not adequately simulated by a mere low-pass filter. Thus, while the exaggerated F0 contours in Miller et al.'s (2010) study were not useful for NH listeners, Grant (1987) suggested that hearing-impaired listeners would not be able to detect subtle F0 contrasts, and would therefore benefit from F0 contours *only* if they were exaggerated by roughly 1.5 to 6 times those observed in natural speech.

Numerous studies suggest the benefit of a natural F0 contour in noise, but explanations for this benefit remain incomplete. The benefit of the natural F0 contour of speech is supported by several studies but exactly what segmental information is transmitted via the F0 contour? In English, F0 changes can direct a listener's attention at focused words in a sentence (Cutler & Foss, 1977), but it is not clear why this attention spotlight would increase the intelligibility of an utterance at the segmental level. There are at least two phonetic features that have been associated with F0 variation. One is vowel height, which has been shown to vary directly (if modestly) with F0 across an enormous catalog of languages, including those where F0 is constrained by tonal phonology (Whalen & Levitt, 1995). As vowel height is raised, F0 is also raised. This is unlikely to drive the benefit observed in noise though, since consonant confusions dominate the errors observed in the clinic and in experimental conditions. A more likely candidate is the use of F0 as a cue for consonant voicing; voiceless consonants tend to be associated with a higher F0 in adjacent vowels (House & Fairbanks, 1953), and this bit of segmental information could aid in the recognition of speech (particularly the voicing feature) in the adverse conditions described above. Phatak and Allen (2007) suggested that speech-shaped noise can mask cues for the voicing feature. The current study explores the possibility that good performance for stop consonant voicing perception in noise is at least partly attributable to the use of F0 as a cue.

The growing focus on acoustic cues

Although most phonetic features have well-known acoustic correlates, it is rare to find analysis of confusion patterns based explicitly on acoustic analysis. Soli and Arabie (1979) suggested that acoustic cues (rather than phonetic features) were better predictors of consonant confusions, owing at least partially to the varying acoustic instantiations of phonetic features across different classes of sounds. For example, the [-voice] cues for /t/ and /p/ can both be described as epochs of aperiodic aspiration noise, but this noise has a notably different spectral structure for /t/ than for /p/ (to be described later), which is further modulated by vowel context (Cooper, Delattre, Liberman, Borst & Gerstman, 1952). These sounds are thus likely to be recovered with varying degrees of success, depending on the spectral shape of the interfering masker. This variability is overlooked by an account of phonetic features; it can only be described in terms of acoustics. Therefore, the treatment of voicing in most analyses oversimplifies the information that listeners need to recover in order to correctly recover phonetic cues.

Dubno and Levitt (1981) examined an extensive list of acoustic cues and their relevance for consonant perception in noise. For example, in addition to considering place of articulation as a discrete variable (e.g 0,1,2,3…), they measured various spectral peaks, transition durations, energy ratios and durations of other acoustic events such as consonant closures. They found that consonant energy, consonant spectral peaks and speech-to-noise ratio were especially good predictors of intelligibility. In view of the previous two studies by Miller and Nicely (1955) and Wang and Bilger (1973), Dubno and Levitt did not test for voicing confusions (they were uncommon and presumably not worthy of separate consideration). Thus, understanding of the robustness of this feature remains incomplete.

The acoustic-driven approach has been rejuvenated in recent years, but remains imperfect. This is likely because of the varying objectives of analyses; decomposing perceptual representations is a different task than optimizing automatic speech recognition (ASR), and yet different from describing basic auditory processing. Sometimes, instead of exploring a wide variety of cooperating cues for features, researchers will aim at uncovering singular events that define individual sounds (i.e. invariant cues). For example, Régnier and Allen (2008) identified a 20ms diffuse high-frequency burst of energy between 4 and 8 kHz that appears to be crucial to perceiving /t/. Li, Menon and Allen (2010) suggested a formalized method by which the crucial acoustic events are identified; their three-dimensional deep search (3DDS) involves time-truncation, spectral filtering and noise masking to identify essential temporal and spectral properties of a sound, and how robust they are to noise. While this method appears to be fruitful for enhancing speech contrasts in noise, it suffers from a number of limitations that weaken its suitable application to human listeners. In particular, it ignores the well-known inter-dependence between acoustic components that give rise to phonetic perceptions. The 3DDS methods (in particular, elision or truncation of signal components) can potentially yield misleading conclusions of the type that slowed the understanding of final consonant voicing contrast (see Chapter 2). Time-truncation (rather than time compression / expansion) may eliminate temporally-constrained spectral cues whose absence relegates a listener to increase dependence on residual cues that were not previously salient. For example, elision of later-occurring information in vowels removes spectral transitions that drive voicing decisions; when this information is removed, the role of duration information is overestimated (see Walsh & Parker, 1984 for a review of this case for voicing perception). Li and Allen (2011) used band-pass filtering to isolate crucial regions of frication spectral energy in the contrast between /s/ and /ʃ/ (among many other speech contrasts). While this is widely regarded as a dominant cue for this contrast, this approach neglects the well-documented influence of adjacent formant transitions (Whalen, 1984) and other vocalic cues that convey vowel and talker gender information (Mann & Repp, 1980). The influence of spectral content in adjacent segments has been observed for the consonant place contrast (Lotto & Kluender, 1998) as well as vowel contrasts (Holt, Lotto and Kluender, 2000). Thus, despite the potential usefulness of the methods proposed by Li, Allen and their colleagues, it may be fruitful to explore phonetic perception in a different way when modeling the experience of human listeners. Essentially, the removal of cues is not the same as changing cues, so conclusions based on either method might not always be in agreement. The strategy used here is to explore the contributions of acoustic cues by changing them rather than eliminating them.

Perception of stop consonant voicing

For word-initial stop sound voicing (the contrast explored in the current experiment), most literature has focused on three acoustic cues: voice-onset-time (VOT), fundamental frequency (F0), and the first formant transition/onset frequency (F1). The perceptual contrast for voicing in stop sounds has been largely attributed to VOT, which is the timing difference between consonant release and the onset of voicing for the following vowel (Lisker & Abramson, 1964). For English stops, large positive VOTs correspond to voiceless sounds, while small or negative VOTs correspond to voiced sounds. Since the early days of speech synthesis, it was clear that the onset frequency of the first formant played a role in this contrast as well. Liberman, Delattre and Cooper (1958) showed that progressive cutback of the rising F1 transition (i.e. progressive raising of the onset frequency) facilitated the perception of voicelessness in synthetic stops. This essentially ascribes the aspiration of voiceless stops to a change in the vowel rather than an additional segment before the vowel; since the vowel onset contains formant transitions (the first of which is always rising), an extended period of voicelessness at the onset will naturally cause those transitions to be all or mostly completed before voicing begins[1]. For the case of the first formant, this means two things: 1) the onset frequency of

[1] *This is consistent with treating onset formant transitions that follow voiced stops as part of the vowel (not consonant) segment. Note that this clarification of the vowel onset (that it begins with the formant transitions after the consonant release, regardless of whether those formants are excited by a periodic or aperiodic glottal source) has implications for analyses and models that use vowel length as a cue for onset stop voicing, such as those by Allen and Miller (1999) and Toscano and McMurray (2009). It would also address the*

F1 will be higher following aspiration, and 2) the rapid change in the spectrum resulting from the F1 transition will be all or mostly completed before the vowel has begun. An experiment by Lisker (1975) suggested that it is the F1 onset frequency rather than the rapid spectral change that is responsible for this effect, although it is not clear whether the results of that study could be explained partly by changes in vowel quality. Regardless, the involvement of F1 in voicing perception is worthy of consideration.

Jiang, Chen and Alwan (2006) revealed that at unfavorable SNRs, listeners are able to use F1 onset transition as a cue for stop voicing in noise. Specifically, F1 transition was beneficial for non-high vowel contexts, owing to the larger transition between the low starting point for F1 in the consonant closure to the higher F1 for low vowels (confirming a prediction by Hillenbrand et al., 1984). For high vowels /i/ and /u/, F1 is relatively stable, as the low steady-state F1 is not much different from the low starting point in the consonant closure. This helps to explain the aforementioned results of Dubno and Levitt (1981), who found that consonants were more confusable in /i/-environments. It also qualifies the many studies that use the vowel /*a*/, a sound that

issue of measuring vowel onset in whispered speech, where phonation should be absent altogether. This would attenuate the differences in length of "vowels" following voiced vs. voiceless stops as described by Allen and Miller (1999) and Toscano and McMurray (2009). Furthermore, this perspective would also challenge Li and Allen's (2011) claim that F2 transitions are unnecessary for perception of voiceless stop contrasts; since the F2 transition is integral in the spectral shape of the aspiration noise, it contributes heavily to the burst/aspiration cue that is generally considered to be essential for the place contrast.

contains a very large F1 transition and is thus presumably the easiest vowel environment in which to recover voicing using the F1 transition cue.

Yet another cue for stop consonant voicing is F0 contour. Following voiceless stops, F0 is relatively higher than that after voiced stops (House & Fairbanks, 1953). This pattern was replicated by Lehiste and Peterson (1961) in their analysis of intonation. They found that the pitch peak for syllables with voiced onsets occurred near the midpoint of the vowel, whereas the pitch peak for syllables with voiceless onsets occurred immediately after the consonant. Hombert (1975) found that this voicing-related F0 contrast is consistent across many languages, and is expressed over the first 100 ms of the vowel. Haggard et al. (1970) suggested that listeners can use this F0 contour to categorize stop consonant voicing, although stimuli in their experiment were unnatural in the sense that they contained a uniform VOT ambiguously between that for a typical /b/ or /p/, and the range of F0 onsets was 163 Hz. In natural speech, the voicing-driven F0 difference has been shown to be much more modest; Hombert (1975) and Ohde (1984) suggest that speakers produce differences of roughly 30-40 Hz. This natural range was tested against the dimension of VOT in a study by Abramson and Lisker (1985), who showed that the influence of F0 was strongest in the ambiguous range of VOT levels, but negligible for unambiguous VOTs. Whalen et al. (1993) suggested however, that even for unambiguous items, F0 that is not consistent with the voicing of the stop sound (according to the VOT) will slow listeners’ judgments. That is, listeners are sensitive to F0 even when the VOT clearly indicates a voiced or voiceless sound.

Summary of literature review

This chapter has thus far described the impacts of noise and hearing loss as they relate to speech perception, and has described how they may impact voicing perception. The difficulties that listeners experience in noise reveal non-random confusions of phonetic segments. Notably, segments that are voiced are rarely confused with those that are voiceless. The beneficial role of F0 as a cue for listening in noise suggests that its role in the voice-voiceless contrast may be promoted at less-favorable SNRs. The additional difficulty of low-pass filtering/hearing loss in noise also implicates F0 as a potential source of benefit, since it is accessible even without high-frequency information. Little work has been done to unpack the perceptual processes that maintain accurate voicing perception in noise, especially with regard to the use of F0. The primary goal of this study is to expand further along the trajectory laid out by previous work by Miller and Nicely (1955), Dubno and Levitt (1981), Régnier and Allen (2008) and Li and Allen (2011) by shifting the focus of analysis from phonetic features and acoustic events to the relative importance of acoustic cues/events as they convey phonetic features. In other words, previous analyses identified the features that are recovered, but this experiment aims at the mechanisms by which they are recovered.

Despite the increased focus on the acoustic aspects of phonetic features, it is still rarely stated explicitly that acoustic events (e.g. a burst centered within a particular frequency range, a durational cue, a particular kind of formant transition, a spectral tilt, etc.) might take on different roles in changing listening conditions. Instead, many investigations have sought the acoustic events that are robust *across* listening conditions, with the assumption that those events are critical to the phonemes in question. The current study takes a different approach; the cues used for phonetic feature identification in noise or band-limited conditions are hypothesized to rely on acoustic events that have only negligible effects on perceptual responses in optimal conditions. Since these cues naturally co-vary with others that are more prominent in favorable conditions, they preserve response patterns that imply “normal” or at least “successful” perception of voicing.

Following the motivation of experiments in Chapter 2, the experiments in this chapter address whether listeners could be more strongly compelled to shift phonetic cue reliance if one of the cues was compromised. The current experiment addresses high-frequency audibility and/or background noise. For the case of stop consonant voicing perception, this is essentially the problem of masking the aspiration noise whose duration signals the contrast. Since this energy tends to be concentrated in the high frequency range (especially for /t/), low-pass filtering would predictably compromise this cue. Since the voicing feature is typically perceived correctly, the question remains as to whether listeners are able to use the compromised cue or whether they can capitalize on the other cues that remain in the low frequency range. The low-frequency portion of the spectrum can be masked heavily by speech-shaped noise, which is the other degradation in this experiment. Previous literature showing the benefit of natural F0 contour in noise suggests that F0 is perceptible in noise and therefore potentially useful as a voicing cue. It stands to reason that F0 might play a role in the high level of success in voicing perception by listeners in noisy and/or filtered conditions.

Experiment 3:
Bandwidth, noise and the stop consonant voicing contrast

Hypotheses

It was predicted that when speech signals were low-pass filtered or masked by speech-shaped noise, listeners' voicing judgments would be driven more heavily by F0 and less by VOT. Additionally, it was predicted that in the full-spectrum condition, the masking noise would have larger effects for the /p/-/b/ contrast than for the /t/-/d/ contrast, since the burst and aspiration noise for the /p/ sound more closely matches (and thus would be more heavily masked by) the speech-shaped noise. Because the /t/ burst and aspiration are nearly exclusively comprised of high-frequency energy, it was predicted that the /t/-/d/ contrast would be more heavily affected by low-pass filtering.

Method

Participants.

Participants included 20 adult listeners (mean age: 24.3 years, 15 females) with normal hearing, defined as having pure-tone thresholds ≤20 dB HL from 250–8000 Hz in both ears (ANSI, 2010). All participants were native speakers of American English and were screened for self-reported familiarity with tonal languages (e.g. Mandarin, Cantonese, Vietnamese, etc.) to ensure that no participant entered with *a priori* increased bias towards using F0 as a lexical/phonetic cue.

Stimuli.

There were two sets of stimuli that were created using natural speech. The words Pete, Beat, Teen and Dean were recorded multiple times by a native speaker of English, and tokens with similar voice quality and inflection were chosen for subsequent modifications. These words were recorded in a double-walled sound-treated room using an AKG C1000 microphone at 44.1 kHz sampling. The stimuli were equated for peak RMS amplitude in the vowel segment, for reasons to be discussed in the next section. Stimuli varied by VOT (in 7 or 8 steps for p/b and t/d, respectively) and F0 (in 8 steps). Following the method used by McMurray (2008), portions of words with /b/ or /d/ onsets were progressively replaced with voiceless aspiration from /p/ or /t/, respectively, in 10 ms increments from the onsets (bound at the closest zero-crossing) to create continua of voice onset time. Thus the vowel from each stimulus item came from the /b/ or /d/-initial tokens. For the d/t continuum, the VOT range spanned from 0ms to 70ms, and the range for the b/p continuum spanned from -10ms (pre-voicing) to 50ms, as indicated by previous studies (Lisker & Abramson, 1964; Abramson & Lisker, 1970). The F0 contour was manipulated using the pitch synchronous overlap-add (PSOLA) method in Praat (Boersma & Weenik, 2011). The range of F0 spanned from 94 – 142 Hz, which was a slight expansion of the ranges indicated by Ohde (1984), Abramson and Lisker (1985), and Whalen et al. (1993). The F0 was interpolated in 8 steps along a log scale. It was kept steady over the first two pitch periods of the vowel, and fell (or rose) linearly until returning to the original contour at the 100 ms point in the vowel. This 100 ms epoch of voicing-related F0 contour matches that observed by Hombert (1977). See Figure 3.1 for a schematic illustration of the F0 contours in these stimuli.

Figure 3.1. F0 contours for stimuli in Experiment 3

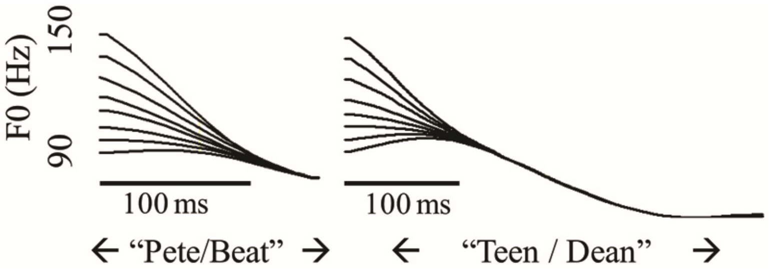

Figure 3.1 Schematic of F0 contours for stimuli in Experiment 3. Duration of the left contours (for Pete/Beat) varied between 148 and 198 ms, depending on the duration of the VOT. Duration of the right contours (for Teen/Dean) similarly varied between 422 and 492 ms. F0 contours were expanded or contracted (rather than truncated) to accommodate for VOT differences.

Background noise.

Speech-shaped noise was extracted offline from the iCAST program (Fu, 2002). Its spectrum was strongest in the 200-600 Hz region, and decreased by roughly 6 dB per octave. The noise began roughly 280 to 360 ms before the onset of the consonant release, and ended roughly 380 to 450 ms after the end of the word. Placement of the stimulus within the noise was varied so that onset relative to noise could not be used as a reliable perceptual cue. The noise contained 70 ms onset and offset volume ramps. The level of the noise was set relative to the level of the vowel segment rather than the entire syllable for two reasons. First, the stimuli with longer VOTs had less overall energy since the loud voiced energy was replaced by lower-level aspiration noise. Thus, referencing the overall energy would have resulted in more favorable noise levels for the long-VOT items (since less noise is required to mask the resulting softer sounds to reach the same SNR). Second, the syllables in the p/b continuum ended in a voiceless stop that contained a considerable epoch of virtual silence (compared to the continuous voicing for the nasal at the end of the Teen-Dean tokens). Therefore, the overall RMS level of the syllables was affected by segments unrelated to the contrast in question. Using the entire syllable to calculate RMS would have thus resulted in a more-favorable noise level for the p/b condition since less noise is required to mask the b/p words to reach the same SNR. Since many studies do not reference one particular point in a syllable to calculate SNR, the reader is encouraged to remain cautious when comparing these SNR levels to those from other publications. The advantage of the current SNR calculation is that it permits the comparison of t/d and p/b contrasts in the same condition.

Low-pass filtering **.**

The stimuli were low-pass filtered using the Hann band filter function in Praat (Boersma & Weenik, 2011), using the parameters in Table 3.1. In contrast to the method used by Stickney and Assmann (2001), this filtering was done *after* the addition of background noise, to more closely simulate the experience of hearing impairment, which would affect both target and masking sounds. This ordering of masking and filtering was done so that the noise level was not made more favorable merely as a result of the stimulus being filtered (volume attenuation would require less noise to mask the filtered sounds at equivalent SNRs). The volume level of the words was not adjusted after filtering, for the same reason; this resulted in filtered stimuli that were noticeably softer in volume than the full-spectrum stimuli.

Table 3.1

Parameters of low-pass filters in Experiment 3.

Condition	Low-pass (Hz)	Sideband smoothing (Hz)
Full-spectrum	N/A	N/A
4000 Hz LPF	3750	400
2000 Hz LPF	1750	400
1000 Hz LPF	800	300

Note: filters were created using the Pass (Hann) Band function in Praat.

Changes in the acoustic cues resulting from noise and filtering.

The presence of noise and/or filtering has demonstrable effects on the temporal envelope of the sound. Since this time-varying envelope contains the crucial relative timing/amplitude changes that carry the VOT cue, VOT is clearly compromised in these conditions. See Figure 3.2a for an illustration of how the temporal envelope cue is clearly visible in a quiet waveform, but obscured when masked by increasing amounts of noise. See Figure 3.2b for an illustration of the effects of the LPF setting on the stimuli in 0 dB SNR masking noise.

Figure 3.2a. Spectrograms of stimuli in different SNR conditions in Experiment 3

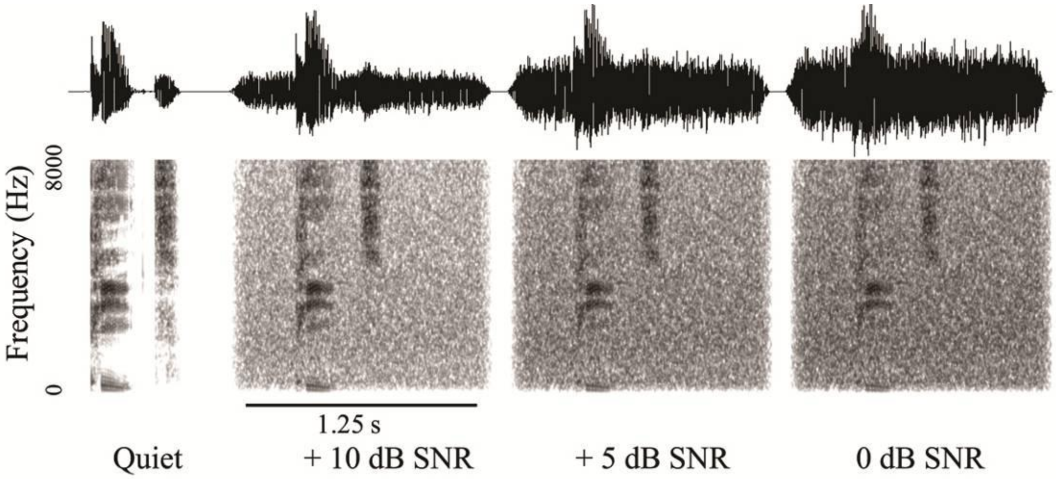

Figure 3.2a. Four waveforms and spectrograms of the word “Pete” in various conditions of masking noise.

Figure 3.3. Spectrograms of stimuli in different LPF conditions in Experiment 3

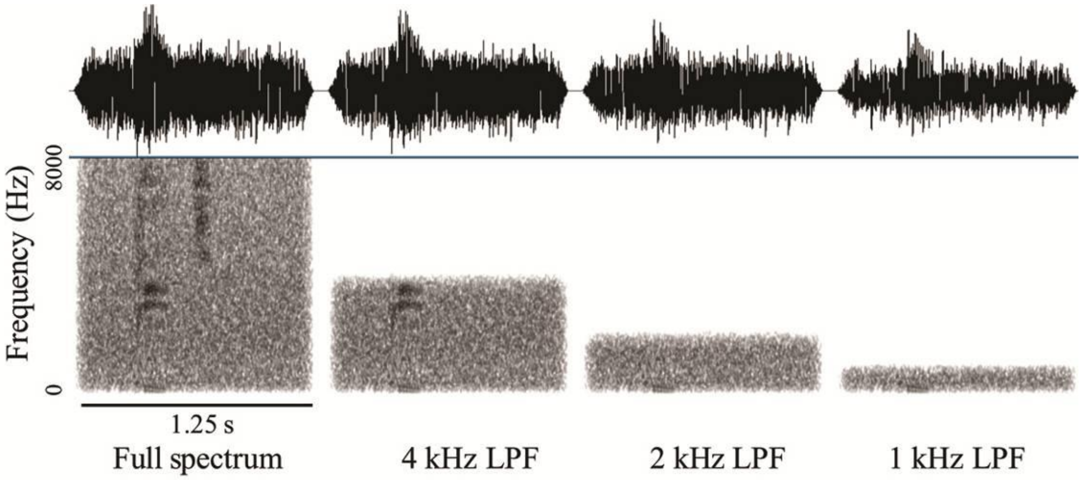

Figure 3.3. Four waveforms and spectrograms of the word "Pete" at 0 dB SNR in various low-pass filter (LPF) settings.

With regard to the masking of the aspiration noise, the t/d and p/b contrasts are affected in similar but subtly different ways. If the problem of voicing judgment is reduced to a problem of detection of aspiration noise, the spectrum of that noise determines the deleterious effect of the masker or filter. The rising spectrum of the /t/ burst is not effectively masked by the speech-shaped noise (Phatak and Allen, 2007); in our stimuli where the vowel-to-noise intensity ratio is 0 dB, the SNR in the frequency range between 4 kHz and 8 kHz (where the /t/ burst is found) is +13 dB. The /p/ burst in comparison has a falling spectral shape, meaning that it would undergo relatively less change due to the filtering, but relatively more masking in speech-shaped noise. See Figure 3.3 for an illustration of how speech-shaped noise differentially masks the aspiration noise for /p/ and /t/ bursts. The spectra of these consonant sounds suggests that speech-shaped noise compromises the /p/ noise more heavily than the /t/ noise, which has energy primarily in the frequency range where the noise is weakest.

Figure 3.3. Spectra of /t/ and /p/ sounds with different levels of masking noise

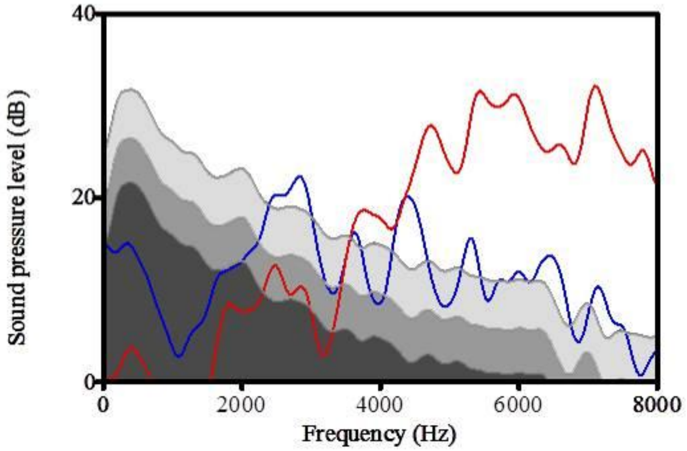

Figure. 3.3. Illustration of aspiration spectra for /p/ (blue) and /t/ (red), with different colors of shading (in grey) to denote the spectral shape of the masking noise at different SNRs. The highest level of noise was used for the 0 dB SNR condition, while the middle and lower lines reflect the levels for the +5 dB and +10 dB SNR conditions, respectively (SNR refers to level of vowel relative to level of the masking noise).

The other challenging aspect of this experiment is low-pass filtering. Just as for the masking noise, the /p/ and /t/ sounds are affected by filtering in different ways. Since the majority of the energy for the /t/ aspiration is contained in frequencies above 4 kHz, the filter settings used in this experiment seriously compromise this segment. For the /p/ aspiration, however, there is sufficient energy at lower frequencies to convey the segment even when the signal is filtered. See Figure 3.4 for an illustration of how low-pass filtering affects the envelope and spectral properties of the p and t bursts and aspirations, which could be regarded as cues for voicing. The spectra of these consonant sounds suggests that filtering significantly compromises the /t/ voicing cue, while it leaves intact a portion of the /p/ voicing cue. Therefore, it is expected that filtering will affect the t/d contrast more heavily than the p/b contrast.

It should be noted that while the vowels and final consonants in all speech stimuli were degraded by the masking noise and filtering, these segments were entirely predictable within testing blocks, and indicated by the visual word choices.

Figure 3.4. Spectra of /t/ and /p/ sounds with different low-pass filters

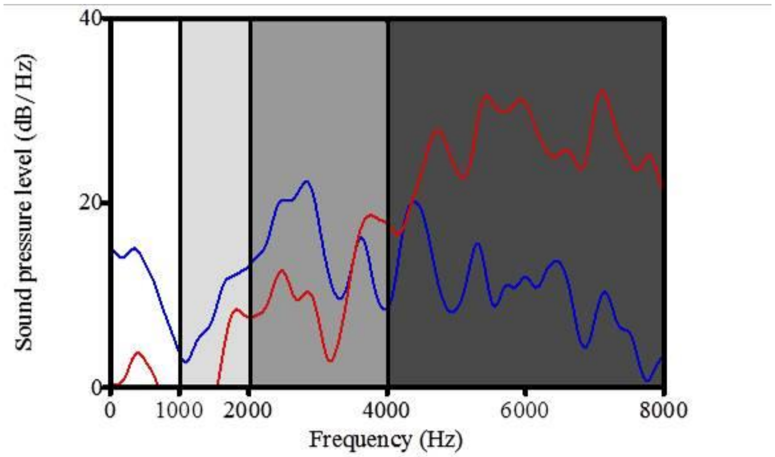

Figure 3.4. Illustration of aspiration spectra for /p/ (blue) and /t/ (red), with vertical lines marking the low-pass filter (LPF) settings used in this experiment. Different colors of shading denote the availability of that spectral region in the various LPF conditions.

Procedure.

All speech recognition testing was conducted in a double-walled sound-treated booth. Volume level was calibrated at the position of the listener’s head using a Radio Shack sound level meter that referenced a 1 kHz tone that was equated in RMS amplitude to the speech stimuli in the optimal condition (full spectrum, quiet) Stimuli in the optimal condition were presented at 65 dBA in the free field through a single Tannoy Reveal studio monitor loudspeaker (frequency response: 65 Hz – 20 kHz) at a distance of 1 – 2 feet placed in front of the listener at eye level. Filtered speech was not amplified to equate loudness, for reasons expounded earlier, and to prevent a confound in the presentation of low-frequency spectral energy. Different SNR conditions were constructed by adding noise at various intensities to constant-amplitude speech rather than mixing speech at lower intensities to constant-amplitude noise. Therefore, conditions at poorer SNRs were louder than those at more favorable SNRs. This was done to represent the same speech energy in different filtering and SNR conditions. Listeners responded to these stimuli by clicking a button on a computer screen labeled with word choices (Teen/Dean or Pete/Beat). There was no time limit on their response, and they were permitted to enable stimulus repetitions up to three times; stimulus repetitions were very rare. After an initial block of words in the optimal condition (full-spectrum, in quiet), stimuli were randomly presented within blocks that were organized by low-pass filter settings and SNR. Each block was heard at least 5 times. Before performing the group analyses, individual listeners' response functions were initially fit to a simple logistic model using Sigmaplot 9.01 (Systat, 2004). When listeners' data for a particular condition did not reach satisfactory convergence to the model, 1 or 2 more repetitions of that condition were conducted to allow a better fit. This was done for 5 of 20 listeners in some of the more challenging conditions (i.e. those where signal degradations were harsh enough to distort consistent use of the cues).

Conditions.

The levels of low-pass filtering and SNR in this experiment were not fully-crossed. Instead, they were motivated by the specific questions highlighted below in Table 3.2.

Table 3.2.

Different conditions tested in Experiment 3, defined by spectral bandwidth and SNR.

	Initial exploration of bandwidth and noise effects			
Bandwidth	Full	1000 Hz	Full	1000 Hz
SNR	Quiet	Quiet	0 dB	0 dB
	Effect of bandwidth in 0 dB SNR noise			
Bandwidth	Full	4000 Hz	2000 Hz	1000 Hz
SNR	0 dB	0 dB	0 dB	0 dB
	Effect of SNR with 1000 Hz low-pass filter			
Bandwidth	1000 Hz	1000 Hz	1000 Hz	1000 Hz
SNR	Quiet	+10 dB	+5 dB	0 dB

Note: Rows are organized by the specific purpose of comparison, stated in italic text. Note that the 1000 Hz, 0 dB SNR condition is present in all three rows.

This arrangement of conditions was inspired by preliminary experiments (top row) that suggested that either 0 dB SNR or a 1 kHz low-pass filter (LPF) permitted use of the VOT cue, while the combination of these factors promoted the use of F0 nearly exclusively. Questions following this pilot testing included 1) (middle row) What bandwidth is necessary to facilitate the use of VOT when the SNR is 0 dB? and 2) (bottom row) What SNR is needed to facilitate the use of VOT when the LPF is 1 kHz? Each of these conditions was tested for the p/b stimuli and for the t/d stimuli, resulting in a total of 16 conditions (some conditions above are repeated across the different comparisons). Listeners heard a variable subset of the conditions (that were not necessarily limited to one contrast), depending on their scheduling availability; most heard between 5 and 10 different conditions. There were a total of 50 condition repetitions across all listeners (i.e. ten listeners for each condition that was heard at least five times each) for a total of over 800 tested conditions. Each repetition of a single condition took roughly 3 - 5 minutes.

Analysis.

Listeners' binary responses (voiced or voiceless) were fit using a generalized linear (logistic) mixed-effects model (GLMM). This was done in the R software interface (R Development Core Team, 2010), using the lme4 package (Bates and Maechler, 2010). A random effect of participant was used, and the fixed-effects were the stimulus factors described above (Consonant place, VOT, F0, LPF, SNR). The binomial family call function was used because the possibility of a "voiceless" response could not logically exceed 100% or fall below 0%. The model incorporated the logit link function, and an assumption that variance increased with the mean according to the binomial distribution. The model incorporated each main factor and all possible interactions (the four-way interaction was significant, necessitating the inclusion of all nested factors and interactions) The goal of this model was similar to that used by Peng et al. (2009); it tested whether the coefficient of the resulting parameter estimate for an acoustic cue was different from 0 and, crucially, whether the coefficient was different across conditions of LPF and SNR levels. Changes in this coefficient represent changes in the log odds of voiceless perceptions resulting from the condition change.

Results

The psychometric functions for the first comparison (the initial exploration of bandwidth and noise effects) are shown in the four panels in Figure 3.5a and 3.5b, and the factor coefficients are illustrated in Figure 3.6. To conserve space, results for the other conditions are represented solely by the factor coefficients in Figures 3.7 and 3.8.

The initial comparison confirms (for both the p/b and t/d contrasts) that the use of VOT declines with either the 1 kHz low-pass filter or with the 0 dB SNR noise. The combination of both these effects resulted in a dramatic decline in VOT use. Conversely, the opposite result held for the F0 cue. For the p/b contrast, either the filtering or noise promoted increased use of F0, while the combination of filtering and noise dramatically increased use of F0. For the t/d contrast, the noise did not facilitate increased F0 use, but the filtering and combination filtering/noise did increase F0 use. All of these effects are consistent with the predictions based on the acoustics of the aspiration noises of /p/ and /t/ segments. Specifically, since the /t/ aspiration spectrum is audible even at 0 dB SNR, it was still usable, minimizing listeners' need to recruit F0 as a cue.

Figure 3.7 illustrates the effect of signal bandwidth in the 0 dB SNR conditions. The use of VOT gradually declined with decreasing bandwidth for the p/b contrast, and declined precipitously for the t/d contrast. That is, the presence of 0 dB SNR noise was especially deleterious for the t/d distinction when it was filtered at 4 kHz or lower. Bandwidth had a less dramatic effect for the p/b contrast.

Figure 3.8 illustrates the effect of SNR in the 1 kHz low-pass filter conditions. For the p/b contrast, there is a gradual decline in the use of VOT as SNR becomes less favorable in the 1 kHz LPF condition. For the t/d contrast, this decline is more abrupt, since the presence of any noise in the 1 kHz LPF condition apparently compelled listeners to abandon the VOT cue in favor of the F0 cue. In general, the decline in VOT use was accompanied by an increase in F0 use. Thus, listeners did not simply randomly guess when VOT was compromised – they recruited helpful information from a different acoustic cue.

Figure 3.5a. Psychometric functions for VOT and F0 cues for the /p/-/b/ contrast in different conditions in Experiment 3.

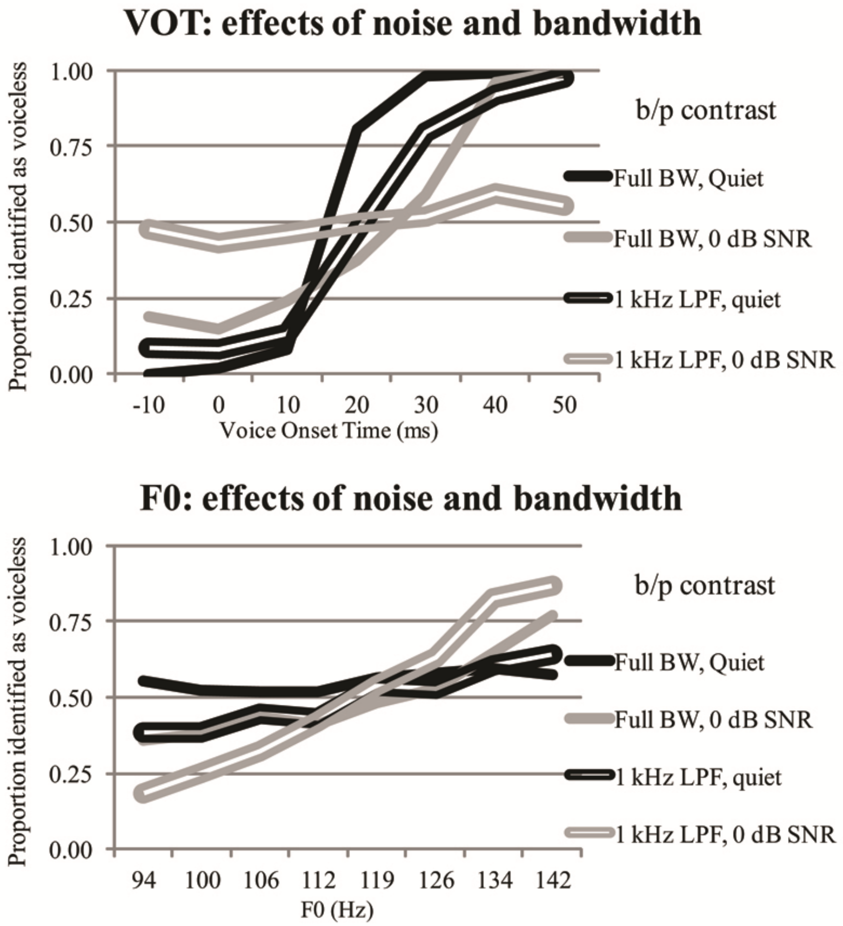

Figure 3.4a. Group mean psychometric functions along the continua of voice-onset time (upper panel) and F0 (lower panel) for the p/b contrast in various listening conditions. Error bars were omitted to maintain clarity.

Figure 3.5a. Psychometric functions for VOT and F0 cues for the /t/-/d/ contrast in different conditions in Experiment 3.

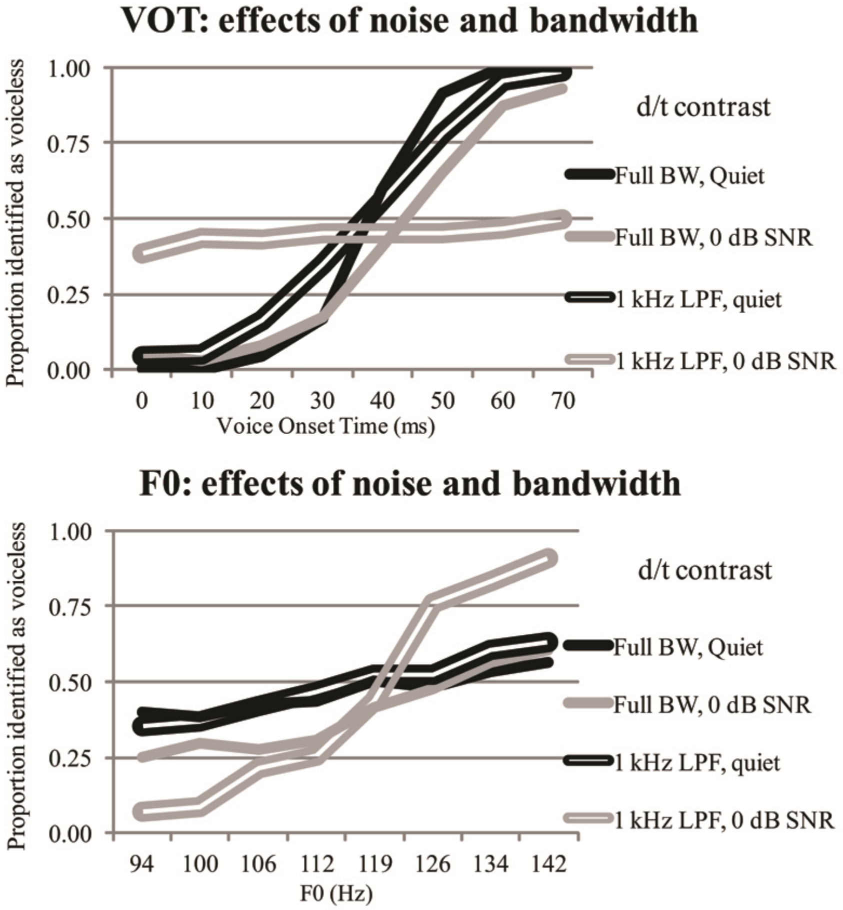

Figure 3.5b. Group mean psychometric functions along the continua of voice-onset time (upper panel) and F0 (lower panel) for the t/d contrast in various listening conditions. Error bars were omitted to maintain clarity.

Figure 3.6. Parameter estimates for VOT and F0 cues in Experiment 3

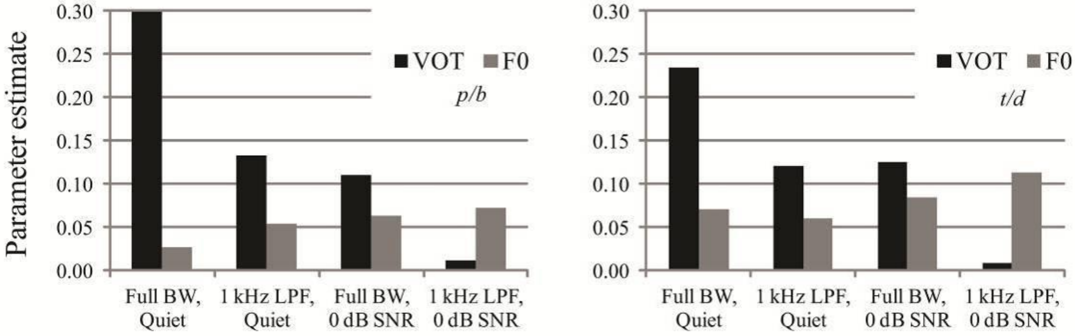

Figure 3.6. Parameter estimates (coefficients) for the logistic model for the first comparison (effects of low-pass filtering and/or masking noise). Black and gray bars represent estimates for the VOT and F0 contrasts, respectively. The left and right panels illustrate estimates for the p/b and t/d contrasts, respectively. The leftmost pairs of bars in each panel represents the “optimal” condition of full-spectrum speech in quiet, while the rightmost pairs of bars indicate the conditions with the least-favorable LPF and SNR settings.

Figure 3.7. Parameter estimates for VOT and F0 cues in Experiment 3

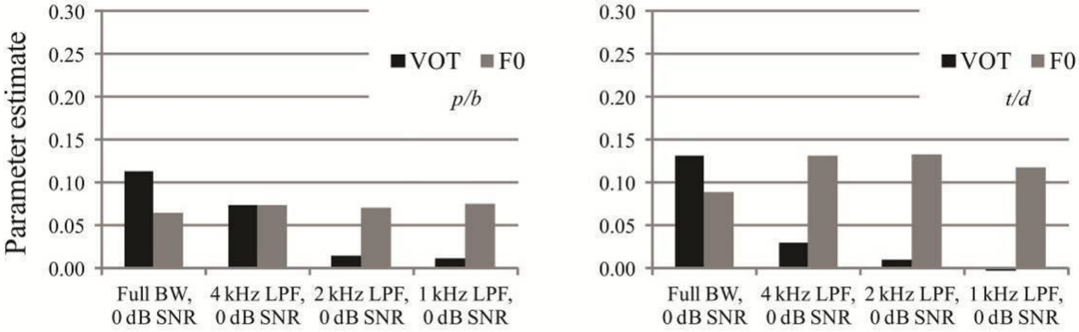

Figure 3.7. Parameter estimates (coefficients) for the logistic model for the second comparison (effects of bandwidth in 0 dB SNR masking noise). Black and gray bars represent estimates for the VOT and F0 contrasts, respectively. The left and right panels illustrate estimates for the p/b and t/d contrasts, respectively.

Figure 3.8. Parameter estimates for VOT and F0 cues in Experiment 3

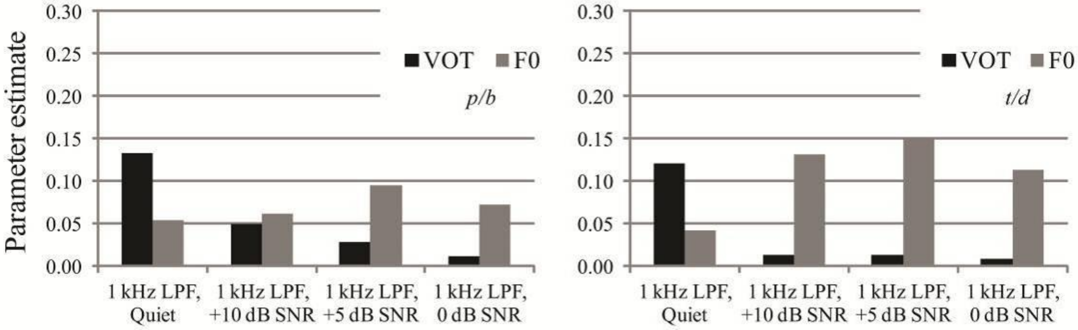

Figure 3.8. Parameter estimates (coefficients) for the logistic model for the third comparison (effects of SNR with 1 kHz low-pass filter). Black and gray bars represent estimates for the VOT and F0 contrasts, respectively. The left and right panels illustrate estimates for the p/b and t/d contrasts, respectively.

Because the analyses presented thus far do not speak to perceptual *accuracy* per se, a final analysis was conducted to evaluate the identification of stimuli where both the VOT and F0 cues cooperated to confer a typical “voiceless” or “voiced” stop consonant. Identification of these continuum-endpoint stimuli could appropriately be evaluated for correctness. Figure 3.9 illustrates performance levels for these stimuli by listeners in all conditions. Results suggest that both the voiceless and voiced stop consonants were identified reliably by listeners in all conditions; with the exception of /b/ in the most challenging condition (1 kHz LPF with 0 ddB SNR noise), all sounds were identified with 80% accuracy or greater in all conditions. Apart from the optimal condition where performance was at ceiling, listeners showed poorest accuracy for /b/ and highest for /t/. However, it should be noted that this was a 2-alternative forced choice task that assessed only voicing perception; in an open- or expanded-set task, it is likely that both of these consonants would be confused with consonants that vary in place of articulation. Essentially, this figure implies that the potentially different perceptual strategies taken by listeners in this experimental task did not necessarily result in substantial differences in voicing perception accuracy.

Figure 3.9. Identification accuracy for voiced and voiceless sounds in Experiment 3

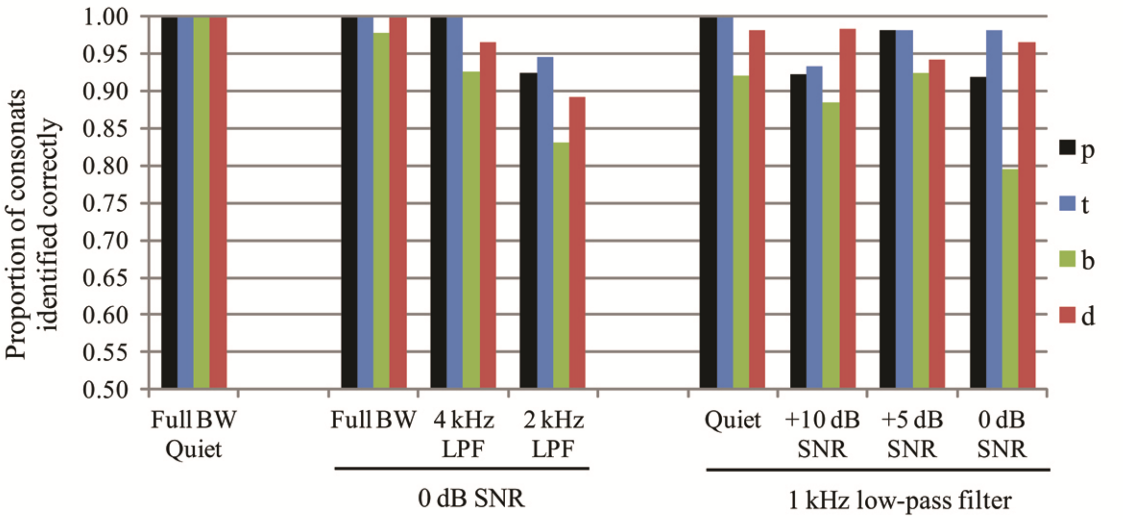

Figure 3.9 Mean accuracy in identification of stimuli at continuum endpoints by different listener groups. Voiceless and voiced items for this analysis were limited to those where both the VOT and F0 cues cooperated appropriately (i.e. long VOT and high F0 or short VOT and low F0) at continuum endpoints to signal the same feature.

Conclusions

In this experiment, listeners' use of VOT and F0 as cues to stop consonant voicing was measured in conditions that varied in terms of SNR of speech-shaped masking noise and in terms of low-pass filtering. It was found that the presence of these challenging conditions both facilitated increased reliance upon F0 contour at the expense of VOT. The decline in VOT use was anticipated in view of the demonstrable effects of low-pass filtering and masking noise on the temporal amplitude envelope, which is presumably essential for perceiving VOT. In question was whether listeners would compensate for that by increasing reliance upon F0. Results revealed that listeners indeed compensated for the decreased VOT cue use with increased reliance upon the F0 cue.

The effects of masking noise and filtering had unequal effects on the p/b and t/d contrasts. The use of cues for the p/b contrast was influenced more by the level of masking noise, presumably because the spectrum of the /p/ aspiration is mostly preserved even when low-pass filtered. The /t/ aspiration spectrum is distinctly different from that of the speech-shaped noise, so the VOT (aspiration) cue was available to listeners even when the (full-spectrum) signal was in 0 dB SNR noise. Conversely, the use of VOT for the t/d contrast appears to be influenced more heavily by low-pass filtering. Consistent with earlier literature on the acoustics and perception of /t/, the audibility of energy above 4 kHz is essential for the perception of /t/ aspiration. The low-pass filtered /t/ aspiration was rendered nearly inaudible by the LPF, resulting in complete masking even at modest SNRs. Thus, although these two contrasts are considered to be equal phonetically (and therefore averaged in most analyses), they rely on different kinds of acoustic energy for their distinction. Collapsing them into the same category for analysis (e.g. for information transfer analysis) therefore overlooks the asymmetries in perception that arise partly due to masking noise spectrum and filtering/audibility.

Summary and Discussion of Experiment 3

The motivation for this experiment was to model potential listening strategies that could arise when a person experiences hearing impairment. Because hearing impairment is more complex than a simple low-pass filter, the results of this study should be interpreted with caution. There are supra-threshold deficits in the spectral and temporal domains that might limit a listener’s ability to utilize either of the acoustic cues explored in this study. For example, some listeners with hearing impairment may not be able to capitalize on the F0 variations in this study (Grant, 1987), owing to poor frequency resolution and/or temporal fine structure coding (Bernstein & Oxenham, 2006; Lorenzi, Gilbert, Carn, Garnier & Moore, 2006). Additionally, because people with hearing loss tend to also be older than the listeners in this study, there could be additional factors that impair the use of acoustic cues in the temporal domain. Older listeners have been shown to experience deficiencies in auditory temporal processing in basic psychophysical tasks (Gordon-Salant and Fitzgibbons, 1993; 1999), and tasks involving perception of temporal phonetic cues (Gordon-Salant et al., 2006). This could potentially result in a predisposition to rely upon spectral cues (e.g. F0, F1) even in the absence of difficulties like masking noise and reduced high-frequency audibility. These aging effects can be larger when the target contrast is embedded in sentential context (Gordon-Salant, Yeni-Komshian and Fitzgibbons, 2008). Extension of these results to sentential contexts would also be subject to F0 contour effects arising from non-segmental influences. This would decrease the contrast of the voicing-related F0 perturbations and thus render the cue less useful.

The current experiment used steady-amplitude speech-shaped noise (SSN) as a masker. The use of steady-amplitude masking was motivated by the temporal specificity of acoustic cues for the contrast explored here. Amplitude-modulated noise or speech might momentarily provide a favorable SNR for these cues via a dip in the amplitude contour concurrent with the onset of the target words. The current experiment could have also used noise with a different power spectrum, such as white noise (WN). Consonant confusions occur more frequently for WN than SSN when they are equated for RMS amplitude (Phatak & Allen, 2007; Phatak et al., 2008), but this observation extends beyond the scope of the current experiment. Because this experiment explored only voicing perception, errors on place of articulation (which comprise the majority of errors in white noise) would not impact this experiment if it were run using white noise. Acoustic comparison of phonetic acoustic cue spectra with masking noise spectra (Figure 3.2) suggested that SSN masked /p/ more heavily than the /t/ sound. White noise would result a different masking pattern, and would thus likely result in a reversal of the asymmetrical effects observed in the current experiment for the /p/-/b/ and /t/-/d/ contrasts. Specifically, with the entire spectrum available and audible, the SNR of stimuli in WN would likely affect the /t/-/d/ contrast more heavily than the /p/-/b/ contrast.

Despite the differences between the current participants and individuals with hearing loss, the current study sheds light on some of the potential reasons why voicing is such a robust feature in the presence of various adverse listening conditions. When one acoustic cue for voicing is compromised, listeners are able to capitalize on the presence of other residual cues that convey the same information. In the case of decreased high-frequency audibility and the presence of moderate masking noise, that residual cue in this study was the F0 contour. It is likely that this cue contributes to the benefit of a natural F0 contour of sentences presented in noise. For other words, it is likely that other cues remain as well, such as the F1 contour.

Because the influence of VOT on the F1 contour was minimized in this experiment (via the use of the high /i/ vowel), the VOT cue can appropriately be described as a temporal cue. That is, the VOT cue was conveyed via the temporally-varying amplitude contour. Decreased use of this cue in the low-pass filtered conditions is consistent with the results of Eddins, Hall and Grose (1992), who showed that temporal resolution (measured via gap detection) improves with increasing signal bandwidth. It is also consistent with the results of Shailer and Moore (1983; 1985), who found that higher-frequency spectral regions facilitate better temporal resolution[2] (although this relationship is strongest for frequencies below 1.5 kHz, which were represented by only one condition in the current experiment). This reflects the well-known trade-off between spectral and temporal resolution in the peripheral auditory system, where lower-frequency regions have better spectral resolution and poorer temporal resolution; high-frequency regions have the opposite pattern. Because the LPF in the current experiment limited the bandwidth and also eliminated high-frequency energy, listeners likely experienced poorer ability to encode VOT on a purely psychophysical level.

Noise-reduction schemes in modern hearing aids commonly attenuate low-frequency energy; this strategy may be at odds with the results presented in this paper. It

[2] The relationship between stimulus frequency and temporal resolution is at least partially qualified by the type of stimulus used. Moore, Peters and Glasberg (1993) used sinusoidal stimuli in a gap detection task and found a weak relationship between frequency and temporal resolution for the frequency range between 200 and 2000 Hz. Results from that study do not rule out the possibility of a frequency – temporal resolution relationship that would be exploited in the comparison between 2 kHz and 4 kHz LPF conditions, or between 4 kHz and full-spectrum conditions. Additionally, the speech signals in this experiment are broadband in nature and thus correspond more directly to the studies that used broadband noise stimuli.

is not clear whether low-frequency attenuation would translate into ostensible detriment though, since everyday listening is much different than the experimental task of two-choice forced identification of single syllables. In normal listening, there is an abundance of influences apart from the basic acoustic input, including visual input and cognitive processing from higher-level linguistic processing (McClelland, Mirman and Holt [2006] review some of these cognitive interactions with speech perception). Future work might explore whether the current findings generalize to sentential context and other scenarios where the response choices are open-set or at least more numerous. It may be the case that the voicing-related F0 perturbations exploited by listeners in the current study would be compromised in running speech for at least two reasons. First, acoustic cues are known to be reduced in running speech as a result of coarticulation and speech timing limitations. Additionally, there are non-segmental factors that govern F0 at the utterance level, including sentence type or intention as well as the location of stressed words. Thus, there is much work to be done to evaluate the usefulness of the F0 contour for voicing perception by listeners in natural listening conditions. Even in the absence of that work, the results from this experiment suggest that listeners can capitalize on the redundancy of acoustic cues in the speech signal to recover the voicing features accurately in adverse listening conditions (Figure 3.9). Importantly, since the recovery of voicing in challenging conditions in this experiment was not done via the ‘conventional’ method of VOT perception, it should be noted that correct perception of this (or any other) feature does not necessarily mean that the listener has access to sufficient information to facilitate optimal or conventional perception.

Chapter 4: Effects of spectral degradation on the use of contextual influences in phonetic identification.

Introduction and Motivation

Variability in the acoustics of phonetic segments is a well-known phenomenon that presents a challenge to those who study automatic speech recognition and also to those who wish to understand the perceptual processes that listeners use to recognize those segments. An efficient system of speech recognition (whether human or machine) must accommodate several kinds of variance in speech acoustics, including those arising from phonetic context and from inter-talker differences, such as gender and vocal tract size. One example of this accommodation is the phonetic context effect, whereby a listener interprets the same sound differently depending on the context in which it is presented. In this paper we explore this behavior as it relates to spectral degradation, which is central to the experience of individuals with cochlear implants.

Auditory and phonetic context effects

One of the earliest published examples of phonetic context effects is from a study by Cooper et al. (1952), who found that a synthetic stop burst centered around 2000 Hz would be heard as either /p/ or /k/, depending on the vowel that followed the burst (generally, it was heard as /p/ if the following vowel did not have a formant peak in the region of the burst). Another heavily explored example of a phonetic context effect is the perception of stop consonants /g/ and /d/, which is affected by the presence of a preceding liquid consonant /l/ or /r/. Following the syllable /ar/, listeners are biased to hear /da/, while they are more likely to hear /ga/ after the syllable /al/ (Mann, 1980).

The /arda-alga/ experiment has been replicated in a number of ways that shed light on perceptual processes that are not bound by phonological or phonetic constraints. For example, Mann (1986) revealed that the influence of the /l/ and /r/ segments persist even for Japanese speakers, for whom the /l/ and /r/ are not phonologically contrastive. Lotto and Kluender (1998) showed that the preceding contextual segment (ar/al) and target segment (da/ga) do not need to emanate from the same voice in order to interact. In that experiment, the talker changed mid-way through the utterance, but the first talker's speech still had some influence on the labeling of the sound of the second talker. This influence was weaker than that from speech produced by the same talker, suggesting at least some role in spectral continuity throughout the perceived segments. A second experiment in that study replaced the preceding syllable with a simple tone glide that matched only the frequency of the third formant in the first syllable (as /l/ and /r/ are distinguished primarily by this third formant). Although the context effects from the tone glide were modest, they promoted bias in the same direction observed with full-spectrum speech. This effect persisted even when the preceding tone was held at a constant frequency instead of being modeled after the dynamic transition found in speech. A consistent observation across these studies is that the source of bias (whether a formant transition or corresponding non-speech tone/glide) enhances frequency contrast for the cue that ultimately determines the listener's judgment. As the precursor sound ends at a higher frequency, the following sound will be enhanced at the lower frequency, and vice-versa. Precursors with a low F3 (/r/) promote the perception of following sounds with high F3 (/d/), while precursors with high F3 (/l/) promote perception of following sounds with low F3 (/g/). This helps to also explain that when listeners are presented with sounds that vary along a /k-t/ continuum, they are biased to hear /t/ (high frequency) after /ʃ/ (low frequency) and to hear /k/ (low frequency) after /s/ (high frequency) (Mann and Repp, 1981).

Frequency contrast has not been uniformly accepted as the only contribution to context effects; Fowler, Brown and Mann (2000) challenged this explanation by pointing out that the non-speech precursor effects observed by Lotto and Kluender (1998) failed to persist in a psychophysical task designed specifically to test frequency contrast, and only yielded perceptual biases for speech when presented in an intensity relation that is likely greater than that found in natural speech. Fowler (2006) argues that the perceptual processes that accommodate coarticulation effects in natural speech stem instead from perception of physical gestures of the talker. This argument is based at least partly in the observation that the /d/-/g/ bias can be observed even when the preceding syllable is entirely omitted, as long as the truncated stimulus was originally produced as a natural /alga/ or /arda/. That is, acoustic evidence for a segment spreads to neighboring segments, and is recoverable by listeners in the absence of the original segment. Although coarticulation compensation can occur both contextually and in isolation, it remains unclear as to what best explains listeners' perceptual accommodation to variability in the speech signal.

Multiple modalities can be used to provide context for phonetic identification. Fowler et al. (2000) as well as Strand and Johnson (1996) showed that perceptual phonetic bias could arise from visual as well as acoustic context. Furthermore, the

"McGurk effect," (traditionally described as the result of fusion of auditory and visual inputs to form one cohesive speech percept) can be elicited using a fusion of auditory and tactile information (Fowler & Dekle, 1991). Although no differences presented between participants who are sighted or visually impaired, the tactile stream became more influential when acoustic masking noise was added (Sato, Cavé, Ménard, & Brasseur 2010), suggesting an increased role for multi-modal perception in the face of acoustic signal degradation.

Multi-modal context effects are not universally embraced by all researchers. Holt et al. (2005) argue that visual influence on phonetic perception need not be tied to the articulatory properties of perceived segments; simple reliable covariance in the visual modality should be sufficient to elicit such context effects. Furthermore, lexically mediated (indirect) context effects identified by Elman and McClelland (1988) failed to replicate in a study by Vroomen and de Gelder (2001) that used visual instead of lexical fricative disambiguation. Holt et al. (2005) explain this by pointing to the likely interaction of concurrent (rather than sequential) information from different modalities. That is, the influence of multiple modalities is implemented as an integration rather than as a context effect.

The current study examines phonetic context effects of a particularly interesting kind. Influences on the perception of /s/ and /ʃ/ include acoustic segments that occur *after* the fricatives have been heard. Although Fowler (2006) argues that this kind of influence rules out an acoustic context account of the effects (since neurons presumably would be inhibited only by previously-occurring, not later-occurring information), that perspective suffers from the bias of treating syllables as being composed of discrete segments rather than entities that are implemented over a longer domain. Even though listeners can readily decompose syllables into segments, they still might capitalize on a longer time window when identifying syllables (or some other linguistic unit), and then *subsequently* decompose the syllable. Evidence for this qualification was observed by Read, Zhang, Nie and Ding (1986), who showed that listeners are not able to segment initial consonants of a syllable without having been trained to use an alphabetic orthographic system; Mandarin-speaking listeners that solely used a logographic writing system appeared to perceive sounds only at the syllabic level, without any further segmental decomposition. This implies that the influence of temporally sequential segments could be inter-dependent, and potentially completed the segment identification is reported.

Despite the lack of consensus on how to explain context effects as they relate to speech perception, the current investigation is undertaken with the view that spectral/phonetic context effects are consistently observed in normal hearing listeners and they appear to be helpful from the perspective of accommodating natural speech variability. This is characterized here as helpful behavior because the adjustments of phonetic identifications are consistently in a direction that is appropriate, given the direction of acoustic change in the signal. Simply, as the acoustics are shifted upward in frequency, the perceptual category boundary is accordingly shifted upward in frequency, and vice versa. Lotto and Holt (2006) argue that context effects arise from general auditory and cognitive mechanisms and imply that they should be impaired in those who use CIs, because their auditory perceptions are heavily compromised. In this investigation, we offer an exploration into just how much CI listeners can demonstrate spectral context effects in the perception of /s/ and /ʃ/ sounds, and whether visual cues can influence this context effect.

Acoustics and perception of sibilant fricatives

The fricatives /s/ and /ʃ/ are voiceless fricatives contrasted by place of articulation; the former is an alveolar sibilant and the latter is a post-alveolar/palatal sibilant. Spectral peak location and spectral mean for /s/ are higher than those for /ʃ/, owing to the smaller front resonating cavity resulting from the more anterior tongue position for /s/, as well as the presence of a sub-lingual resonance cavity in /ʃ/ (Jongman, Wayland and Wong, 2000). The exact locations of these spectral peaks are not entirely stable though, as the fricative noise spectrum is affected by vowel context. Talkers consistently assume articulatory postures for later sounds (i.e. coarticulate) during speech production, and this regressive coarticulation is particularly evident for these two sounds. The anticipatory rounded lip posture of the /u/ vowel has a frequency-lowering effect on the frication spectrum of the fricative (Kunisaki and Fujisaki, 1977). Fittingly, listeners make appropriate perceptual boundary shifts in accordance with contextual cues stemming from lip rounding; fricatives are more likely judged to be /s/ before round vowels like [u] and [o] than they are before unround vowels like [e], [i] or [a] (Kunisaki & Fujisaki, 1977; Mann & Repp, 1980). This boundary shift is commonly thought to arise from accommodation to coarticulation, although it is consistent with more general spectral contrast as well.

Apart from vowel context, there are other sources of variability in fricative acoustics, such as gender-related differences. Fricatives produced by women have higher-frequency spectral energy than those produced by men (Jongman et al., 2000). Although

Schwartz (1968) attributes this spectral difference to vocal tract size, results from Gonzalez (2004) and van Dommelen and Moxness (1995) show poor correlation between body-size and either formant frequencies or F0, both of which could be intuitively thought to vary with vocal-tract length. Furthermore, vocal tract size cannot be the only factor at play in gender-driven fricative variability, as there is evidence that gender-related spectral differences are at least partly explained by learned behaviors. For example, gender identification for children's speech is reliably good even prior to pubescent physical changes in the vocal tract (Perry, Ohde & Ashmead, 2001). Gender-related F0 differentiation is also mediated by culture; larger F0 separation exists between males and females in Japanese talkers compared to Dutch talkers (van Bezooijen, 1995). Additionally, the acoustic differences between male and female voices vary across languages in a way that is possibly attributable to varying degrees of gender role differentiation in different societies (Bladon, Henton & Pickering, 1984; Johnson, 2006).

Listeners are sensitive to the aforementioned gender-related spectral differences when perceiving fricative sounds. All other factors being equal, a listener is more likely to hear a fricative as /s/ if it is spoken by a male voice (Mann & Repp, 1980; Strand, 1999), particularly if that fricative is ambiguously between /s/ and /ʃ/. In an identification experiment by Strand and Johnson (1996) that used a continuum of words ranging from sod to shod, listeners heard more /s/-onset words when the talker was male; listeners in that study were also influenced by visual stimuli (male or female faces) in a way that was consistent with (but far less powerful than) the direction of the auditory effect. A follow-up study (Johnson et al., 1999) using a vowel continuum (between the vowels in "hud" and "hood") suggested that this gender compensation effect is not driven by F0; the gradient effect of gender influence was strongest for the talker judged to be most feminine, who actually had a lower F0 than the feminine talker judged to be less typical. In that study, an arbitrary male or female label on an ambiguous voice also had effects on phonetic labeling, providing further evidence that contextual influences can include non-acoustic stimulus attributes. This result contributed to an assertion by Strand (1999) that gender-related context effects are influenced by conceptions of stereotypicality formed by experience.

Gradient effects of gender-driven talker context effects were further explored by Munson, Jefferson and McDonald (2006), who used a continuum of /s-ʃ/ sounds appended to syllable codas spoken by 44 talkers previously judged on perceived sexuality. Results revealed that perceived sexuality influenced phonetic judgments in a way that is consistent with common societal impressions; fricatives appended to female voices were more likely to be heard as /s/ (the “male” direction of the effect) if the voice was rated as less feminine. Although this effect did not clearly emerge for male voices, the study suggested that listeners are sensitive not only to perceived gender, but also to the association between sexuality and fricative production.

The influence of talker gender on fricative perception is not limited to the contributions of voice acoustics. Strand and Johnson (1996) showed that the image of a male or female face was sufficient to influence /s/-/ʃ/ judgments by normal-hearing listeners even when auditory information was incongruent. This effect was quite modest compared to the acoustic (voice) context, but it still implicated the role of multi-modal influence in phonetic context effects. An important component of this result is that the visually mediated effects were identified for voices previously judged to be less-typically male or female. That is, the voice gender was somewhat unclear, which possibly promoted the influence of disambiguating visual information that may not have been used in a typical listening situation. With regard to the current study, it could be said that gender cues in the voice are compromised for CI listeners in a wholly different way, but still a way that promotes the recruitment of extra facilities, such as visual cues.

In addition to the effects of vowel rounding and talker gender / perceived sexuality, perception of fricatives is also influenced by the formant transitions at the onset of the following vowel. Although Heinz and Stevens (1961) and Harris (1958) downplayed the role of formant transitions in the /s/-/ʃ/ contrast (at least compared to that for the contrast between /f/ and /θ/), the transitions found in vowels that originally followed /s/ lead to more /s/ judgments across a synthetic fricative continuum, particularly for ambiguous tokens (Mann & Repp, 1980, Repp, 1981, Strand & Johnson, 1996). This effect is at least partly modulated by the age of the listener; Nittrouer and Studdert-Kennedy (1987) and Nittrouer and Miller (1997) have shown that children younger than 7 years old are influenced relatively more by formant transitions and relatively less by the spectrum of the fricative noise, compared to adults. Additionally, a mismatch between the fricative energy and the formant transitions slows judgment of fricatives that are unambiguous, suggesting that listeners are influenced by these transitions even when the fricative segments are unequivocal (Whalen, 1984). Furthermore, the aforementioned effect of vowel context is heavily diminished when the vowel formants are devoid of the onset transitions appropriate to the consonant constriction (Mann & Repp, 1980). Despite these findings, the role of formant transitions has not always been clear. When Repp (1981) replicated the Mann and Repp (1980) study with a more natural-sounding fricative continuum, the effects of the vowel environment and formant transitions were confirmed, but a small number of listeners seemingly neglected the vocalic context, suggesting that it is not entirely compulsory to accommodate coarticulation effects. Furthermore, the effect of vowel context becomes less clear when stimuli include unfamiliar vowels (Whalen, 1981).

Other explanations of vocalic context effects for the /s/-/ʃ/ distinction have emerged, including ones that stem from general ideas of auditory spectral contrast rather than articulation gesture perception (Fowler, 2006) or the perception of physical attributes like vocal track size or gender (Munson and Coyne, 2010). Consequently, some interpretations of the contextual effects of vocalic information concern the relative change in the spectral energy of the frication and the vowel segments in the region of the 3rd or 4th formant (Stevens, 1985). The energy in the fricative in this spectral region (roughly 2500 to 3400 Hz) is virtually nil for /s/, and therefore increases substantially as the vowel segment begins. For /ʃ/ on the other hand, there is strong frication energy in this region during the consonant segment, so there is relatively less change across the consonant-vowel transition (Jongman et al., 2000). The segmental contrast (and context effect) can thus be re-characterized as being driven by high F3 energy contrast (for /s/) versus low F3 energy contrast (for /ʃ/). This perspective is compatible with some explanations of speech perception as a process driven by cochlear (i.e. spectral) entropy (Lotto and Kluender, 1998; Kluender, Coady and Kiefte, 2003; Stilp and Kluender, 2010) in addition to the previous explanations of spectral context effects (Lotto & Holt, 2006). Naturally, studies that explore the role of relative spectral change are designed slightly differently than those previously discussed. For example, Hedrick and Ohde (1993) synthesized a continuum of /s/-/ʃ/ sounds that varied not along a low-high frequency continuum, but instead in relative amplitude of the frication energy in the region of the adjacent third vowel formant.

Context effects in cochlear implant listeners

Only a small number of studies have explored the influence of vocalic context in the identification of fricatives by CI listeners. Hedrick and Carney (1997) found that 4 adult cochlear implant (CI) users relied heavily upon the relative amplitude cue in the fricative segment, to the apparent neglect of formant transition cues. In fact, only one of the CI users in that study showed any use of the formant transition cue, and that was only when the relative amplitude cue was set to an ambiguous level. Compared to the CI users, NH adults in that study showed greater sensitivity to formant transitions in the vowel, but it should be noted that these vowels were synthetic; listeners consistently show artificially lower sensitivity to formant transitions when listening to synthetic speech (Nittrouer, 2005; Nittrouer & Lowenstein, 2008; Assmann & Katz, 2005). Thus, perhaps the use of this cue by NH participants in Hedrick and Carney's (1997) study may have been underestimated because of the stimuli used. A considerable number (12 of 26) of the adult CI users tested by Summerfield, Nakisa, McCormick, Archbold, Gibbon and Donaghue (2002) made significant use of vocalic information in tokens created from natural speech. However, adult and child CI users tested by Summerfield et al. made less use of vocalic information than the NH adults listening to normal speech or to stimuli processed to simulate the SPEAK coding strategy.

Although results from Summerfield et al. (2002) are in disagreement with those of Hedrick and Carney (1997), it should be noted that listeners in these studies used different processing strategies, and were exposed to stimuli that contained different kinds of vocalic cues. Listeners in Summerfield et al.’s (2002) experiment were exposed to three mutually cooperating natural vocalic cues rather than just the single synthetic formant transition cue used by Hedrick and Carney (1997). Specifically, those in Summerfield et al.’s experiment heard a male vowel /u/ spoken after a natural /s/ (the gender, vowel context and formant transitions all yield bias toward /s/), and a female-spoken vowel /i/ spoken after a natural /ʃ/ (the gender, vowel context and formant transitions all yield bias toward /ʃ/), attached to ambiguous fricative sounds. Thus, it is unclear from these results exactly what kind of vocalic cue was being used by the implanted listeners in their study, or whether CI users can benefit from any of these cues in isolation. Additionally, only 4 of 13 children tested by Summerfield et al. were able to use vocalic information – all four of these children used the implant for over 2.5 years. Thus, it remains apparent that the use of contextual (vocalic) information by CI listeners is impaired, and that CI users require at least some experience with their devices and perhaps the cooperation of multiple cues in order to facilitate contextual phonetic boundary shifting.

Summary of literature review

The /s/ and /ʃ/ sounds are distinguished primarily by the relative frequency of spectral peaks in the fricative noise, but the precise location of these peaks in the spectrum is affected by various influences, including vowel context and the gender of the talker producing the sound. Listeners accommodate to these and other sources of variance by adjusting their criterion for what is perceived as /s/ or /ʃ/. This is an example of a more general class of phonetic context effects that is explained at least partially by spectral contrast between adjacent segments. Because cochlear implant users are known to experience poor resolution in the spectral domain, they are not expected to exhibit the same amount of contextual accommodation. Not surprisingly, in the limited previous work that tested for contextual accommodation, CI listeners did not show effects that matched those observed from normal-hearing listeners. In the current study, we seek to clarify the separate effects of gender, formant transitions and vowel context in the perception of voiceless fricatives by people who use cochlear implants. In line with voluminous previous literature addressing speech perception by this population, normal-hearing listeners were also tested with the same stimuli, and they were also tested in CI simulations using noise-band vocoding.

Rationale and hypotheses

There are at least two rationales for testing context effects. First is to add a layer beyond phonetic cue-weighting (Chapters 2 and 3) that provides even more finely-grained insight on the perception of acoustic cues. Specifically, in moving beyond the perception of phonetic features and acoustic cues in isolation, experiments presented in this chapter examine the perception of cues in changing contexts, where absolute cue levels convey different information depending on other parts of the signal. It is likely that a listener experiencing signal degradation would not demonstrate context-dependent listening strategies since the degradation would obscure not only the cues in question, but also the distinctions between the various contexts that call for dynamic use of those cues.

In view of the limitations of CI listeners in the spectral domain, it is hypothesized that they will not take full advantage of spectral context, because the cues that define context are themselves degraded. On the other hand, they are expected to recruit information from domains that remain relatively less affected by the signal degradation inherent in electric hearing. Because gender-related spectral differences may operate at the gross level (such as overall spectral shape), they should resist spectral degradation, whereas finer spectral properties (like formant frequencies for vowels or consonant transitions) are likely to be heavily compromised. Therefore, CI listeners and NH listeners in CI simulation were predicted to show some accommodation to gender context, but not for vowel and formant transition contexts.

Experiment 4: Auditory context effects for the /s/-/ʃ/ contrast: Effects of spectral degradation and electric hearing

Hypotheses

Phonetic context effects have been measured in various previous experiments, so the effects published by Mann and Repp (1980) were expected to replicate for NH listeners. In view of the poor spectral resolution available to listeners with cochlear implants, it was predicted that they would show extremely small (if any) accommodation to acoustic context. Since the cochlear implant simulations do not incorporate the additional difficulties of spectral warping (the upward shifting of the spectrum resulting from variable or imperfect surgical implant insertion depth), it was expected that normal-hearing listeners in the spectrally degraded conditions would show context effects that were smaller than those in the unprocessed condition, but greater than those demonstrated by listeners with cochlear implants.

Method

Participants.

Participants included 10 adult (mean age 21.9 years; 8 female) listeners with normal hearing, defined as having pure-tone thresholds ≤20 dB HL from 250–8000 Hz in both ears (ANSI, 2010). A second group of participants included 7 adult (age 50-73; mean age 63.7 years; 3 female) recipients of cochlear implants. CI users were all post-lingually deafened. Six were users of the Cochlear Freedom or N24 devices; one used the Med-El device. All participated in Experiments 1 and 2. See Table 2.1 for demographic information and speech processor parameters for each CI user. All participants were native speakers of American English.

Stimuli.

Stimuli were comprised of a fricative and a vowel. The fricative was aperiodic noise that was synthesized to match /s/, /ʃ/, or an ambiguous intermediate sound. The vowels were natural utterances excised from recordings of the words “see,” “she,” “sue,” and “shoe” spoken by four native speakers of American English (two females, two males). These words were recorded in a double-walled sound-treated room using an AKG C1000 microphone at 44.1 kHz sampling.

Fricative synthesis. The fricative continuum contained nine synthetic items that varied between /s/ and /ʃ/, modeled after productions of sounds in real words produced by various native speakers (both male and female) recorded at the University of Maryland. Items in the continuum were comprised of three spectral peak prominences (SPPs) that varied in three dimensions: frequency (center of the noise band), bandwidth, and relative amplitude of the peaks (spectral tilt). SPP frequencies and bandwidths were linearly interpolated along a log-frequency scale (to model the non-linearity in the human auditory system) between unambiguous levels for /s/ and /ʃ/. Spectral tilt was linearly interpolated between values observed in the natural utterances. Table 4.1 contains details of all the parameters of this continuum. Figure 4.1 is a schematic representation of the SPPs and Figure 4.2 shows example spectrograms of the synthetic and natural speech tokens.

To create the fricatives, 180ms of white noise was first multiplied by an amplitude contour modeled after that for a natural fricative. It contained a non-linear ramping envelope that reached its peak 115 ms after onset. The offset decaying amplitude ramp was 18 ms in duration. All fricatives were equal in duration. This broadband noise was then filtered into individual SPPs using the following procedure: A SPP with frequency *f* and bandwidth *b* was created using the bandpass filter function in the Praat software (Boersma & Weenik, 2011) in four sequential steps. The basic premise of this process was to choose the center frequency as the intersecting boundary of adjacent low-pass and high-pass filters, each having a specific smoothing function that defined the bandwidth of the resulting band of noise. First, the low pass was set by filtering from 0 to *f* Hz, with *b* Hz smoothing. Next, the high pass was set from *f* to 12000 Hz, with *b* Hz smoothing. Then, a second bandpass filter was created that contained only a 200 Hz-wide narrow band of energy centered by *f*. Finally this narrow band was added to the previously-described wider-bandwidth SPP at +3 dB to create a well-defined spectral peak at *f*. This process was repeated for all three SPPs for each stimulus item. Center frequencies *f* were varied in equal steps along a log scale between those that define clear /ʃ/ and clear /s/.

Following this bandpass filtering, all three SPPs for each stimulus were roughly the same amplitude since the original white noise spectrum was flat. To apply appropriate spectral tilts, the middle SPP was unchanged, but the first and third were relatively amplified or attenuated relative to the middle SPP, depending on their place in the continuum. Consistent with measurements of the natural utterances, fricatives at the /ʃ/-end of the continuum contained falling spectral tilts (lower-frequency components were more intense than higher-frequency components) while those at the /s/-end contained rising spectral tilts (higher-frequency components were more intense than lower-frequency components). The tilt was linearly interpolated along the 9 continuum steps.

Following the attenuation of each SPP, they were added together to produce the final fricative segment. These fricative segments were appended to the onset of natural utterances of /i/ and /u/ segmented from recordings of “see,” “she,” “sue” and “shoe” by four native speakers of English (two male, two female). The resulting 144-item stimulus set consisted of 9 fricatives x 2 vowel environments x 2 formant (natural consonant) environments x 4 talkers. All of the vowel, formant and talker (gender) information was conveyed via the vocalic segment following the fricative, and each factor was fully crossed.

Figure 4.1. Continuum of fricative sounds in Experiment 4

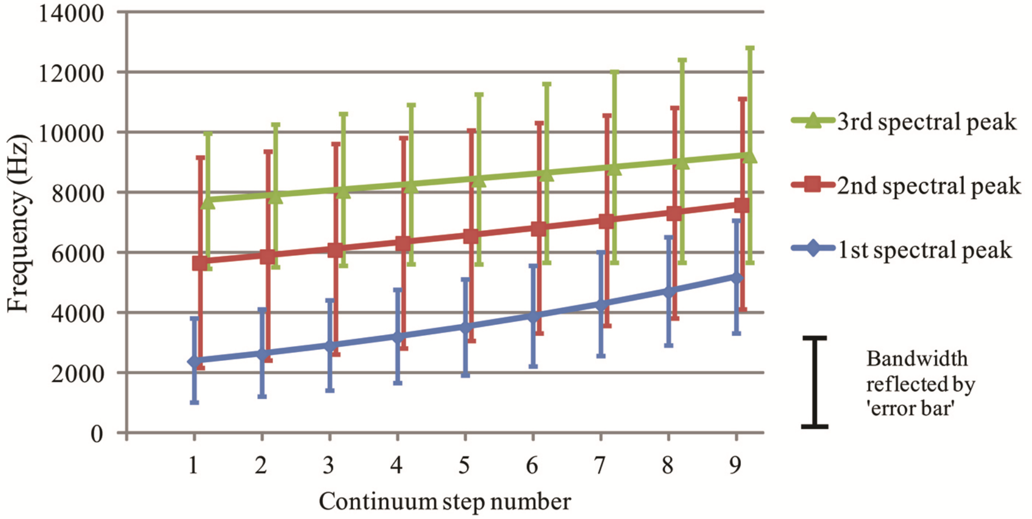

Figure 4.1. Schematic illustration of the fricative continuum, with each step consisting of three spectral peaks of varying bandwidths.

Figure 4.2. Spectrograms of stimuli in Experiment 4

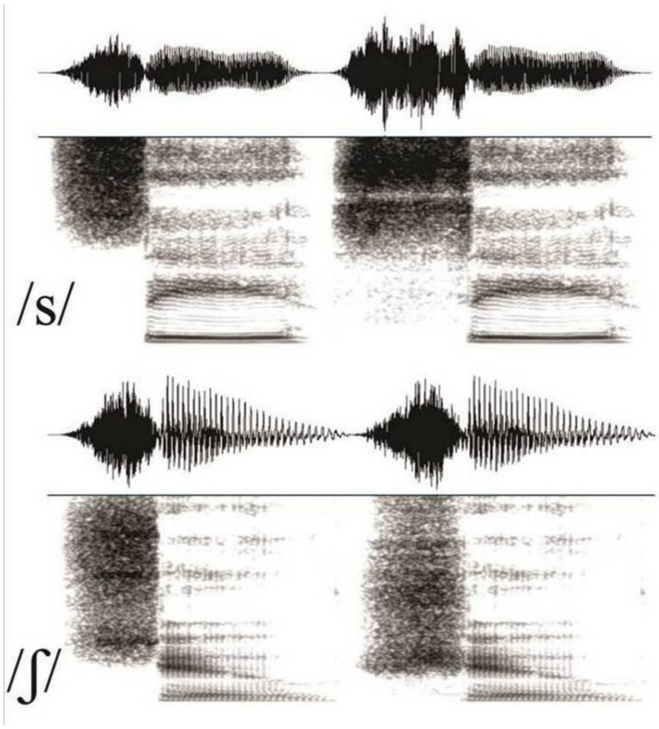

Figure. 4.2. Spectrograms of synthetic and natural fricatives that were appended to natural vocalic contexts in the experiment. The upper panel depicts /s/ in the word “see” spoken by a female. The lower panel depicts /ʃ/ in the word “she” spoken by a male. Left spectrograms depict synthetic/manipulated stimuli; natural utterances are depicted to the right.

Table 4.1.

Acoustic description of the fricative continuum in Experiment 4

Continuum step:	1	2	3	4	5	6	7	8	9
SPP1 (Hz)	2421	2664	2932	3226	3550	3906	4298	4729	5203
SPP2 (Hz)	5700	5911	6130	6357	6592	6837	7090	7352	7625
SPP3 (Hz)	7741	7918	8099	8283	8472	8666	8863	9065	9272
BW1 (Hz)	1397	1448	1501	1556	1612	1671	1732	1796	1861
BW2 (Hz)	3500	3500	3500	3500	3500	3500	3500	3500	3500
BW3 (Hz)	2245	2378	2520	2670	2828	2997	3175	3364	3564
SPP1 mix (dB)	3.3	2.5	1.7	0.8	0	-0.8	-1.3	-1.7	-2.1
SPP2 mix (dB)	0	0	0	0	0	0	0	0	0
SPP3 mix (dB)	-3.3	-2.5	-1.7	-0.8	0	0.8	1.3	1.7	2.1
SPP1 (Log Hz)	3.38	3.43	3.47	3.51	3.55	3.59	3.63	3.67	3.72
SPP2 (Log Hz)	3.76	3.77	3.79	3.80	3.82	3.83	3.85	3.87	3.88
SPP3 (Log Hz)	3.89	3.90	3.91	3.92	3.93	3.94	3.95	3.96	3.97
BW1 (Log Hz)	3.15	3.16	3.18	3.19	3.21	3.22	3.24	3.25	3.27
BW2 (Log Hz)	3.54	3.54	3.54	3.54	3.54	3.54	3.54	3.54	3.54
BW3 (Log Hz)	3.35	3.38	3.40	3.43	3.45	3.48	3.50	3.53	3.55

Note: the continuum steps were interpolated using a log scale. Those values are presented here along with a conversion to Hz values for ease of interpretation. Numerals 1, 2 and 3 refer to the spectral peak prominences from lowest to highest.

Spectral degradation: Noise-band vocoding. For the NH listeners in simulated conditions, spectral resolution was degraded using noise-band vocoding (NBV), which has become a common way to simulate a cochlear implant (see Shannon et al., 1995). This was accomplished using online signal processing within the iCAST stimulus delivery software (version 5.04.02; Fu, 2006). Stimuli were bandpass filtered into 8 frequency bands using sixth-order Butterworth filters (24 dB/octave). This number of bands was chosen to best approximate the performance of CI users (Friesen et al., 2001).

The temporal envelope in each band was extracted by half-wave rectification and low-pass filtering with a 300-Hz cutoff frequency, which is sufficient for good speech understanding (Shannon et al., 1995), and for temporal coding of the F0 of all the talkers. The envelope of each band was used to modulate the corresponding bandpass-filtered noise. Specific band frequency cutoff values were determined assuming a 35 mm cochlear length (Greenwood, 1990) and are listed in Table 4.2. The lower and upper frequency cutoffs for the analysis and carrier bands were 150 and 10000 Hz, respectively, to approximate those commonly used in modern CI speech processors (Başkent & Shannon, 2003).

Table 4.2.

Analysis & carrier filter bands for the noise-band vocoding scheme in Experiment 4

Channel	1	2	3	4	5	6	7	8
High-pass (Hz)	150	314	570	967	1586	2549	4046	6376
Low-pass (Hz)	314	570	967	1586	2549	4046	6376	10000

Procedure.

All speech recognition testing was conducted in a double-walled sound-treated booth. Volume level was calibrated at the position of the listener’s head using a Radio Shack sound level meter that referenced a 1 kHz tone that was equated in RMS amplitude to the speech stimuli. Stimuli were presented at 65 dBA in the free field through a single Tannoy Reveal studio monitor loudspeaker (frequency response: 65 Hz – 20 kHz) at a distance of 1 – 2 feet placed in front of the listener at eye level. Each token was presented once, and listeners subsequently used a computer mouse to select the word that they perceived. Stimuli were presented in alternating blocks of spectral resolution (unprocessed or 8-channel NBV), and presentation of tokens within each block was randomized. In this self-paced task, the 144 stimuli were each heard 5 times in each condition of spectral resolution. Listeners with cochlear implants only heard the unprocessed items.

Analysis.

Listeners’ responses (coded as 1 or 0 referring to /s/- or /ʃ/-onset word choices) were fit using a generalized linear (logistic) mixed-effects model (GLMM). This was done in the R software interface (R Development Core Team, 2010), using the lme4 package (Bates & Maechler, 2010). A random effect of participant was used, and the fixed-effects were the stimulus factors described above. The binomial family call function was used because the possibility of an /s/ response could not logically exceed 100% or fall below 0%. This resulted in the use of the logit link function, and an assumption that variance increased with the mean according to the binomial distribution. Parameter levels were centered around 0, since the R GLM call function sets “0” as the default level while estimating other parameters. Although the spectral peak prominences are represented on the figures using the Hz frequency scale, they were identified in the statistical model using log Hz. SPP of 5203 Hz is 3.716 in the log Hz scale, and was identified in the analysis as 0.166, since it was 0.166 greater than the central value in the continuum (3.55 log Hz). This factor level stood to represent all of the co-varying changes in the upper SPPs as well as the spectral tilt. Although all of these factor components are tied together, no claims are being made regarding which is the most relevant for perception.

Starting with an intercept-only model, factor selection (i.e. the inclusion of talker gender as a response predictor) was done using a forward-selection hill-climbing process whereby candidate factors were added one-by-one; that which yielded the highest significance was kept. Subsequent factors (or factor interactions) were retained in the model if they significantly improved the model without unnecessarily over-fitting. The ranking metric was the Akaike information criterion (AIC) (Akaike, 1974), as it has become a popular method for evaluating mixed effects models (Vaida & Blanchard, 2005; Fang, 2011). This criterion measures relative goodness of fit of competing models by balancing accuracy and complexity of the model. The model tested whether the coefficient of the resulting estimating equation for an acoustic cue was different from 0 and, crucially, whether the coefficient was different across conditions of spectral resolution. There were two sets of data: 1) NH listeners in different conditions of spectral resolution and 2) CI users listening to the unprocessed sounds.

Although the use of cues for phonetic perception are commonly assessed using psychometric slope coefficients (Mayo & Turk, 2005; Munson & Nelson, 2005; Morrison, 2005; Morrison & Kondaurova, 2009), the goal of the current study was not to measure the efficiency of listeners in their use of the cues for the fricative continuum. Instead, the goal was to assess whether the binary contextual factors (talker gender, vowel context, vocalic formant transitions) influenced the /s/–/ʃ/ identification function shift along the frequency scale. There are two simple measures that test for this. The first is to simply take the proportion of tokens along the fricative continuum that were labeled as /s/ in the different vocalic contexts (Summerfield et al., 2002). Along the entire continuum of fricatives, more of them should be labeled as /s/ when appended to the male voices (or the /u/ vowels), compared to those appended to the female voices (or the /i/ vowels). Another way is to interpolate the category crossover boundary (i.e. 50% crossover point) in all of these various contexts; a listener's category boundary for fricatives spoken by male talkers should be observed at a lower frequency than that for female talkers. The former method incorporates data from the entire continuum, while the latter method yields a result that is more interpretable with regard to speech acoustics. Both of these methods were explored here in addition to the generalized linear model.

Results

Identification functions for NH listeners, NH listeners in the CI simulations, and the CI listeners in the various vocalic contexts for are shown in figures 4.3, 4.4 and 4.5, respectively. Difference scores quantifying the use of vocalic cues are illustrated in figure 4.6 and the category boundary effects are illustrated in figure 4.7. These data all replicate earlier findings of context effects for normal-hearing listeners; the effect of talker gender had the strongest effect on the identification functions, followed by vowel context and formant transitions. The generalized linear mixed-effects models were described optimally as follows:

1) *Perception by NH listeners in different conditions:*
/s/ perception ~ SPP + gender + SR + gender:SR + vowel + vowel:SR + formant + gender:vowel + SPP:SR + formant :SR + SPP:formant + SPP:vowel

2) Perception by CI listeners:
/s/ perception ~ SPP + gender + vowel + SPP:gender + SPP:vowel + SPP:gender:vowel

For these two models, the interaction between two factors A and B is indicated by A:B. Independent factors are indicated by "+." "SR" refers to spectral resolution (either unprocessed/natural speech or signals degraded with the noise-band vocoder). SPP means spectral peak prominence (defining the continuum of fricative spectra) and (1|Participant) is a random effect of participant.

For the normal hearing listeners, there were significant main effects of fricative spectrum ($p < 0.001$), talker gender ($p < 0.01$), vowel context ($p < 0.05$) and formant transition ($p < 0.05$). All of these effects went in the predicted direction; there were more /s/ perceptions when talkers were male, when the vowel context was /u/, and when formant transitions were from a natural /s/ environment. There was also a significant effect of spectral degradation (Fig 4.3 v. 4.4; $p < 0.001$); there were more /s/ perceptions in the degraded condition. Spectral degradation interacted significantly with all four main effects (all $p < 0.001$). There were virtually no context effects in the degraded condition; context effects were only significant when the signals were unprocessed. The effect of vowel context was stronger for the male talkers ($p < 0.001$), perhaps because of stronger /u/-fronting by the female talkers. There were also significant interactions between fricative spectrum and both formant transition and vowel context (both $p < 0.05$); these contexts exerted strongest effects for items in the center of the fricative continuum.

For the cochlear implant listeners, there were significant main effects of fricative spectrum, talker gender and vowel context (all $p < 0.001$). The effect of formant transition did not reach significance ($p = 0.103$), and its inclusion did not improve the model according to the AIC metric. There were significant interactions between fricative spectrum and gender ($p < 0.01$) and between fricative spectrum and vowel ($p < 0.001$). There was a significant three-way interaction between fricative spectrum, gender and vowel ($p < 0.05$).

Figure 4.3 Context effects for NH listeners in the unprocessed condition in Experiment 4

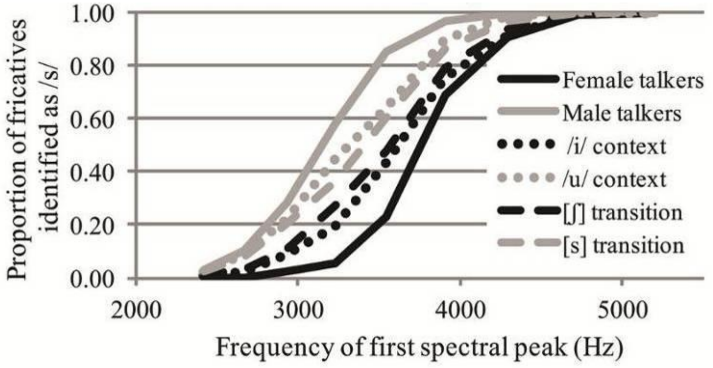

Figure 4.3. Response function for /s/ labeling along the continuum of fricative sounds for the NH listener group in the unprocessed (natural) sound condition. Solid lines indicate gender effects, dotted lines indicate vowel context effects and barred lines indicate formant transition effects. For each cue, responses are collapsed across all levels of the other cues. Black lines depict cue levels expected to produce fewer /s/ responses; all effects went in the predicted direction.

Figure 4.4 Context effects for NH listeners in the spectrally-degraded condition in Experiment 4

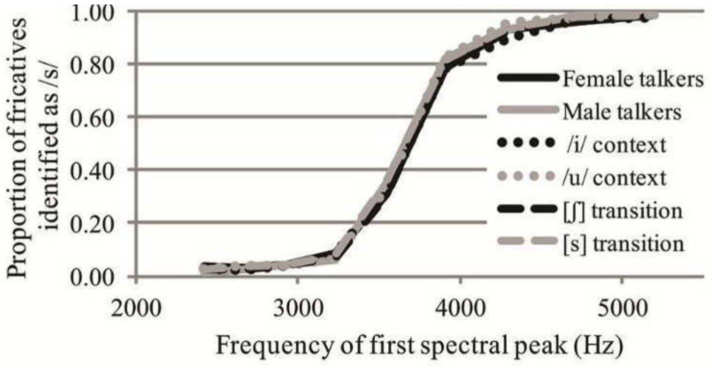

Figure 4.4. Response function for /s/ labeling along the continuum of fricative sounds for the NH listener group in the degraded (noise-band vocoder) condition. Solid, dotted and barred lines in both black and gray are virtually indistinguishable because of the general lack of context effects arising from any of the cues.

Figure 4.5 Context effects for CI listeners in Experiment 4

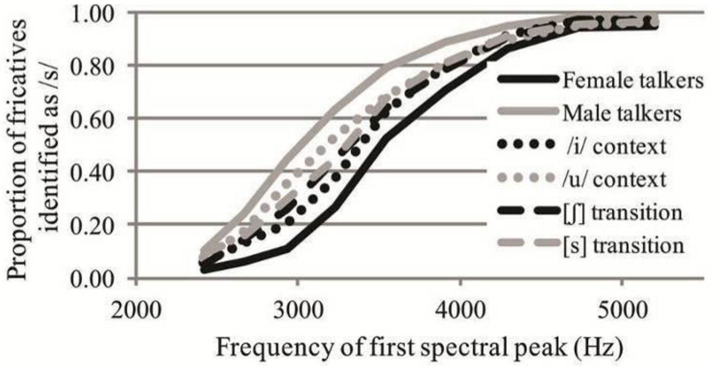

Figure 4.5. Response function for /s/ labeling along the continuum of fricative sounds for the CI listener group. Solid lines indicate gender effects, dotted lines indicate vowel context effects and barred lines indicate formant transition effects. For each cue, responses are collapsed across all levels of the other cues. Black lines depict cue levels expected to produce fewer /s/ responses; all effects went in the predicted direction.

Direct parametric statistical comparisons were not done between the NH and CI listener groups because of the age differences and high amount of within-group subject variability in the CI group, including duration and etiology of deafness, processing schemes, etc. It appears that CI listeners show less accommodation to vocalic context effects since the overall difference scores and boundary shifts were lower for the CI group compared to NH group. CI listeners did out-perform NH listeners in simulations, who were virtually unaffected by the changing contexts. Figure 4.6 illustrates a rough quantification of the context effects in this experiment. Across the entire continuum of fricatives, the number of /s/ perceptions in the context of female voices, /i/ vowels or /ʃ/ formant transitions is subtracted from the corresponding number of /s/ perceptions in the context of male voices, /u/ vowels or /s/ formant transitions, respectively. Thus, if 65% of the fricatives spoken by males are heard as /s/, but only 43% are heard as /s/ when spoken by females, the difference score would be 22%. Figure 4.6 illustrates the difference scores attributable to each of the three context effects for all three listener groups.

Figure 4.6 Context effects in Experiment 4: Difference scores

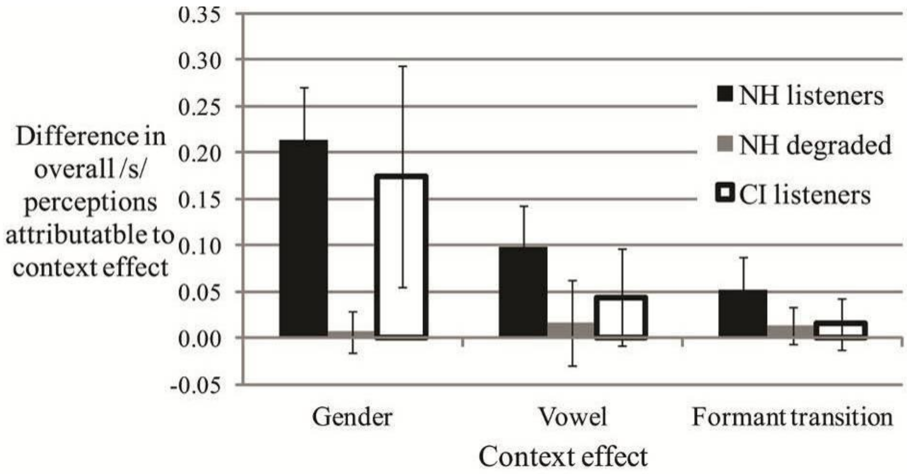

Figure 4.6. Difference score bar graphs for three context effects in Experiment 4. Bar height indicates the proportion of /s/ perceptions of one level of the context factor (e.g. male voice) minus the proportion of /s/ perceptions at the other level (e.g. female voice). Positive numbers reflect boundary shifts in the direction consistent with voice acoustics. Error bars indicate standard deviation.

An advantage of the generalized linear model is that it permits the interpolation of the 50% crossover point in /s/-/ʃ/ identification. The spectral peak frequency that yielded 0.5 odds of /s/ perception was identified for each context. Boundaries for the expected low-frequency boundary context (e.g. male voice) were subtracted from that for the predicted high-frequency boundary context (e.g. female voice) to produce boundary shifts in Figure 4.7. The figure complements the results reported above by providing a conversion of context to the frequency domain, which was used to specify the fricatives in question. The graph reinforces the conclusion that CI listeners demonstrated greater contextual adjustment than NH listeners that were listening in the spectrally degraded condition, but not the same amount as NH listeners hearing the regular unprocessed stimuli.

Figure 4.7 Context effects in Experiment 4: Boundary shifts

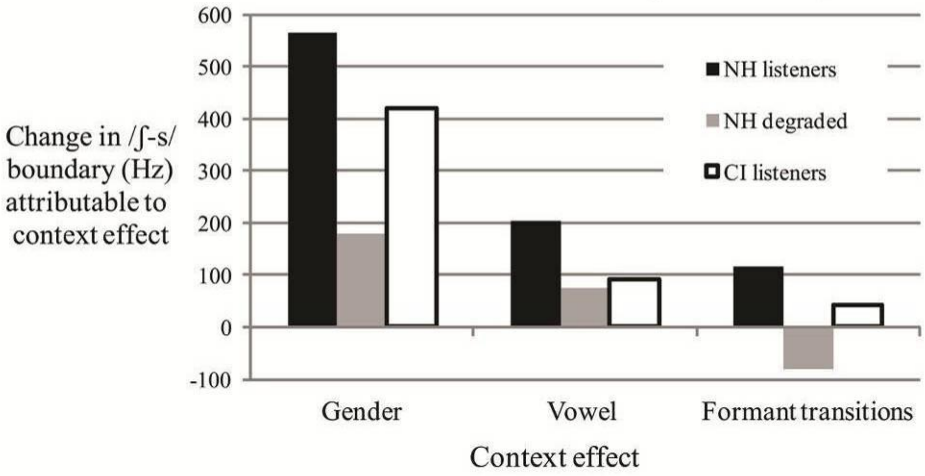

Figure 4.7. Boundary shifts predicted by the logistic models fit to the data for each listener group in Experiment 4. Bar height reflects the shift (in Hz) in the 50% crossover point between /s/ and /ʃ/ for one level of the context factor (e.g. male voice) minus the proportion of /s/ perceptions at the other level (e.g. female voice). Positive numbers reflect boundary shifts in the direction consistent with voice acoustics.

Because the analyses presented thus far do not speak to perceptual *accuracy* per se, a final analysis was conducted to evaluate the identification of stimuli where all the acoustic cues cooperated to confer a typical /s/ or /ʃ/ segment. Identification of these stimuli could appropriately be evaluated for correctness. Figure 4.8 illustrates performance levels for all listener groups for the stimuli at the continuum endpoints for both the female and male talker groups. Results suggest that both the /s/ and /ʃ/ fricatives were identified reliably by listeners in all conditions; performance was always above 90% and was lowest for /ʃ/ fricatives spoken by male talkers heard by CI listeners (this was still roughly 95% correct). Essentially, this figure implies that the potentially different perceptual strategies taken by listeners in this experimental task did not necessarily result in differences in identification accuracy.

Figure 4.8 Accuracy of /s/ and /ʃ/ identification in Experiment 4

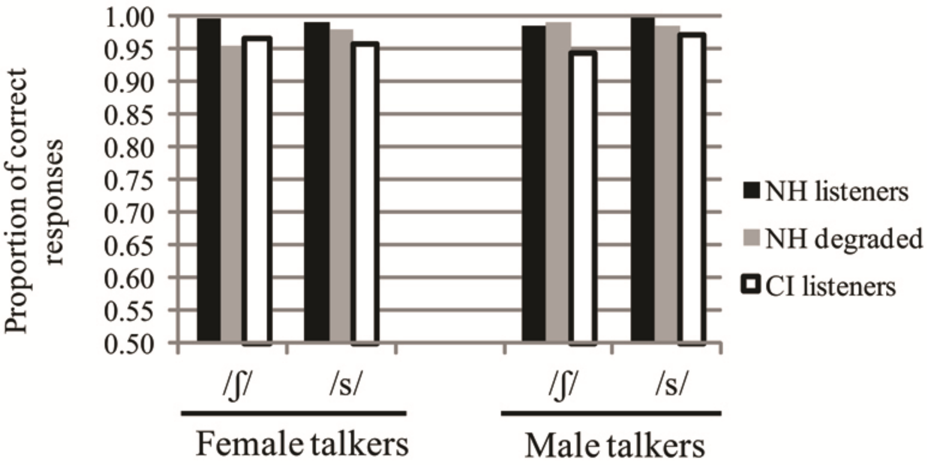

Figure 4.8 Mean accuracy in identification of stimuli at continuum endpoints by different listener groups for female and male talkers in Experiment 4. The /s/ and /ʃ/ items for this analysis were limited to those in which fricative spectrum and formant transition cues cooperated appropriately at continuum endpoints to signal the same segment.

Conclusions

In this experiment, listeners were presented with syllables consisting of a fricative and a vowel. Fricatives varied gradually between /s/ and /ʃ/ sounds; these were appended to various vocalic contexts consisting of talkers of both genders speaking two vowels (/i/ and /u/) spliced from natural utterances that contained formant transitions appropriate for either /s/ or /ʃ/. Normal-hearing listeners heard these words with clear unprocessed spectral resolution and also through an 8-channel noise-band vocoding scheme to simulate the use of a cochlear implant. Cochlear implant listeners heard only the natural/unprocessed words.

Consistent with earlier literature, listeners with normal hearing were more likely to identify fricatives as /s/ when they were 1) perceived to be spoken by male voices, 2) in the context of a rounded vowel, or 3) followed by formant transitions appropriate for /s/. That is, listeners were influenced by the context in which the fricative was heard.

Cochlear implant listeners showed context effects of a similar type but to a lesser degree. The presence of context effects at all was somewhat surprising, given the limitations of spectral resolution in cochlear implants. The differences between alternate contexts in this experiment (male/female voices, /i/-/u/ vowels, /s/-/ʃ/ formants) are all described mainly by spectral properties such as formant spacing, spectral tilt, voice pitch, vowel formants and dynamic formant movements. These are exactly the types of acoustic cues that are compromised in CI listeners as a result of cochlear nerve damage or dead regions, a limited number of active implant electrodes, and electrical interactions between those electrodes. Additionally, the spectral information delivered by an implant is frequently shifted upwards in frequency (a basal shift) as a result of surgical and physical limitations; this further distorts the spectral resolution in electric hearing. Implanted listeners in this experiment demonstrated that in spite of all these apparent adversities, it is possible to exhibit sensitivity to vocalic context cues when listening with a CI. Mann and Repp (1980) showed that the effects of vowel context are diminished when the vowel is devoid of appropriate onset formant transitions. Consistent with this observation, reduced effects of vowel context were observed in CI listeners in this study, who did not reach significance in their use of the formant transitions. Thus, although CI listeners can identify the /s/ and /ʃ/ sounds with excellent accuracy (Figure 4.8), they exhibit perceptual patterns that deviate from those of NH listeners.

In conditions that are thought to roughly simulate the use of a cochlear implant, normal-hearing listeners showed virtually no accommodation of vocalic context when labeling fricative sounds. There are a number of possible explanations for this dramatic difference between results of CI listeners and NH listeners in CI simulations. Even though previous literature suggests that NH listeners in simulations should do at least as well as CI listeners (Friesen et al., 2001), performance was typically measured with word recognition or phoneme recognition, where success is possible without very good spectral resolution. In the task presented here, the performance metric depended much more strongly on the perception of spectral cues. Thus although the noise-band vocoder with 8 channels is a relatively good predictor of CI listener performance for consonant and vowel recognition, it might not accurately predict how CI listeners use some spectral cues, particularly when they only affect consonant identification in subtle ways. In other words, the CI simulations show that while NH listeners are better at some abilities that underlie speech perception generally (such as the use of alternative acoustic cues), they might not be better than CI listeners at solving the problem of spectral degradation specifically. Essentially the noise-band vocoder condition is akin to an “initial activation” of an implant, since most of the normal-hearing listeners had no prior experience with this kind of signal. The implanted listeners had years of experience with their devices, and may have learned to take advantage of signal properties that are not easily accessible to the naïve listener.

Experiment 5: Auditory and visual context effects for the s-ʃ contrast: Effects of spectral degradation and electric hearing

A second phonetic context experiment was designed to test the influence of visual context cues on the perception of fricatives. The current experiment contained important elements used by Strand and Johnson (1996), with a number of important differences. Strand and Johnson selected talkers whose gender characteristics were ambiguous; the two talkers chosen for Experiment 5 were those that produced the most dissimilar response functions in the audio-alone condition. That is, they were the most "male" or "female" voices[3]. As a result, the strong auditory context effects rendered the experiment very conservative with regard to the visually mediated effects, since they are likely to be overpowered by strong acoustic cues rather than being promoted by ambiguous acoustic cues. The design of Experiment 5 was very similar to Experiment 4, but with only two talkers and only one configuration of formant transitions (those for /s/), to reduce the stimulus set that was multiplied by the inclusion and crossing of multiple visual cues.

Hypotheses

It was hypothesized that the acoustic context effects of talker gender and vowel environment found in Experiment 4 would persist with the addition of visual cues in

[3] *The female voice that yielded the fewest /s/ responses from Experiment 1 was not actually that with the higher mean F0, (consistent with Johnson et al., 1999). The female voice with the lower F0 had a breathier voice quality than the other female voice; the male voice chosen for Experiment 2 was lower-pitched and creakier than the other male voice.*

Experiment 5. Since the acoustic gender cues in this experiment were chosen to be unambiguous, it was expected that listeners with normal hearing would not need to rely on any additional cues to discern gender. Therefore, the visual gender (face) cues observed by Strand and Johnson (1996) were expected to have only negligible (if any) effects. The effects of visual cues for gender and for lip-rounding were expected to be larger for listeners with cochlear implants or in CI simulations, compared to normal-hearing listeners in the unprocessed condition. Gender-related context effects were expected to be largest when auditory and visual cues were complementary rather than conflicting.

Method

Participants.

Participants included 10 adult (mean age 22.2 years; 6 female) listeners with normal hearing, defined as having pure-tone thresholds ≤20 dB HL from 250–8000 Hz in both ears (ANSI, 2010). Seven of these listeners also participated in Experiment 4. A second group of participants included 7 adult CI users who also participated in Experiment 4. See Table 4.1 for demographic information and speech processor parameters for each CI user. All participants were native speakers of American English.

Stimuli.

Acoustic components. The fricative sounds in Experiment 5 were identical to those used in Experiment 4, with three-quarters of them omitted. The entire fricative continuum was used, but Experiment 5 featured only two of the four talkers, and only half of the formant configurations (omitting the vowels with ʃ formant transitions). Both vowels /i/ and /u/ were used.

Spectral degradation: Noise-band vocoding. Noise-band vocoding was done offline using TigerCIS (Fu, 2010). The analysis and carrier filter parameters were exactly the same as those used for Experiment 4.

Video stimulus recording. Video and reference audio (for temporally aligning audio and video) were recorded concurrently with a Canon DM-XL1 video camera onto digital videotape (mini DV; 29.97 fps). An adult female and adult male (both native speakers of English) were recorded while seated in front of a light gray background. No special effort was made to highlight the mouth of the talkers (i.e., no spotlights or any other special lighting or camera angles were used to emphasize the oral cavity). The talkers were instructed to start and end from a neutral "resting" mouth position and to avoid blinking during syllable articulation. The list consisted of the syllables /si su ʃi ʃu/ repeated at least five times each by each talker, prompted by the experimenter (off camera). See Figure 4.9 for examples of screenshots of the videos used in Experiment 5.

Video Editing. Video files were imported in audio-video interleave (AVI) format to a MacBook Pro running Windows XP using VirtualDub (www.virtualdub.org) and segmented as individual AVIs. For each talker, one video token of each syllable type was selected that best matched the synthesized fricative and vowel durations and that contained the least amount of movement or blinks during articulation. This resulted in eight video tokens (four for each speaker) to be combined with the auditory stimuli. Segmented AVI files were converted to grayscale to reduce visual complexity (to be compatible with a planned electrophysiological experiment), cropped to include only the talkers' upper shoulders and face (final dimensions: 425x425 pixels), and compressed with Cinepak in VirtualDub. The average duration of the audio-visual stimuli was 2.0 seconds (range: 1.87 s – 2.26 s, standard deviation = 0.14 s).

Audio-visual dubbing. The alignment was controlled tightly so that the audio from the video recordings could be replaced with the audio signals from Experiment 4, in view of the perceptual effects already established for these sounds. The waveform of the reference audio was first extracted from each video file (see above). The audio materials from Experiment 4 were then modified with onset or offset silence using Praat (Boersma & Weenik, 2011) to properly align to this source reference. The source reference audio was then replaced with the modified pre-existing audio for each condition and saved in AVI format in Virtual Dub.

Audio-visual crossings. Because context effects on fricative perception were the focus of this investigation, conflicting cross-modal vowel quality was not presented (i.e., no /u/ auditory tokens were dubbed onto /i/ visual tokens). However, all other crossings were performed: 2 visual talkers (male, female) x 2 audio talkers (male, female) x 9 auditory continuum tokens x 2 visual lip configurations (rounded or retracted). This resulted in a total of 144 dubbed audio-visual tokens for each condition (unprocessed/natural and spectrally-degraded).

Figure 4.9 Video screenshots for Experiment 5

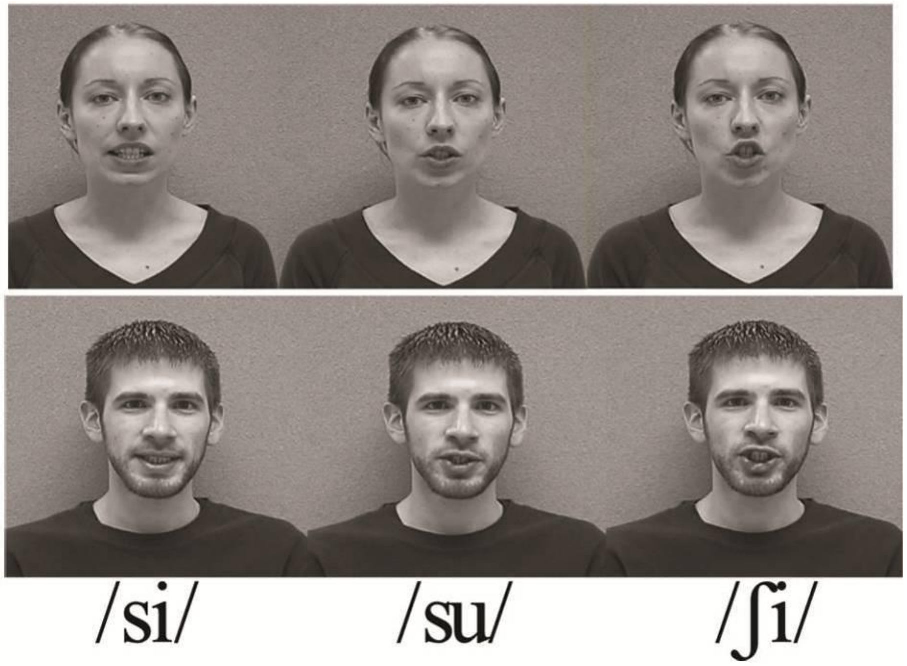

Figure 4.9. Screenshots of the videos used in Experiment 5. Images on the left show stimuli with no lip-rounding, visually consistent only with /si/. Center images show stimuli with some lip-rounding that arose from vowel coarticulation in /su/, but is also visually consistent with /ʃu/; consonant lip-rounding could arise because of the consonant itself or coarticulation from the vowel. Images on the right show stimuli with heavy lip-rounding that arose from the /ʃ/ articulation; these are consistent with either /ʃi/, /ʃu/ or /su/, but not with /si/. Lip-rounding for /ʃ/ sounds followed by the /u/ vowel were virtually indistinguishable from those followed by the /i/ vowel.

Procedure.

All speech recognition testing was conducted in a double-walled sound-treated booth. Volume level was calibrated at the position of the listener's head using a Radio Shack sound level meter that referenced a 1 kHz tone that was equated in RMS amplitude to the speech stimuli. Stimuli were presented at 65 dBA in the free field through a single Tannoy Reveal studio monitor loudspeaker (frequency response: 65 Hz – 20 kHz) at a distance of 1 – 2 feet placed in front of the listener at eye level. Each token was presented once, and listeners subsequently used a computer mouse to select the word that they perceived (see, she, sue or shoe). Although the task was single-interval four-alternative forced choice, responses were coded only for initial consonant (i.e. see and sue both were coded as /s/ perceptions, while she and shoe were both coded as /ʃ/ perceptions). Stimuli were presented using the Alvin software package (Hillenbrand & Gayvert, 2005) in alternating blocks of spectral resolution (unprocessed or 8-channel NBV), and presentation of tokens within each block was randomized. CI listeners only heard the unprocessed stimuli. In this self-paced task, the 144 stimuli were each heard 5 times in each condition of spectral resolution.

Analysis.

Listeners' responses (coded as 1 or 0 referring to s- or ʃ-onset words) were fit using a generalized linear (logistic) mixed-effects model (GLMM). This was done in the R software interface (R Development Core Team, 2010), using the lme4 package (Bates and Maechler, 2010). A random effect of participant was used, and the fixed-effects were the stimulus factors described above. This was the same kind of analysis used for Experiment 4, with different fixed-effects. As for Experiment 4, two additional analyses were conducted to calculate basic cue-driven difference scores and frequency boundary shifts.

Results

Identification functions in the various vocalic contexts are shown in Figures 4.10, 4.11 and 4.12. Accommodation to context is indicated by the separation of psychometric functions of the same line type. Consistent with Experiment 4, NH listeners were affected by talker gender. CI listeners were affected slightly less, and NH listeners in CI simulated were affected the least. The functions corresponding to lip-rounding posture suggest that this effect is small for NH listeners, and larger for both CI listeners and NH listeners in CI simulations. Difference scores quantifying the use of vocalic cues are illustrated in Figure 4.13 and the category boundary effects are illustrated in Figure 4.14. The generalized linear mixed-effects models were described optimally as follows:

1) *Perception by NH listeners in different conditions:*
/s/ perception ~ SPP + lip + audio + vowel + video + SR + SPP:SR + lip:SR + audio:SR + vowel:SR + video:SR + SPP:audio:SR + lip:video:SR + SPP:vowel + SPP:lip + SPP:lip:SR + SPP:vowel:SR + (1|Participant)

2) *Perception by CI listeners:*
/s/ perception ~ SPP + lip + audio + video + SPP:video + lip:video + SPP:audio + SPP:lip + vowel + SPP:vowel + SPP:lip:vowel + (1 | Participant)

For these two models, the interaction between factors A and B is indicated by A:B. Independent factors are indicated by "+." SR refers to spectral resolution (either unprocessed/natural speech or signals degraded with the noise-band vocoder). Audio is the auditory gender cue and video is the visual gender cue. Lip is the configuration of lip-rounding, referring to whether the visual component originated from a ʃ-onset (rounded) word or an /s/-onset (retracted) word. SPP means spectral peak prominence (defining the continuum of fricative spectra) and (1|Participant) is a random effect of participant.

For the normal hearing listeners, there were significant main effects of fricative spectrum, lip rounding and auditory gender cues (all $p < 0.001$). The effect of visual gender cues did not reach significance for NH listeners in the unprocessed condition. Importantly, there were interactions between the context factors and spectral degradation. The effects of visual gender cues (Fig. 4.10) and lip-rounding (Fig. 4.11) were significantly stronger in the degraded condition ($p < 0.01$ and $p < 0.001$, respectively). The effects of auditory gender cues (Fig. 4.9) and vowel context (Fig. 4.12) were significantly weaker in the degraded condition (both $p < 0.001$). The effect of lip-rounding was stronger for the female speaker overall ($p < 0.01$), and this effect was significantly stronger when in the degraded condition ($p < 0.001$). The use of the fricative spectrum cue was weaker in the degraded condition. In the degraded condition, the slope of the identification function was shallower for the /i/ context than for the /u/ context, presumably because the lip configuration for this vowel combined with the retracted lip posture for the consonant is incompatible with /ʃ/, preventing true floor effects for that subset of stimuli; the rounding of the /u/ vowel is compatible with either phoneme, so floor and ceiling were reached.

Figure 4.10a Gender-related context effects for NH listeners in the unprocessed condition in Experiment 5

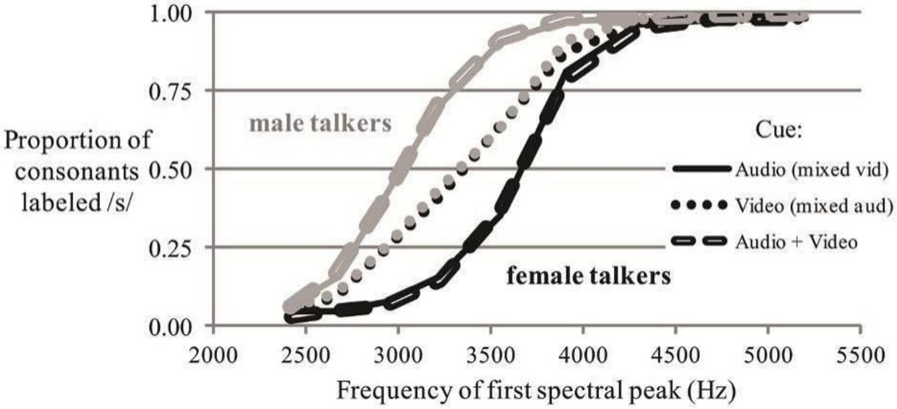

Figure 4.10a. Perception of the fricative continuum in the context of various auditory, visual and audio-visual cues to gender in Experiment 5 for normal-hearing listeners in the unprocessed condition. Black lines indicate contexts predicted to elicit fewer /s/ perceptions (female talker cues), and gray lines indicate contexts predicted to elicit more /s/ perception (male talker cues).

Figure 4.10b Gender-related context effects for NH listeners in the spectrally-degraded condition in Experiment 5

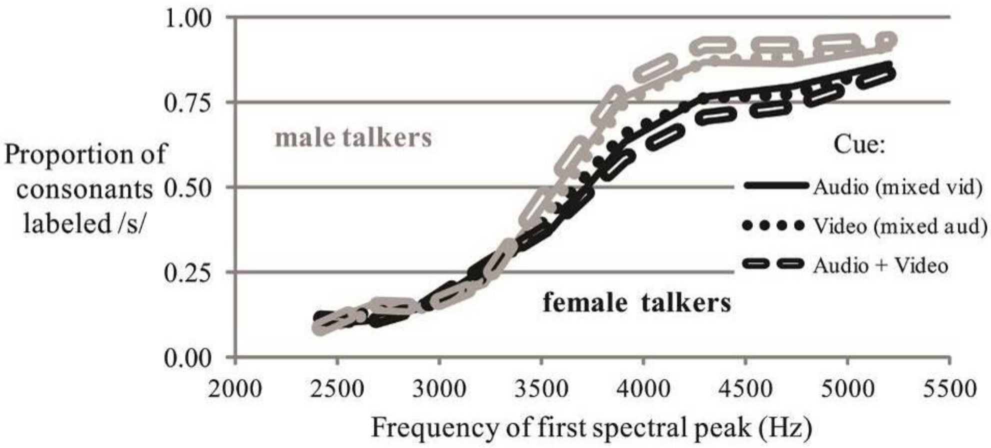

Figure 4.10b. Perception of the fricative continuum in the context of various auditory, visual and audio-visual cues to gender in Experiment 5 for normal-hearing listeners in the spectrally-degraded condition. Black lines indicate contexts predicted to elicit fewer /s/ perceptions (female talker cues), and gray lines indicate contexts predicted to elicit more /s/ perception (male talker cues).

Figure 4.10c Gender-related context effects for CI listeners in Experiment 5

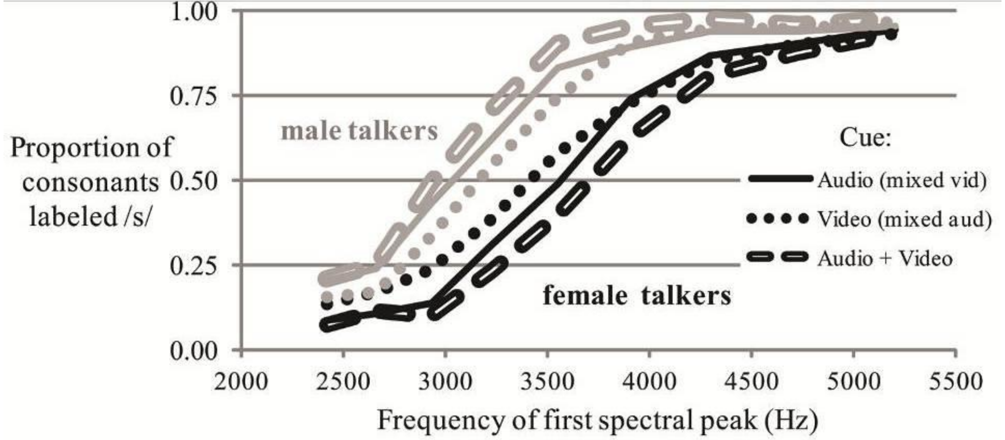

Figure 4.10c. Perception of the fricative continuum in the context of various auditory, visual and audio-visual cues to gender in Experiment 5 for cochlear implant listeners. Black lines indicate contexts predicted to elicit fewer /s/ perceptions (female talker cues), and gray lines indicate contexts predicted to elicit more /s/ perception (male talker cues).

Figure 4.11a. Effects of visual cues to lip-rounding by NH listeners in the unprocessed condition in Experiment 5

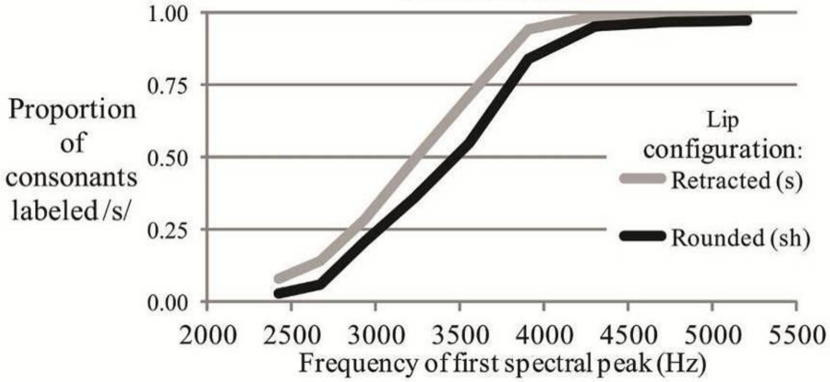

Figure 4.11a. Perception of the fricative continuum in the context of visual lip-rounding cues for normal-hearing listeners in the unprocessed condition in Experiment 5.

Figure 4.11b. Effects of visual cues to lip-rounding by NH listeners in the spectrally-degraded condition in Experiment 5

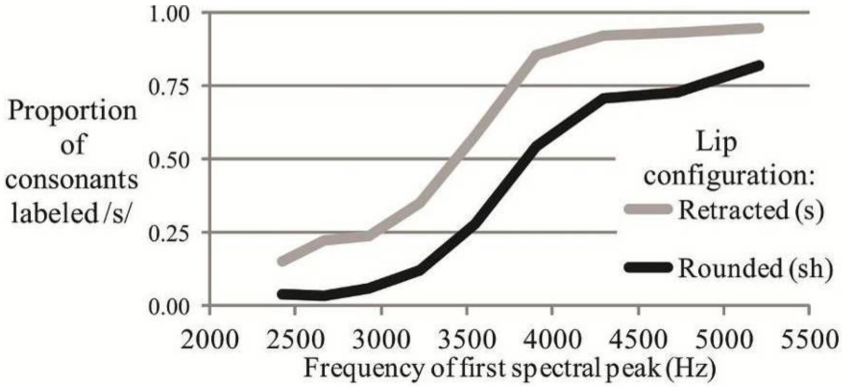

Figure 4.11b. Perception of the fricative continuum in the context of visual lip-rounding cues for normal-hearing listeners in the spectrally-degraded condition in Experiment 5.

Figure 4.11c. Effects of visual cues to lip-rounding by CI listeners in Experiment 5

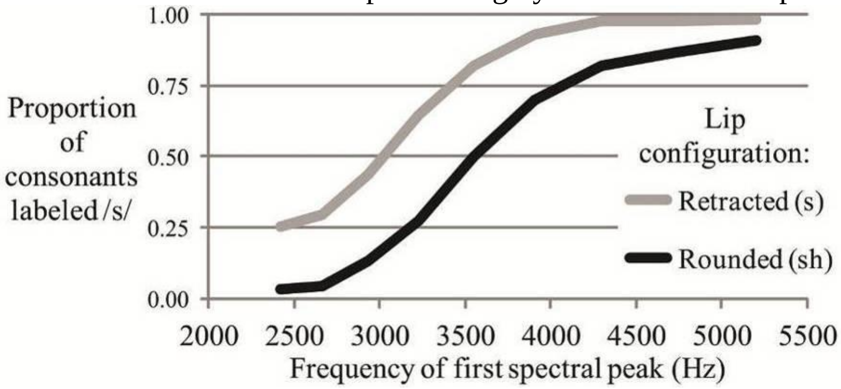

Figure 4.11c. Perception of the fricative continuum in the context of visual lip-rounding cues for cochlear implant listeners in Experiment 5.

Figure 4.12a Vowel context effects for NH listeners in the unprocessed condition in Experiment 5

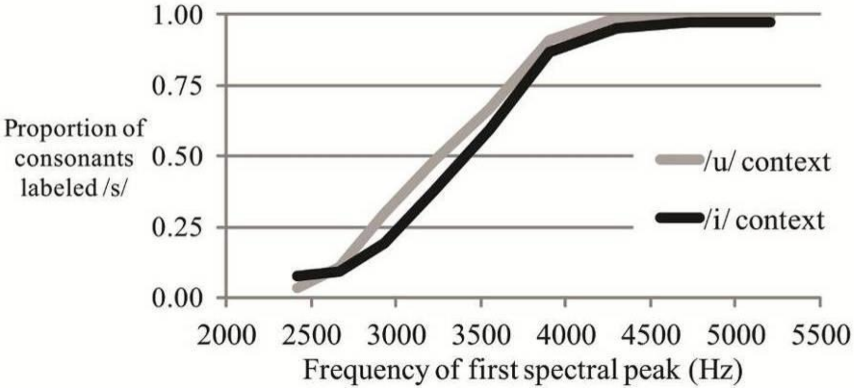

Figure 4.12a. Perception of the fricative continuum in the context of audio-visual cues to vowel context for normal-hearing listeners in the unprocessed condition in Experiment 5.

Figure 4.12b Vowel context effects for NH listeners in the spectrally-degraded condition in Experiment 5

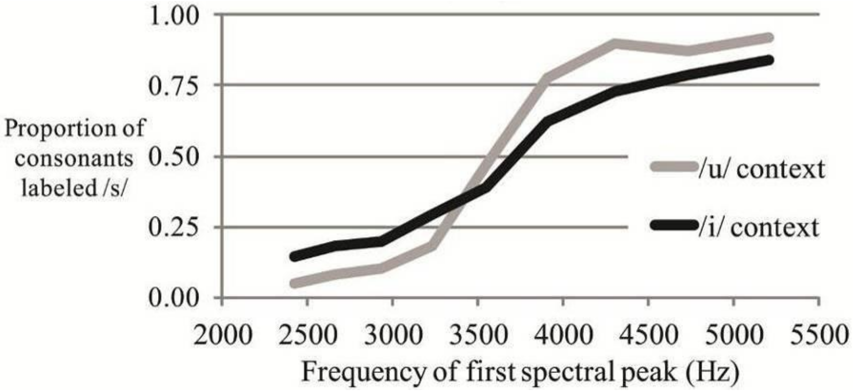

Figure 4.12b. Perception of the fricative continuum in the context of audio-visual cues to vowel context for normal-hearing listeners in the spectrally-degraded condition in Experiment 5.

Figure 4.12c Vowel context effects for CI listeners in the unprocessed condition in Experiment 5

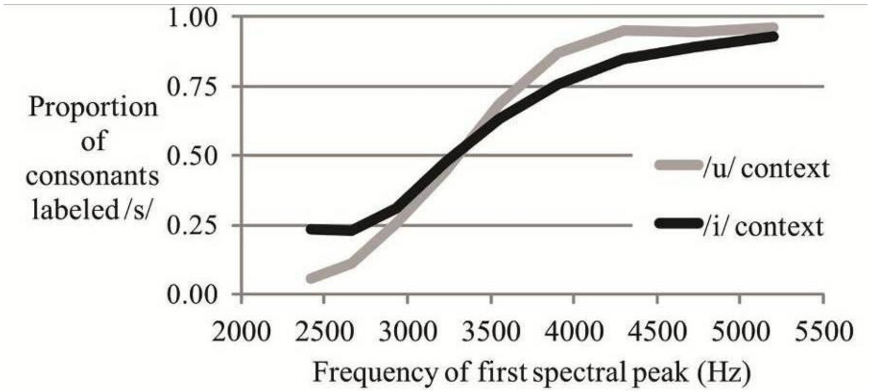

Figure 4.12c. Perception of the fricative continuum in the context of audio-visual cues to vowel context for cochlear implant listeners in Experiment 5.

Figure 4.13 Context effects in Experiment 5: Difference scores

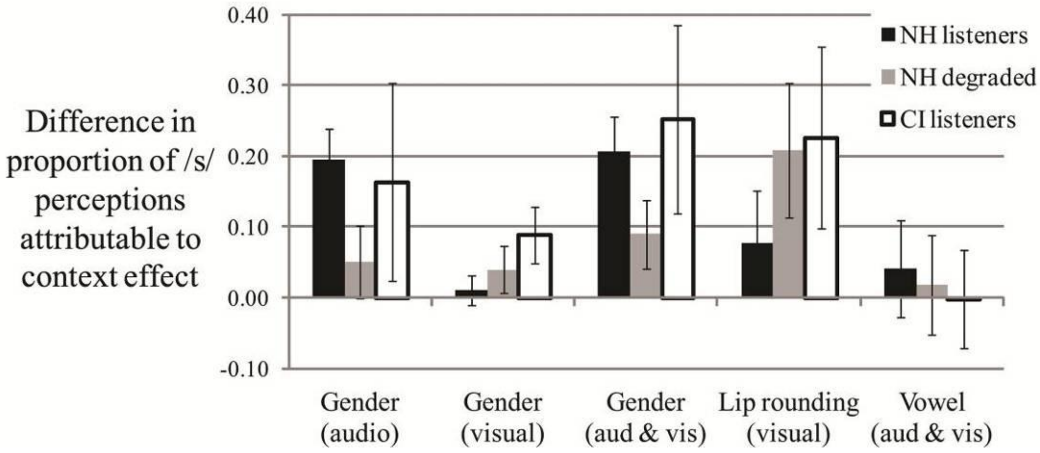

Figure 4.13. Quantification of various auditory, visual, and audio-visual phonetic context effects in Experiment 5. Bars represent the average proportion of /s/ perceptions for the /s/-bias levels of the cues (e.g. male voice, male face, retracted lips, /u/ vowel context) minus the proportion of /s/ perceptions in the /ʃ/-bias levels of the same cues. Errors bars indicate standard deviations.

Figure 4.14 Context effects in Experiment 5: Boundary shifts

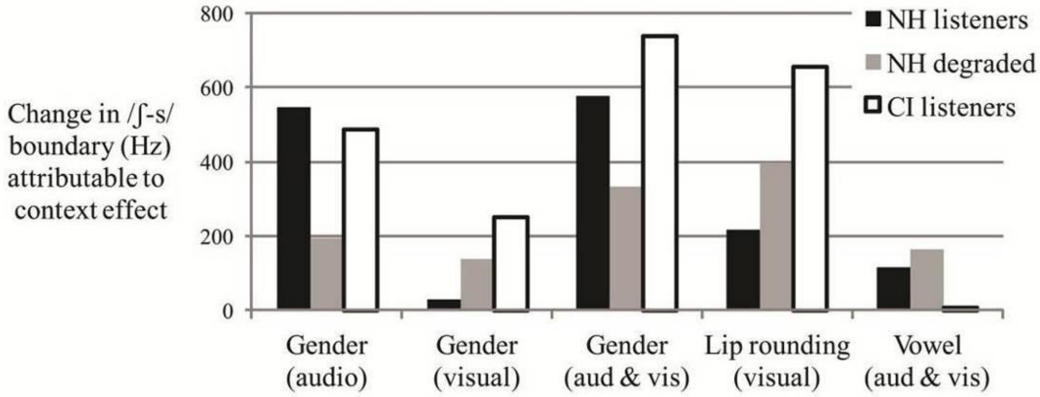

Figure 4.14. Effects of various auditory, visual, and audio-visual phonetic context effects on the /s/-/ʃ/ boundaries predicted from the group logistical models in Experiment 5. Bars represent the separation of boundaries in the /s/-bias levels of the cues (e.g. male voice, male face, retracted lips, /u/ vowel context) from the boundaries in the /ʃ/-bias levels of the same cues. Positive numbers indicate change in the expected direction.

Conclusions

In this experiment, the measurement of contextual effects in phonetic identification was measured in a task that incorporated visual cues of various types. It was found that the acoustic context effects that were already known to exist for NH listeners (and replicated in experiment 4) persisted in the audio-visual task. Additionally, visual cues played a role in phonetic perception, particularly when the acoustic signal was degraded via electric or simulated electric hearing. Specifically, lip-rounding (or lip-retraction) was a strong cue for the /ʃ/ sound (or the /s/ sound) for CI listeners and NH listeners in the degraded condition. Furthermore, the presence of a male face or female face in the visual domain brought about shifts in phonetic judgments in a manner consistent with the acoustic properties of male and female voices. The effect of vowel context was not as strong in the audio-visual task, perhaps because the lip-rounding cue was spread across both consonant and vowel segments in a way that produced cue interactions consistent with physical coarticulation.

Consistent with the main hypothesis in this dissertation, listeners compensated for degraded spectral cues by making more use of cues in the non-spectral domain. In this case, those cues were from a different sensory modality altogether.

Summary and Discussion of Experiments 4 and 5

The presence of acoustic context effects in these experiments supported findings of earlier literature and generalized the phenomenon to cochlear implant listeners. The context effects did not emerge for NH listeners that heard spectrally degraded signals. It is likely that experience with spectrally degraded speech plays a role in the differences between the CI and NH (degraded) groups.

Consistent with the results of Experiments 1, 2 and 3, the experiments in this chapter suggest that listeners can identify phonemes with similar accuracy but still exhibit different perceptual behaviors. In this case, CI listeners showed excellent accuracy in /s/ and /ʃ/ identification (Figure 4.8) while showing reduced sensitivity to some acoustic components that help define this segment, such as adjacent formant transitions (Figures 4.5, 4.6, 4.7). Experiments in this chapter benefited from multiple kinds of analysis, and this was most evident for the effect of vowel context in Experiment 5. Calculation of the s/-/ʃ/ boundary and basic difference scores suggested virtually no accommodation to vowel context by CI listeners (Figure 4.13, 4.14), but the psychometric functions were morphologically dissimilar (4.12c). This difference was captured in the GLMM, since the curve slope is a variable in the model estimating equation.

The visually mediated effects in this study are of two kinds. The first is a type of auditory-visual fusion consistent with the classic McGurk effect (McGurk & MacDonald, 1976). Lip-rounding indicates the segment /ʃ/, while lip-spreading indicates /s/. Consequently, the auditory perception of the fricative is affected in a way that is consistent with the concurrent visual perception. The effect of the visual gender cue in this experiment is of a different kind than that of the visual lip-rounding cue because while lip-rounding has direct phonetic correspondence, gender does not. That is, there is nothing inherent in a female or male face that should make the /ʃ/ or /s/ sounds more probable. The influence of this cue cannot be completely understood just on the basis of this experiment, but it is likely that it arises either from a general effect of cue covariance or by a learned association between phonetic segments and typical gender-related differences in fricative and/or vowel acoustics. Although Holt et al. (2005) suggest that mere variation in concurrent visual information is sufficient to induce context effects, it is noteworthy that the effect in this experiment went in the direction predicted by the acoustics that correspond to the gender of the visual stimulus (recall that the visual and auditory gender cues in this experiment were fully-crossed). In other words, listeners had no reason to prefer /s/ for male faces (even when presented with concurrent female voices) other than having been exposed to the natural association between visual cues and the auditory spectral properties of voices emanating from those faces.

The influence of visual cues on the perception of fricatives by CI listeners implicates the potentially enhanced role of higher-level cognitive processes (specifically multi-modal cue integration and potential influence of long-term learned covariance) in speech perception by this population. The auditory information delivered to these listeners is heavily compromised, but they appear to capitalize on the associations between segments and 1) direct phonetic cues like lip-rounding and 2) indirect phonetic biases like the male/female influence on the /s/-/ʃ/ boundary shift. The additional visual cues for gender did not benefit NH listeners, presumably because the auditory cues are sufficient to enable the context-dependent identification shift. Similarly, the visual lip-rounding cue did not exert heavy influence on the NH listeners presumably because the auditory cues are clear and reliable. For NH listeners in the degraded conditions and also for CI listeners, the visual lip-rounding cue was a more potent cue, likely because the auditory information was less reliable. These results suggest that listeners are able to exploit numerous dimensions of stimulus properties using multiple modalities in order to best accommodate the variability inherent in natural speech. Furthermore, the results show promise for the potential coding of speech information driven by spectral contrast.

Because frequency resolution (for cues like formant transitions) is heavily impaired for CI listeners, they predictably make frequent errors on consonant place contrasts driven by such cues. Experiments 4 and 5 show that CI users can benefit from spectral contrast (as well as other cues); perhaps the deliberate enhancement of spectral contrast in CI processors can facilitate increased success in consonant place recognition, because it is a cue already present in the signal. Implications for this will be discussed in more detail in Chapter 5. Overall, the results from Experiments 4 and 5 are consistent with those from Experiments 1, 2 and 3: when phonetic cues are compromised by simulated or real hearing impairment, listeners can capitalize on the presence of other cues that naturally co-vary in the signal in order to perceive phonetic segments in a way that could help to facilitate correct perception of words.

Chapter 5: General Summary and Discussion

This dissertation has described experiments that help to illuminate perceptual strategies that underlie phonetic perception by listeners with normal hearing, listeners with cochlear implants, and listeners with simulated hearing impairment/cochlear implants. These experiments were motivated by the existence of multiple acoustic cues for phonetic contrasts, and the observation that some of those cues are compromised if a listener experiences spectral degradation due to hearing impairment or cochlear implantation or if the speech is masked by noise. Participants in these experiments demonstrated that the influence of various phonetic cues (in either the auditory or visual domains) can be modulated by the quality of the acoustic input in terms of spectral resolution, bandwidth, or signal-to-noise ratio.

The concept of trading relations between spectral and temporal information is not a new one, but it has been largely neglected in previous literature on listeners with hearing impairment. The argument is made here that there is a fine distinction between perceiving phonetic features and perceiving acoustic cues for those features. Perception of “lax / tense,” “voicing” or other features by an impaired listener does not imply that it was because of the same perceptual cue used by normal-hearing listeners. In view of the multiple acoustic cues available for any particular phonetic segment, the contrasts explored in this study may represent just a fraction of those for which CI listeners could employ alternative perceptual strategies. Thus, caution should be used when comparing results of NH listeners and CI listeners in word recognition tasks; similar performance (Figures 2.6, 2.11, 4.8) may not verify similar perception or perceptual processes. Additionally, high transmission of a phonetic feature in a speech-in-noise task (Figure

3.9) does not verify typical use of acoustic phonetic cues. Furthermore, if CI listeners reliably distinguish /s/ and /ʃ/ sounds in words (Figure 4.8), it does not mean that they can accommodate variability in these segments and adjust perception accordingly, the way that NH listeners do. All of these qualifications imply the usefulness of speech perception analysis that extends beyond the level of the phonetic feature.

It could be argued that the difference in cue-weighting or cue usage might make no difference in the "bottom line" of word recognition. After all, if a listener correctly perceives a word, he/she might not care about the method by which it was done. However, it is not clear whether all perceptual cue weighting strategies are equally reliable, efficient or taxing for the listener. The data in this dissertation cannot speak to any potential differences in processing speed, efficiency or listening effort, but it should be noted that if normal-hearing listeners tend to rely on a particular cue for a contrast, there is probably a reason for that tendency (it may be explained by acoustic reliability; see Holt & Lotto, 2006; Toscano & McMurray, 2010). Future work might address this issue by exploring neurological responses to multidimensional speech stimuli (see Pakarinen, Takegata, Rinne, Huotilainen, & Näätänen, 2007; Pakarinen, Lovio, Huotilainen, Alku, Näätänen, & Kujala, 2009), or by more sophisticated measurements that show sensitivity to listening effort, such as pupil dilation (Koelewijn, Zekveld, Festen & Kramer, 2011).

More generally, this work accords with previous literature that indicates greater use of vowel duration by listeners with hearing impairment (Revoile, 1982), and adds a new layer to work comparing the use of cues in natural and synthesized signals (Assmann & Katz, 2005; Nittrouer, 2004; Nittrouer, 2005). The variability in the CI listener data is problematic for drawing general conclusions, but it might potentially be a fruitful avenue of exploration. A small number of CI listeners in this study appeared to rely heavily on the same cues used by NH listeners, while the others were relatively more influenced by other cues. While auditory prostheses and amplification devices are designed generally to transmit the acoustic cues used by normal-hearing listeners, not all listeners use the cues in the same way. Thus, for listeners that cannot or do not make use of some spectral cues, computational effort (i.e. battery power, processing speed, etc) might be wasted in the delivery of some extra spectral information in the implant. It is not known whether successful CI listeners are those that are able to extract and decode spectral cues despite device limitations, or if they are diverting attention/resources away from those ("dead-end") cues in favor of those that remain intact in the temporal domain.

The issue of age differences between the NH and CI groups in Experiments 1, 2, 4 and 5 introduces some complications in the analysis of the current data. It has been shown numerous times that older listeners show deficiencies in auditory temporal processing in basic psychophysical tasks (Gordon-Salant & Fitzgibbons, 1993; 1999), and tasks involving perception of temporal phonetic cues (Gordon-Salant et al., 2006). They therefore might be less able to capitalize on the duration cue available in this study and in natural speech. Furthermore, older listeners have been shown to experience more difficulty with spectrally-degraded speech in general (Schvartz, Chatterjee & Gordon-Salant, 2008). If one presumes that psychophysical capabilities/deficiencies influence behavior in this identification task, the trend of the CI listeners in this study is opposite to that which might be predicted by their age; they showed increased use of durational cues compared to the young NH listeners. However, it is evident that capability is not entirely predictive of cue usage; the CI listeners in this study did not use the fricative voicing cue even though this population has been shown to exhibit very fine sensitivity to temporal modulations. Perhaps younger CI listeners, with hypothetical advantages in temporal processing, would show more reliable use of the vowel duration and/or voicing cues than the older listeners in this study. Young post-lingually-deafened CI listeners are generally more scarce in the population though, and were not available at the time of this experiment; the question of the role of aging in the use of phonetic cues and/or temporal processing in electric hearing invites future work. Unfortunately, even if a group of implanted and non-implanted listeners were age-matched, other sources of variability would remain. Across-group variability could include cognitive functioning, attention, temporal resolution, neural degeneration, or other language-related abilities. Within-group variability for a group of CI listeners could also include differences in the duration of deafness, duration of CI use, motivation, lifestyle, neural survival, and speech processor settings. For these reasons, it is extremely difficult to obtain rigorously matched groups of NH and CI listeners. Even in the face of this limitation, results from the current study and previous literature still provide some value in understanding the experience of electric hearing.

When simulating electric hearing, it has become commonplace in the field to use noise-band vocoders (NBV) or sine-wave vocoders (SWV) to process speech that is played for NH listeners. Results presented in this dissertation expose some limitations of the NBV in modeling perception of speech by CI listeners. While the deficit in spectral cue processing was modeled fairly well in Experiments 1 and 2 (NH listeners in the NBV condition showed decreased use of formant cues; see Figures 2.3, 2.4, 2.7, Tables 2.4, 2.7a), this deficit was heavily over-estimated in Experiments 4 and 5, where CI listeners showed markedly greater accommodation to phonetic contexts that were presumably driven by spectral contrasts (compare Figures 4.6, 4.7, 4.13 and 4.14). Although experience with the implant is a likely contributor to this difference, it is notable that NH listeners in the NBV condition showed virtually zero spectral accommodation. Clearly, the NBV fails to deliver some information that is used by listeners with CIs.

Some caution should be used when interpreting the current results in the context of hearing impairment in general. For example, even though the F0 cue facilitated correct voicing perception in Experiment 3, this cue might be less useful for a listener that experiences cochlear hearing impairment. Turner and Brus (2001) showed that while the amplification of low-frequency energy (which includes F0) provided benefit for listeners with hearing impairment, this benefit was smaller than that observed for those with normal hearing. This is in agreement with Grant (1987), who downplays the potential benefit of F0 contour for impaired listeners because they are less sensitive to small changes in F0. Furthermore, because stimuli in Experiment 3 were presented in isolation, it is not known whether the use of F0 cues would generalize to longer utterance contexts, where F0 is likely compromised by other sources of variability and phonetic reduction. Previous work has shown that temporal processing deficits for phonetic cues by older listeners is exacerbated in sentence contexts (Gordon-Salant, Yeni-Komshian & Fitzgibbons, 2008). Thus, if the F0 cues are reduced in sentence context, it is likely that older listeners would be at a great disadvantage, since they are already less likely to utilize the temporal cues efficiently.

Despite some limitations in extension of this work directly to listeners with hearing impairment, results of Experiment 4 and 5 have implications for the future of speech processing strategies in CIs. The small number of CI listeners in these experiments might show promise for speech feature coding strategies that capitalize on time-varying spectral contrast. Stilp and Kluender (2010) suggest that short-term variations in cochlear-scaled entropy account for much of the intelligibility of speech. The results presented here suggest that CI listeners can take advantage of such dynamic spectral changes, albeit perhaps to a lesser extent than those with normal hearing. If implant speech processors further exploit and emphasize these variations, implant recipients may enjoy greater success in perception. This strategy, of course, would be limited by the compressed electric dynamic range of the implant.

Dynamic amplitude cues for gross spectral shape can potentially facilitate perception of the consonant place contrast, which remains among the most urgent unresolved problems for CI listeners. Phoneme groups such as /p-t-k/ and /b-d-g/ are consistently among the most difficult to distinguish among this population. Among the acoustic cues that are used to discriminate these sounds is relative spectral change over time (Alexander & Kluender, 2008). Specifically, the change in spectral tilt of the stop release (and subsequent 30 ms) and the following vowel segment can influence listeners' perception of these place contrasts. Despite being described acoustically as early as 1957 (Halle, et al.) and emphasized by Blumstein and Stevens (1979, 1980), this acoustic cue has not received adequate attention in the CI literature. It is likely that the dominance of formant transitions over relative spectral tilt in NH perception is the cause for the neglect of the tilt cue. Gross spectral shape (spectral tilt) was described by Blumstein, Isaacs and

Mertus (1982) as a potentially invariant cue for stop place contrasts. Kewley-Port (1983) and Lahiri, Gewirth and Blumstein (1984) showed that while dynamic spectral tilt changes can influence place perception, these spectral changes are subordinate to formant transitions for NH listeners (Dorman & Loizou, 1996). Importantly, CI listeners in the current experiment did not take advantage of formant transitions in their labeling of /s/-/ʃ/ place contrasts in Experiment 5 (see Figures 4.6 and 4.7; CI listeners also did not show substantial use of dynamic formant information in Experiment 1 or 2 – see Table 2.4). Perhaps efforts to deliver these quick time-varying formants could be supplemented by efforts to deliver the gross spectral shape (tilt), because that is likely what the CI listener will ultimately recover. The tradeoff between formant transitions and other cues by hearing-impaired listeners is highlighted by recent work by Alexander and Kluender (2008; 2009), who showed that hearing-impaired listeners were more heavily influenced by spectral tilt than normal-hearing listeners (whose judgments were based primarily on formant transitions). Because hearing impairment is associated with poorer spectral resolution, the likely explanation is that hearing-impaired listeners made greater use of the cue that required relatively less-fine spectral detail (i.e. spectral tilt is a coarse spectral property accessible even with only limited resolution). Because CI listeners experience spectral resolution that is even poorer than the typical listener with hearing loss (Henry et al., 2005), the importance of dynamic gross spectral shapes is potentially greater for listeners with CIs. Future research may uncover additional examples of situations in which hearing-impaired listeners adopt slightly altered acoustic-phonetic perceptual strategies relative to normal-hearing listeners.

The translation of the current findings into tangible clinical utility is a difficult process. Clinicians are frequently under tremendous time constraints, as well as constraints of policies (legally, professionally or otherwise) that dictate the types of testing to be performed to warrant official diagnosis. It is unlikely that the experiments presented here in their current form would provide enough benefit to outweigh the cost (in terms of time). On the other hand, if such testing could be modified to be more efficient and easy to administer, it could potentially be a useful tool for the programming of hearing aids and/or cochlear implants.

A more direct clinical utility of this work would be to evaluate the effectiveness of CI processing strategies that are designed specifically to improve spectral resolution. For example, the Advanced Bionics corporation produces a processing scheme called HiRes Fidelity 120 (HR120) that is purported to provide 120 “virtual” spectral channels (resulting from variable current steering between electrodes). While the spread of electrical excitation along the basilar membrane likely decimates the effective number of perceptual channels, the issue of spectral resolution is still arguably the most critical room for improvement in modern CIs and thus deserving of attention. Although evaluation of HR120 verifies some improvement in speech recognition and appreciation of music (Firszt, Holden, Reeder and Skinner, 2009), traditional speech perception tasks are limited in their efficacy to evaluate improvements in spectral resolution because listeners already perform very well with devices (or simulated processing strategies) that are known to have poor resolution. Unfortunately, speech tasks that focus primarily on perception of spectral cues do not always show benefit from the increased number of virtual channels in HR120. For example, Han, Liu, Zhou, Chen, Kong, Liu et al. (2009)

tested perception of Mandarin tones, which are identified primarily according to F0 contour (although other cues exist in the temporal domain). They observed no improvement in listeners that used the HR120 processing scheme. Even if there were significant improvements in that task, it would be difficult to generalize that benefit to English, where F0 does not play the same crucial role in the lexicon as in Mandarin. Frequently, experimenters will avoid speech altogether, and instead assess spectral resolution using pitch perception (Firszt, Kosh, Downing & Litvak, 2007) or pitch ranking (Vandali, Sucher, Tsang, McCay, Chew & McDermott, 2005; Luo, Landsberger, Padilla and Srinivasan, 2010). Although pitch ranking is a clean method that is likely free from some of the temporal-cue confounds of basic word recognition, it is a substantially different task than speech perception. The basic approach to evaluating the use of spectral cues in this dissertation could be an avenue through which spectral resolution is evaluated for CI listeners in a task that is directly related to speech perception. For example, experimenters can assess the perception of specific spectral cues in speech contrasts such as those in Experiments 1, 2 and 4 (the tense/lax contrast, final fricative voicing contrast, and the accommodation of talker gender in fricative identification). If it is true that a processing scheme like HR 120 offers better spectral resolution than traditional coding strategies, then listeners should show higher use of spectral cues (i.e. higher cue estimates) in these listening tasks with HR120 than with other coding strategies. A complication in this matter is that it may take time for a CI listener to show increased use of spectral cues. In the current study, NH listeners showed some altered use of cues in an *acute* testing situation where there was no training or explicit instructions. If there were evidence that this type of perceptual strategy-shifting leads to better success at general speech recognition, further research could address whether this aspect of speech perception is amenable to auditory training.

Other areas of future research that stem from the current findings include the application of these experimental methods to other populations of interest. These include learners of a second language, who might show success while attending to acoustic cues in a non-native fashion. An example of this can be found in the case of native Spanish speakers that learn English as a second language. Morrison (2007) showed that the non-native language learners attended more to temporal cues for the /I/-/i/ contrast, while native listeners attended more to the spectral cues. This could potentially be used as a performance metric for language acquisition and mastery. Additionally, subtle differences in perception that might be overlooked by more basic methods might be captured using logistical modeling, since this method is apparently sensitive to understated differences in perception.

Another possible extension of this work could be to track the use of acoustic cues by developing language learners. It might be fruitful to see if auditory processes underlie common patterns of speech difficulties in a way that is not captured by traditional testing. This would essentially be the perceptual counterpart to the concept of “covert contrasts” (Munson, Edwards, Schellinger, Beckman and Meyer, 2010), where a talker appears to neutralize a phonetic contrast that actually exists in the production. The erroneous judgment in this matter could arise because acoustic cues are not produced by the talker with enough specificity, or perhaps because the talker exploits the wrong acoustic cues (i.e. not the cues that adult native speakers use to perceive the contrast). In a similar sense, the approach used in this dissertation could be used for longitudinal assessment of

CI listeners who might show progressive shifts in their use of acoustic cues in speech perception.

In Chapter 5, CI listeners showed sensitivity to spectral variability in the speech signal. The apparently elevated use of temporal cues by CI listeners in chapter 2 warrants some predictions about the sensitivity to sources of temporal variability in the speech signal. Miller and Volaitis (1989) have shown that listeners adjust for different speaking rates by adjusting category boundaries between temporally-contrastive segments, and also adjust perception of phonetic category structure distal to the boundary. Accordingly, listeners also adjust the location of a best exemplar of a phonetic category in response to speaking rate as well (Wayland, Miller & Voltais, 1992). Because CI listeners present with temporal resolution that is as good as or better than that of NH listeners, it is likely that they would yield comparable results in tests of phonetic perception using variable-rate stimuli. Furthermore, it may be the case that increased reliance upon temporal cues would bring about even more robust temporal accommodation to speaking rate because the spectral information would be less influential. Although this is mere speculation, it is an interesting extension of the current results to a situation that is more life-like in terms of input variability.

For decades, researchers have struggled to identify invariant acoustic cues for speech sounds. The high amount of within-talker and across-talker variability has frustrated this search, leaving some to conclude that there are no invariant cues. This remains an active and lively debate among some researchers who advocate for the existence of acoustic landmarks (Stevens, 2002; Li & Allen, 2011) and others who posit that object perception is the physical gesture rather than the acoustics (Fowler, 1995, 2006). The presence of variable perceptual strategies in phonetic perception in this dissertation further undermines the concept of invariant acoustic properties of phonemes. Instead, results presented here support the view that the listener incorporates multiple acoustic cues in perception and furthermore, that the influence of these cues can be modulated by the signal quality with which the speech is represented. Although this statement is in stark opposition to the view expressed by Li and Allen (2011) that non-primary acoustic cues (i.e. those not deemed essential according to the 3DDS method of cue isolation) cause confusion, it could merely be the case that invariant cues do exist and that listeners simply don't rely on them exclusively.

Results presented in this dissertation suggest that the multi-dimensional composition of the speech signal in the spectral, temporal, spectro-temporal and visual domains can potentially be a source of benefit for listeners with hearing impairment, listeners with cochlear implants, or listeners in degraded situations such as noise. The current experiments reinforce long-standing characterizations of speech perception as an extremely robust ability (Assmann & Summerfield, 2004) and strongly support the notion that alternative acoustic cues are available and are used effectively by those with compromised (or simulated compromised) auditory systems.

References

Abramson, A., & Lisker, L. (1970). Discriminability along the voicing continuum: Cross language tests," *Proceedings of the 6th International Congress of Phonetic Sciences, Prague,* 569-573.

Abramson, A., & Lisker, L. (1985). Relative power of cues: F0 shift versus voice timing, In Fromkin, V. (Ed.) *Phonetic Linguistics.* New York: Academic.

Ainsworth, W. (1972). Duration as a cue in the recognition of synthetic vowels. *Journal of the Acoustical Society of America*. *51,* 648–651.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control 19,* 716–723.

Alexander, J. & Kluender, K. K. (2008). Spectral tilt change in stop consonant perception. *Journal of the Acoustical Society of America, 123,* 386–396.

Alexander, J., & Kluender, K. K. (2009). Spectral tilt change in stop consonant perception by listeners with hearing impairment. *Journal of Speech, Language and Hearing Research*, 52, 653–670.

Allen, J., & Miller, J. L. (1999). Effects of syllable-initial voicing and speaking rate on the temporal characteristics of monosyllabic words. *Journal of the Acoustical Society of America,* 106, 2031–2039.

American National Standards Institute (2007). American National Standard Methods for Calculation of the Speech Intelligibility Index [ANSI S3.5-1997 (R 2007)]. New York: American National Standards Institute.

American National Standards Institute [ANSI] (2010). American National Standard Specification for Audiometers. (ANSI S3.6-2010) (New York: American National Standards Institute.

Assmann, P,. & Katz, W. (2005). Synthesis fidelity and time-varying spectral change in vowels. *Journal of the Acoustical Society of America, 117*, 886-895.

Assmann, P., & Summerfield, A. (2004). The perception of speech under adverse conditions. In Greenberg, S., Ainsworth, W., Popper, A. & Fay, R. (Eds.) *Speech Processing in the Auditory System* (231–308). New York Springer.

Baer, T., Moore, B.C.J., & Kluk, K. (2002). Effects of low pass filtering on the intelligibility of speech in noise for people with and without dead regions at high frequencies. *Journal of the Acoustical Society of America, 112*, 1133–1144.

Başkent, D., & Shannon, R. (2003). Speech recognition under conditions of frequency-place compression and expansion. *Journal of the Acoustical Society of America, 113*, 2064–2076.

Bates, D., & Maechler, M. (2010). lme4: Linear mixed-effects models using S4 classes. R package version 0.999375-37 [Software package]. Available from http://CRAN.R-project.org/package=lme4

Bernstein, J., & Oxenham, A. (2006). The relationship between frequency selectivity and pitch discrimination: Sensorineural hearing loss. *Journal of the Acoustical Society of America, 120,* 3929–3945.

Bilger, R., & Wang, M. (1976). Consonant confusions in patients with sensorineural hearing loss. *Journal of Speech and Hearing Research, 19*, 738–748.

Binns, C., & Culling, J. (2007). The role of fundamental frequency contours in the perception of speech against interfering speech. *Journal of the Acoustical Society of America, 122,* 1765-1776.

Bladon, R., Henton, C. & Pickering, J. (1984). Towards an auditory theory of speaker normalization. *Language & Communication*, 4(1), 59–69.

Blumstein, S., Isaacs, E., & Mertus, J. (1982). The role of the gross spectral shape as a perceptual cue to place of articulation in initial stop consonants. *Journal of the Acoustical Society of America, 72,* 43–50.

Blumstein, S., & Stevens, K. (1979). Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants. *Journal of the Acoustical Society of America, 66,* 1001–1017.

Blumstein, S., & Stevens, K. (1980). Perceptual invariance and onset spectra for stop consonants in different vowel environments. *Journal of the Acoustical Society of America* , 67, 648–662.

Boersma, P., & Weenink, D. (2011). Praat: doing phonetics by computer [Software]. Version 5.1.23, Available from http://www.praat.org/

Bohn, O.-S., (1995). Cross-language speech perception in adults; First language transfer doesn't tell it all. In Strange, W. (Ed.) *Speech Perception and Linguistic Experience; Issues in Cross-Language Research* (279-304). Baltimore, MD: New York Press.

Bohn, O.-S., & Flege, J. (1990). Interlingual identification and the role of foreign language experience in L2 vowel perception. *Applied Psycholinguistics, 11*, 303-328.

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

Carhart, R., Tillman, T. W. (1970). Interaction of competing speech signals with hearing losses. *Archives of Otolaryngology, 91,*273–279.

Chatterjee, M., & Shannon, R. (1998). Forward masked excitation patterns in multielectrode cochlear implants. *Journal of the Acoustical Society of America, 103,* 2565–2572.

Chen, M. (1970). ‘‘Vowel length variation as a function of the voicing of the consonant environment,’’ *Phonetica, 22,* 129–159.

Cooper, F., Delattre, P., Liberman, A., Borst, J. & Gerstman, L. (1952). Some experiments on the perception of speech sounds. *Journal of the Acoustical Society of America, 24,* 597-606.

Cutler, A., & Foss, D. (1977). On the role of sentence stress in sentence processing. *Language and Speech, 20,* 1–10.

Denes, P. (1955). Effect of duration on the perception of voicing. *Journal of the Acoustical Society of America, 27,* 761-764.

Derr, M. A., & Masaaro, D. (1980). "The contribution of vowel duration, F0 contour, and fricative duration as cues to the /juz/-/jus/distinction." *Perception & Psychophysics, 27,* 51-59.

Dorman, M., & Loizou, P (1996). Relative spectral change and formant transitions as cues to labial and alveolar place of articulation. *Journal of the Acoustical Society of America, 100,* 3825–3830.

Dorman, M., & Loizou, P. (1997). Mechanisms of vowel recognition for Ineraid patients fit with continuous interleaved sampling processors. *Journal of the Acoustical Society of America, 102,* 581–587.

Dorman, M. & Loizou, P. (1998). The identification of consonants and vowel by cochlear implant patients using a 6-channel continuous interleaved sampling processor and by normal-hearing subjects using simulations of processors with two to nine channels. *Ear & Hearing, 19,* 162-166.

Dorman, M., Dankowski, K., McCandless, G., Parkin, J. & Smith, L. (1991). Vowel and consonant recognition with the aid of a multichannel cochlear implant. *Quarterly Journal of Experimental Psychology, Sect. 43A,* 585-601.

Dubno, J., Dirks, D., and Langhofer, L. (1982). Evaluation of hearing-impaired listeners using a Nonsense-syllable Test. II. Syllable recognition and consonant confusion patterns. *Journal of Speech and Hearing Research, 25,* 141–148.

Dubno, J. R., & Levitt, H. (1981). Predicting consonant confusions from acoustic analysis. *Journal of the Acoustical Society of America, 69,* 249–261.

Elman, J., & McClelland, J. (1988). Cognitive penetration of the mechanisms of perception: Compensation for coarticulation of lexically restored phonemes. *Journal of Memory & Language*, *27*, 143-165.

Fang, Y. (2011). Asymptotic equivalence between cross-validations and Akaike Information Criteria in mixed-effects models, *Journal of Data Science, 9,* 15-21.

Festen, J., & Plomp, R. (1990). Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing. *Journal of the Acoustical Society of America, 88,* 1725–1736.

Firszt. J., Holden, L., Reeder, R., Skinner, M. (2009). Spectral Channels and Speech Recognition in Cochlear Implant Recipients using HiRes 120 Sound Processing. *Otoloy and Neurotology, 30,* 146-152.

Firszt, J., Koch, D., Downing, M., & Litvak, L. (2007) Current steering creates additional pitch percepts in adult cochlear implant recipients. *Otology and Neurotology, 28,* 629-636.

Fishman, K., Shannon, R. & Slattery, W. (1997). Speech recognition as a function of the number of electrodes used in the SPEAK cochlear implant speech processor. *Journal of Speech, Language and Hearing Research, 40,* 1201-1215.

Flege, J. & Hillenbrand, J. (1985). Differential use of temporal cues to the /s/-/z/ contrast by non-native speakers of English. *Journal of the Acoustical Society of America, 79,* 508-517.

Fowler, C. (1995). Speech production. In J. Miller, P. Eimas (Eds.), *Handbook of Perception and Cognition: Speech, Language, and Communication*. San Diego: Academic Press.

Fowler, C. (2006). Compensation for coarticulation reflects gesture perception, not spectral contrast. *Perception & Psychophysics,* 68, 161-177.

Fowler, C., Brown, J., & Mann, V. (2000). Contrast effects do not underlie effects of preceding liquids on stop-consonant identification by humans. *Journal of Experimental Psychology: Human Perception & Performance, 26,* 877-888.

Fowler, C., & Dekle, D. (1991). Listening with eye and hand: Crossmodal contributions to speech perception. *Journal of Experimental Psychology – Human Perception and Performance*, *17,* 816–828.

Francis, A., Baldwin, K., & Nusbaum, H. (2000). Effects of training on attention to acoustic cues. *Perception & Psychophysics*, *62*, 1668–1680.

Francis, A., Kaganovich, N., & Driscoll-Huber, C. (2008). Cue-specific effects of categorization training on the relative weighting of acoustic cues to consonant voicing in English. *Journal of the Acoustical Society of America, 124,* 1234-1251.

French, N., & Steinberg, J. (1947). 'Factors governing the intelligibility of speech sounds. *Journal of the Acoustical Society of America, 19,* 90–119.

Friesen, L., Shannon, R., Başkent, D., & Wang, X. (2001). Speech recognition in noise as a function of the number of spectral channels: Comparison of acoustic hearing and cochlear implants. *Journal of the Acoustical Society of America*, *110,* 1150-1163.

Fu, Q-J. (2006). Internet-Based Computer-Assisted Speech Training (iCAST) by TigerSpeech Technology (version 5.04.02). [Computer program]. Available from http://www.tigerspeech.com/tst_icast.html

Fu, Q-J. (2010). TigerCIS: Cochlear implant and hearing Loss Simulation (Version 1.05.03) [Computer program]. Available from http://www.tigerspeech.com/tst_tigercis.html

Fu, Q.-J., & Shannon, R. (1999). Recognition of spectrally degraded and frequency-shifted vowels in acoustic and electric hearing. *Journal of the Acoustical Society of America, 105,* 1889–1900.

González, J. (2004). Formant frequencies and body size of speaker: a weak relationship in adult humans. *Journal of Phonetics, 32*, 277–287.

Gordon-Salant, S. Fitzgibbons P. (1993). Temporal factors and speech recognition performance in young and elderly listeners. *Journal of Speech and Hearing Research, 36,* 1276-85.

Gordon-Salant, S., & Fitzgibbons, P. (1999). Profile of auditory temporal processing in older listeners. *Journal of Speech Language and Hearing Research, 42*, 300 - 311.

Gordon-Salant, S. Yeni-Komshian, G., Fitzgibbons, P., & Barrett, J. (2006). Age-related differences in identification and discrimination of temporal cues in speech segments. *Journal of the Acoustical Society of America, 119,* 2455-2466.

Gordon-Salant, S., Yeni-Komshian, G., & Fitzgibbons, P. (2008). The role of temporal cues in word identification by younger and older adults: Effects of sentence context. *Journal of the Acoustical Society of America*, *124,* 3249–3260.

Grant, K. (1987). Identification of intonation contours by normally hearing and profoundly hearing-impaired listeners. *Journal of the Acoustical Society of America, 82*, 1172–1178.

Greenwood, D. (1990). A cochlear frequency-position function for several species—29 years later. *Journal of the Acoustical Society of America*, *87,* 2592–2605.

Gruenenfelder, T. & Pisoni, D. (1980). Fundamental frequency as a cue to postvocalic consonantal voicing: Some data from perception and production. *Perception & Psychophysics*, *28,* 514 520.

Haggard, M. (1978). The devoicing of voiced fricatives. *Journal of Phonetics, 6,* 95-102.

Haggard, M., Ambler, A. & Callow, M. (1970). Pitch as a voicing cue. *Journal of the Acoustical Society of America, 47,* 613-617.

Halle, M., Hughes, G., & Radley, J.-P. (1957). Acoustic properties of stop consonants. *Journal of the Acoustical Society of America*, *29,* 107–116.

Han, D., Liu, B., Chen, X., Kong, Y., Liu, H., Zheng, Y. & Xu, L. (2009). Lexical tone perception with HiResolution and HiResolution 120 sound-processing strategies in pediatric Mandarin-speaking cochlear implant users. *Ear and Hearing, 30,* 169-177.

Hanson, H., & Stevens, K. (2002). A quasiarticulatory approach to controlling acoustic source parameters in a Klatt-type formant synthesizer using HLsyn. *Journal of the Acoustical Society of America, 112,* 1158-1182.

Hanson, H., Stevens, K., & Beaudoin, R. (1997). ‘New parameters and mapping relations for the HLsyn speech synthesizer. *Journal of the Acoustical Society of America, 102,* 3163.

Harris, K. (1958). Cues for the discrimination of fricatives in spoken syllables. *Language and Speech*, 1, 1-7.

Hawkins, S., & Nguyen, N. (2004). Influence of syllable-coda voicing on the acoustic properties of syllable-onset /l/ in English. *Journal of Phonetics, 32,* 199–231.

Hedrick, M. & Carney, A. (1997). Effect of relative amplitude and formant transitions on perception of place of articulation by adult listeners with cochlear implants. *Journal of Speech, Language and Hearing Research, 40,* 1445-1457.

Hedrick, M., & Ohde, R. (1993). Effect of relative amplitude of frication on perception of place of articulation. *Journal of the Acoustical Society of America, 94,* 2005–2026.

Heinz, J., & Stevens, K. (1961). On the properties of voiceless fricative consonants. *Journal of the Acoustical Society of America, 33,* 589–593.

Henry, B., Turner, C., & Behrens, A. (2005). Spectral peak resolution and speech recognition in quiet: Normal hearing, hearing impaired, and cochlear implant listeners. *Journal of the Acoustical Society of America, 118,* 1111–1121.

Hillenbrand, J. (2003). Some effects of intonation contour on sentence intelligibility. *Journal of the Acoustical Society of America, 114, 2338*.

Hillenbrand, J., Clark, M., & Houde, R. (2000). Some effects of duration on vowel recognition. *Journal of the Acoustical Society of America, 108,* 3013–3022.

Hillenbrand, J., & Gayvert, R. (1993). Identification of steady-state vowels synthesized from the Peterson–Barney measurements. *Journal of the Acoustical Society of America, 94,* 668–674.

Hillenbrand, J., & Gayvert, R. (2005). Open-source software for experiment design and control. *Journal of Speech Language and Hearing Research, 48,* 45-60.

Hillenbrand, J., Getty, L., Clark, M., & Wheeler, K., (1995). Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America, 97,* 3099-3111.

Hillenbrand, J., Ingrisano, D., Smith, B. & Flege, J. (1984). Perception of the voiced-voiceless contrast in syllable-final stops. *Journal of the Acoustical Society of America, 76,* 18-26.

Hillenbrand, J., & Nearey, T. (1999). Identification of resynthesized /hVd/ utterances: Effects of formant contour. *Journal of the Acoustical Society of America, 105,* 3509–3523.

Hogan, J. & Rozsypal, A. (1980). Evaluation of vowel duration as a cue for the voicing distinction in the following word-final consonant. *Journal of the Acoustical Society of America, 67,* 1764-1771.

Holt, L. & Idemaru, K. (2011). Generalization of dimension-based statistical learning. *Proceedings of the International Congress of Phonetic Sciences,* Hong Kong.

Holt, L. & Lotto, A. (2006). Cue weighting in auditory categorization: Implications for first and second language acquisition. *Journal of the Acoustical Society of America, 119,* 3059-3071.

Holt, L., Lotto, A., & Kluender, K. (2000). Neighboring spectral content influences vowel identification. *Journal of the Acoustical Society of America*, 108, 710-722.

Holt, L., Stephens, J., & Lotto, A. (2005). A critical evaluation of visually-moderated phonetic context effects. *Perception & Psychophysics*, 67, 1102-1112.

Hombert, J. (1975). Towards a theory of tonogenesis: An emipirical, physiologically and perceptually-based account of the development of tonal contrasts in language. (Unpublished doctoral dissertation). University of California, Berkely.

Hombert, J. (1978). Consonant types, vowel quality, and tone. In Fromkin, V. (Ed.) *Tone: A Linguistic Survey* (77-111). New York: Academic.

House, A., & Fairbanks, G. (1953). The influence of consonant environment upon the secondary acoustical characteristics of vowels. *Journal of the Acoustical Society of America, 25,* 105–113

House, A. (1961). On vowel duration in English. *Journal of the Acoustical Society of America, 33,* 1174-1178.

Houtgast, T., & Steeneken, H. (1973). The modulation transfer function in room acoustics as a predictor of speech intelligibility. *Acustica, 28,* 66–73.

Iverson, P., Smith, C., & Evans, B. (2006). Vowel recognition via cochlear implants and noise vocoders: Effects of formant movement and duration. *Journal of the Acoustical Society of America, 120,* 3998-4006.

Jenkins, J., Strange, W., & Edman, T. (1983). Identification of vowels in ‘vowelless’ syllables. *Perception and Psychophysics, 34,* 441–450.

Jiang, J., Chen, M. & Alwan, A. (2006). On the perception of voicing in syllable-initial plosives in noise. *Journal of the Acoustical Society of America, 119,* 1092-1105.

Johnson, K. (2006). Resonance in an exemplar-based lexicon: The emergence of social identity and phonology. *Journal of Phonetics, 34,* 485-499.

Johnson, K., Strand, E., & D’Imperio, M. (1999). Auditory-visual integration of talker gender in vowel perception. *Journal of Phonetics, 27,* 359–384.

Jongman, A., Wayland, R., & Wong, S. (2000). Acoustic characteristics of English fricatives. *Journal of the Acoustical Society of America, 108,* 1252–1263.

Kewley-Port, D., Pisoni, D., & Studdert-Kennedy, M. (1983). Perception of static and dynamic acoustic cues to place of articulation in initial stop consonants. *Journal of the Acoustical Society of America, 73,* 1779–1793.

Kewley-Port, D. & Zheng, Y. (1998). Modeling formant frequency discrimination for isolated vowels. *Journal of the Acoustical Society of America, 103,* 1654-1666.

Kirk, K., Tye-Murray, N., & Hurtig, R. (1992). The use of static and dynamic vowel cues by multichannel cochlear implant users. *Journal of the Acoustical Society of America, 91,* 3487–3498.

Klatt, D. (1976). Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *Journal of the Acoustical Society of America, 59,* 1208-21.

Kluender, K., Coady, J., & Kiefte, M. (2003). Sensitivity to change in perception of speech. *Speech Communication, 41,* 59-69.

Koelewijn, T., Zekveld, A., Festen, J. & Kramer, S. (2011). Pupil dilation uncovers extra listening effort in the presence of a single-talker masker. *Ear and Hearing, 32(5).* doi: 10.1097/AUD.0b013e3182362790

Kunisaki, O., & Fujisaki, H. (1977). On the influence of context upon perception of voiceless fricative consonants. *Annual Bulletin of the Research Institute for Logopedics and Phoniatrics, University of Tokyo, 11,* 85-91.

Ladefoged, P., & Maddieson, I. (1996). *The sounds of the world's languages*. Oxford: Blackwell.

Lahiri, A., Gewirth, L., & Blumstein, S. (1984). A reconsideration of acoustic invariance for place of articulation in diffuse stop consonants: Evidence from a cross-language study. *Journal of the Acoustical Society of America, 76,* 391–404.

Laures, J., & Weismer, G. (1999). The effects of a flattened fundamental frequency on intelligibility at the sentence level. *Journal of Speech Language and Hearing Research, 42,* 1148–1156.

Lehiste, I., & Peterson, G. (1961). Some basic considerations in the analysis of intonation. *Journal of the Acoustical Society of America, 33,* 419–425.

Li, F. & Allen, J. (2011). Manipulation of consonants in natural speech. *IEEE Transactions on Audio, Speech, and Language Processing, 19,* 496-504.

Li, F., Menon, A. & Allen, J. (2010). A psychoacoustic method to find the perceptual cues of stop consonants in natural speech. *Journal of the Acoustical Society of America, 127,* 2599–2610

Liberman, A. M., Delattre, P. C., and F. S. Cooper. (1958). Some cues for the distinction between voiced and voiceless stops in initial position. *Language and Speech*, 1, 153-167.

Lisker, L. (1975). Is it VOT or a first-formant transition detector? *Journal of the Acoustical Society of America, 57,* 1547–1551.

Lisker, L. (1978). Rapid vs. rabid: A catalogue of acoustic features that may cue the distinction. *Haskins Laboratories Status Report on Speech Research, SR-54,* 127-132.

Lisker, L. & Abramson, A. (1964). A cross-language study of voicing in stops: Acoustical measurements. *Word, 20,* 384-422.

Loizou, P. & Poroy, O. (2001). Minimum spectral contrast needed for vowel identification by normal hearing and cochlear implant listeners. *Journal of the Acoustical Society of America, 110,* 1619-1627.

Lorenzi, C., Gilbert, G., Carn, H., Garnier, S., & Moore, B. (2006). Speech perception problems of the hearing impaired reflect inability to use temporal fine structure. *Proceedings of the National Academy of Sciences, 103,* 18866– 18869.

Lotto, A., & Holt, L. (2000). The illusion of the phoneme. In S. J. Billings, J. P. Boyle, & A. M. Griffith (Eds.), *Chicago Linguistic Society, Volume 35: The Panels.* (pp. 191-204). Chicago Linguistic Society: Chicago.

Lotto, A. & Holt, L. (2006). Putting phonetic context effects into context: A commentary on Fowler (2006). *Perception & Psychophysics, 68,* 178-183.

Lotto, A., & Kluender, K. (1998). General contrast effects in speech perception: Effect of preceding liquid on stop consonant identification. *Perception & Psychophysics, 60,* 602-619.

Mann, V. (1980). Influence of preceding liquid on stop-consonant perception. *Perception & Psychophysics, 28,* 407-412.

Mann, V. (1986). Distinguishing universal and language-dependent levels of speech perception: Evidence from Japanese listeners' perception of English "l" and "r." *Cognition, 24,* 169-196.

Mann, V., & Repp, B. (1980). Influence of vocalic context on perception of the /ʃ/ - /s/ distinction. *Perception and Psychophysics, 28,* 213–228.

Mann, V., & Repp, B. (1981). Influence of preceding fricative on stop consonant perception. *Journal of the Acoustical Society of America, 69,* 548-558.

Mayo, C. & Turk, A. (2005). The influence of spectral distinctiveness on acoustic cue weighting in children's and adults' speech perception. *Journal of the Acoustical Society of America, 118,* 1730-1741.

McAdams, S. (1989). Segregation of concurrent sounds I: Effects of frequency modulation coherence. *Journal of the Acoustical Society of America, 86,* 2148-2159.

McClelland, J., Mirman, D., & Holt, L. (2006). Are there interactive processes in speech perception? *Trends in Cognitive Science, 10,* 363–369.

McGurk, H. & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264(5588),* 746–748

McMurray, B., Clayards, M., Tanenhaus, M., & Aslin, R. (2008). Tracking the timecourse of phonetic cue integration during spoken word recognition. *Psychonomic Bulletin and Review, 15,* 1064–1071.

Miller, G., & Nicely, P. (1955). An analysis of perceptual confusions among some English consonants. *Journal of the Acoustical Society of America, 27,* 338–352.

Miller, J. & Volaitis, L. (1989). Effect of speaking rate on the perceptual structure of a phonetic category. *Perception & Psychophysics, 46*, 505-512.

Miller, S., Schlauch, R. & Watson, P. (2010). The effects of fundamental frequency contour manipulations on speech intelligibility in background noise. *Journal of the Acoustical Society of America, 128,* 435-443.

Morrison, G. (2005). An appropriate metric for cue weighting in L2 speech perception: Response to Escudero & Boersma (2004). *Studies in Second Language Acquisition, 27,* 597-606.

Morrison, G. (2007). Logistic regression modelling for first- and second-language perception data. In M. Solé, P. Prieto, & J. Mascaró (Eds.), *Segmental and prosodic issues in Romance phonology* (pp. 219–236). Amsterdam: John Benjamins.

Morrison, G. & Kondaurova, M. (2009). Analysis of categorical response data: Use logistic regression rather than endpoint-difference scores or discriminant analysis. *Journal of the Acoustical Society of America*, *126,* 2159–2162.

Morrison, G. & Nearey, T. (2007). Testing theories of vowel inherent spectral change. *Journal of the Acoustical Society of America, 122,* EL15-22.

Munson, B. & Coyne, A. (2010). The influence of apparent vocal-tract size, contrast type, and implied sources of variation on the perception of American English voiceless lingual fricatives. *Journal of the Phonetic Society of Japan, 14*, 48-59.

Munson, B., Jefferson, S. & McDonald, E. (2006) The influence of perceived sexual orientation on fricative identification. *Journal of the Acoustical Society of America* 119, 2427–2437.

Munson, B. & Nelson, P. (2005). Phonetic identification in quiet and in noise by listeners with cochlear implants. *Journal of the Acoustical Society of America, 118,* 2607–2617.

Nearey, T., & Assmann, P. (1986). Modeling the role of vowel inherent spectral change in vowel identification. *Journal of the Acoustical Society of America, 80,* 1297–1308.

Nilsson, M., Soli, S. & Sullivan, J. (1994). Development of the Hearing in Noise Test for the Measurement of Speech Reception Thresholds in Quiet and in Noise. *Journal of the Acoustical Society of America, 95,* 1085-1099.

Nittrouer, S. (2004). The role of temporal and dynamic signal components in the perception of syllable-final stop voicing by children and adults. *Journal of the Acoustical Society of America, 115,* 1777–1790.

Nittrouer, S. (2005). Age-related differences in weighting and masking of two cues to word-final stop voicing in noise. *Journal of the Acoustical Society of America, 118,* 1072-1088.

Nittrouer, S. & Lowenstein, J. (2008). Spectral structure across the syllable specifies final-stop voicing for adults and children alike. *Journal of the Acoustical Society of America, 123,* 377-385.

Nittrouer, S., & Miller, M. (1997). Developmental weighting shifts for noise components of fricative-vowel syllables. *Journal of the Acoustical Society of America, 102,* 572–580.

Nittrouer, S., & Studdert-Kennedy, M. (1987). The role of coarticulatory effects in the perception of fricatives by children and adults. *Journal of Speech and Hearing Research 30*, 319–329.

Ohde, R. (1984) Fundamental frequency as an acoustic correlate of stop consonant voicing. *Journal of the Acoustical Society of America, 75,* 224–230.

Oxenham, A., &Simonson, A. (2009). Masking release for low- and high-pass-filtered speech in the presence of noise and single-talker interference. *Journal of the Acoustical Society of America*, *125,* 457-468.

Pakarinen, S., Lovio, R., Huotilainen, M., Alku, P., Näätänen, R. & Kujala, T. (2009). Fast multi-feature paradigm for recording several mismatch negativities (MMNs) to phonetic and acoustic changes in speech sounds. *Biological Psychology, 82,* 219-226.

Pakarinen, S., Takegata, R., Rinne, T., Huotilainen, M. &Näätänen, R. (2007). Measurement of extensive auditory discrimination profiles using the mismatch negativity (MMN) of the auditory event-related potential (ERP). *Clinical Neurophyschology, 118,* 177-185.

Parker, E., & Diehl, R. (1984). Identifying vowels in CVC syllables: Effects of inserting silence and noise. *Perception & Psychophysics, 36,* 369-380

Peng, S-C., Lu, N. & Chatterjee, M. (2009). Effects of cooperating and conflicting cues on speech intonation recognition by cochlear implant users and normal hearing listeners. *Audiology and Neurotology, 14,* 327-337.

Perry, T., Ohde, R., & Ashmead, D. (2001). "The acoustic bases for gender identification from children's voices. *Journal of the Acoustical Society of America, 109,* 2988–2998.

Phatak, S., & Allen, J. (2007). Consonant and vowel confusions in speech-weighted noise. *Journal of the Acoustical Society of America, 121,* 2312–2316.

Phatak, S., Lovitt, A., & Allen, J. (2008). Consonant confusions in white noise. *Journal of the Acoustical Society of America*, *124,* 1220-1233.

R Development Core Team (2010). R: A language and environment for statistical computing. [R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0]. [Computer software]. Available from http://www.R-project.org/

Raphael, L. (1972). Preceding vowel duration as a cue to the perception of the voicing characteristic of word-final consonants in American English. *Journal of the Acoustical Society of America, 51,* 1296-1303.

Read, C., Zhang, Y., Nie, H., & Ding, B. (1986). The ability to manipulate speech sounds depends on knowing alphabetic writing. *Cognition,* 24, 31-44.

Régnier, M., & Allen, J. (2008). A method to identify noise-robust perceptual features: Application for consonant /t/. *Journal of the Acoustical Society of America, 123,* 2801–2814.

Repp, B. (1982). Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception. *Psychological Bulletin, 92,* 81-110.

Revoile, S., Pickett, J. & Holden, L. (1982). Acoustic cues to final stop voicing for impaired- and normal-hearing listeners. *Journal of the Acoustical Society of America, 72,* 1145-1154.

Rosen, S., Faulkner, A., & Wilkinson, L. (1999) Adaptation by normal listeners to upward spectral shifts of speech: Implications for cochlear implants. *Journal of the Acoustical Society of America, 106,* 3629-3636.

Sato, M., Cavé, C., Ménard, L., & Brasseur., A. (2010). Auditory-tactile speech perception in congenitally blind and sighted adults. *Neuropsychologia*, 48, 3683-6.

Schvartz, K., Chatterjee, M., & Gordon-Salant, S. (2008). Recognition of spectrally degraded phonemes by younger, middle-aged, and older normal-hearing listeners. *Journal of the Acoustical Society of America*, *124,* 3972 – 3988.

Schwartz, M. (1968). Identification of speaker sex from isolated, voiceless fricatives. *Journal of the Acoustical Society of America*, *43,* 1178–1179.

Scott, T., Green, W. B., & Stuart, A. (2001). Interactive effects of lowpass filtering and masking noise on word recognition. *Journal of the American Academy of Audiology 12,* 437–444.

Shannon, R. (1989). Detection of gaps in sinusoids and pulse trains by patients with cochlear implants. *Journal of the Acoustical Society of America*, 85, 2587-2592.

Shannon, R. (1992). Temporal modulation transfer functions in patients with cochlear implants. *Journal of the Acoustical Society of America, 91,* 2156–2164.

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

Smith, C. (1996). The devoicing of /z/ in American English: effects of local and prosodic context. *Journal of Phonetics, 25,* 471-500.

Soli, S. (1982). Structure and duration of vowels together specify fricative voicing. *Journal of the Acoustical Society of America*, *72,* 366-378.

Soli, S., & Arabie, P. (1979). Auditory versus phonetic accounts of observed confusions between consonant phonemes. *Journal of the Acoustical Society of America, 66,* 46–59.

Stevens, K. (2002). Toward a model of lexical access based on acoustic landmarks and distinctive features. *Journal of the Acoustical Society of America*, *111,* 1872–1891.

Stevens, K., Blumstein, S., Glicksman, L., Burton, M., & Kurowski, K. (1992). Acoustic and perceptual characteristics of voicing in fricatives and fricative clusters. *Journal of the Acoustical Society of America, 91,* 2179-3000.

Stickney, G. & Assmann, P. (2001). Acoustic and linguistic factors in the perception of bandpass-filtered speech. *Journal of the Acoustical Society of America, 109,* 1157-1165.

Stilp, C. & Kluender, K. (2010). Cochlea-scaled entropy, not consonants, vowels, or time, best predicts speech intelligibility. *Proceedings of the National Academy of Sciences, 107,* 12387–12392

Strand, E. (1999). Uncovering the role of gender stereotypes in speech perception. *Journal of Language and Social Psychology, 18,* 86–99.

Strand, E., & Johnson, K. (1996). Gradient and visual speaker normalization in the perception of fricatives. In Gibbon, D. (Ed.) *Natural Language Processing and Speech Technology: Results of the 3rd KONVENS Conference, Bielfelt* (14-26). Berlin: Mouton de Gruyter.

Stuart, A., Phillips, D., & Green, W. (1995). Word recognition performance in continuous and interrupted noise by normal-hearing and simulated hearing-impaired listeners. *Journal of Otology, 16,* 658-663.

Summerfield, A.Q, Nakisa, M., McCormick, B., Archbold, S., Gibbon, K. & O’Donoghue, G. (2002). Use of Vocalic Information in the Identification of /s/ and /ʃ/ by Children with Cochlear Implants. *Ear & Hearing, 23,* 58–77.

Summers, W. V. (1988). F1 structure provides information for final consonant voicing. *Journal of the Acoustical Society of America, 84,* 485–492.

Systat (2004). SigmaPlot (version 9.01). [Computer program].

Toscano, J. & McMurray, B. (2010). Cue integration with categories: Weighting acoustic cues in speech using unsupervised learning and distributional statistics. *Cognitive Science, 34,* 434-464.

Turner, C., & Brus, S. (2001). Providing low- and mid-frequency speech information to listeners with sensorineural hearing loss. *Journal of the Acoustical Society of America, 109,* 2999–3006.

Vaida, F. & Blanchard, S. (2005). Conditional Akaike information for mixed-effects models. *Biometrika, 92,* 351-370.

van Bezooijen, R. (1995). Sociocultural aspects of pitch differences between Japanese and Dutch women. *Language and Speech, 38,* 253-256.

van Dommelen, W., & Moxness, B. (1995). Acoustic parameters in speaker height and weight identification: sex-specific behaviour. *Language and Speech, 38,* 267–287.

Vandali, A., Sucher, C., Tsang, D., McKay, C., Chew, J. &, McDermott, H. (2005). Pitch ranking ability of cochlear implant recipients: a comparison of sound-processing strategies. *Journal of the Acoustical Society of America*, *117*, 3126-3138.

Vickers, D., Moore, B., & Baer, T. (2001). Effects of low-pass filtering on the intelligibility of speech in quiet for people with and without dead regions at high frequencies. *Journal of the Acoustical Society of America, 110,* 1164–1175.

Vroomen, J., & de Gelder, B. (2001). Lipreading and the compensation for coarticulation mechanism. *Language & Cognitive Processes*, *16*, 661-672.

Walsh, T. & Parker, F. (1984). A review of the vocalic cues to [+- voice] in post-vocalic stops in English. *Journal of Phonetics 12,* 207-218.

Wang, M. D., & Bilger, R. C. (1973). Consonant confusions in noise: a study of perceptual features. *Journal of the Acoustical Society of America, 54,* 1248–1266.

Wardrip-Fruin, C. (1982). On the status of temporal cues to phonetic categories: Preceding vowel duration as a cue to voicing in final stop consonants." *Journal of the Acoustical Society of America, 71,* 187-195.

Wardrip-Fruin, C. (1985). The effect of signal degradation on the status of cues to voicing in utterance-final stop consonants. *Journal of the Acoustical Society of America, 77,* 1907-1912.

Warren, P. & Marslen-Wilson, W. (1989). Cues to lexical choice: Discriminating place and voice. *Perception & Psychophysics, 43,* 21-30.

Wayland, S., Miller, J. & Volaitis, L. (1992). The influence of sentence articulation rate on the internal structure of phonetic categories. *Journal of the Acoustical Society of America, 128*, 2465.

Whalen, D. (1984). Subcategorical phonetic mismatches slow phonetic judgments. *Perception & Psychophysics, 35,* 49-64.

Whalen, D., Abramson, A., Lisker, L., & Mody, M. (1993). F0 gives voicing information even with unambiguous voice onset times. *Journal of the Acoustical Society of America, 93,* 2152–2159.

Whalen, D. & Levitt, A. (1995). The universality of intrinsic F0 of vowels. *Journal of Phonetics,* 23, 349-366.

Xu., L. & Pfingst (2003). Relative importance of temporal envelope and fine structure in lexical-tone perception. *Journal of the Acoustical Society of America, 114,* 3024–3027.

Xu, L., Thompson, K. & Pfingst, B. (2005). Relative contributions of spectral and temporal cues for phoneme recognition. *Journal of the Acoustical Society of America, 117,* 3255-3267.

Zahorian, S. A., & Jagharghi, A. J. (1993). Spectral-shape features versus formants as acoustic correlates for vowels. *Journal of the Acoustical Society of America, 94,* 1966–1982.

Zeng, F-G, Rebscher, S., Harrison, W., Sun, X., & Feng, H. (2008). Cochlear implants: System design, Integration and Evaluation. *IEEE Reviews in Biomedical Engineering, 1,* 115-142.

Zwicker, E., & Terhardt, E. (1980). Analytical expressions for critical-band rate and critical bandwidth as a function of frequency. *Journal of the Acoustical Society of America, 68,* 1523-1525.