Role of F0 in Speech Reception in the Presence of Interference: Simulating Aspects of Cochlear-Implant Processing

by
Michael Kaige Qin

B. S. Electrical Engineering
Purdue University, 1996

SUBMITTED TO THE HARVARD-MIT
DIVISION OF HEALTH SCIENCES AND TECHNOLOGY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE
DEGREE OF

DOCTOR OF PHILOSOPHY
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
[February 2005]
January 7, 2005

Author..
Harvard-Massachusetts Institute of Technology, Division of Health Sciences and Technology
January 7, 2005

Certified by..
Andrew J. Oxenham
Principal Research Scientist, Research Laboratory of Electronics
Lecturer, Harvard-MIT Division of Health Sciences and Technology
Thesis Supervisor

Accepted by..
Martha Gray
Edward Hood Taplin Professor of Medical & Electrical Engineering
Co-Director, Harvard-MIT Division of Health Sciences and Technology

Abstract

Speech is perhaps the most ecologically important acoustic stimulus to human beings, because it remains the primary means by which people interact socially. Despite many significant advances made in the development of cochlear implants, even the most successful cochlear-implant users do not hear as well as normal-hearing listeners. The differences in performance between normal-hearing listeners and cochlear-implant users are especially pronounced in understanding speech in complex auditory environments. For normal-hearing listeners, voice pitch or the fundamental frequency (F0) of voicing has long been thought to play an important role in the perceptual segregation of speech sources. The aim of this dissertation was to examine the role of voice pitch in speech perception in the presence of background interference, specifically simulating aspects of envelope-vocoder style implant processing. The findings of the studies show that F0 encoded via envelope periodicity does not provide a sufficiently salient cue for the segregation of speech. This suggests that the poor speech reception performance of implant users in background interference may, at least in part, be due to the lack of salient voice pitch cues. When low-frequency fine-structure information was added to envelope-vocoder processed high-frequency information, some F0 segregation benefits returned and the reception of speech in complex backgrounds improved. Taken as a whole, the dissertation suggests that low frequency fine-structure information is important to the task of speech segregation, and that every effort should be made to present such information to cochlear-implant users.

Acknowledgments

This work was supported by the National Institute on Deafness and Other Communication Disorders (NIDCD) grant number DC05216 and the Harvard-MIT Division of Health Sciences and Technology Speech & Hearing Bioscience and Technology Fellowship.

I would like to first express my gratitude to Andrew Oxenham, my thesis advisor. His vast knowledge and clarity of communication were indispensable to my graduate experience. I am forever indebted to him for his mentorship, tutorage, and considerable patience.

I would like to acknowledge with much appreciation the other members of my committee, Louis Braida, Barbara Shinn-Cunningham, Don Eddington, and Christophe Micheyl for their assistance at all levels of the research project. I would also like to thank those who have provided me with advice at times of critical need, Joshua Bernstein, Evan Chen, Joe Frisbie, Jean Krause, Peninah Rosengard, and Andrea Simonson.

Thanks also go out to my friends at the Sensory Communication Group and the Speech & Hearing Program, for our philosophical debates, exchanges of knowledge, and venting of frustrations. I would also like to thank the office staff for all their kindness and assistance along the way.

I like to thank my mother and father for the support and perpetual forbearance they have provided me throughout my life. I must acknowledge my sine qua non, Laura, without whose love, encouragement, and assistance, I would not have finished this thesis.

Table of content

"The comfort and advantage of society not being to be had without communication of thoughts, it was necessary that man should find out some external sensible signs, whereof those invisible ideas, which his thoughts are made up of, might be made known to others. For this purpose nothing was so fit, either for plenty or quickness, as those articulate sounds, which with so much ease and variety he found himself able to make."

John Locke, "An essay concerning human understanding"

Chapter 1: General introduction

For normal-hearing listeners, speech is a highly effective and extremely robust medium for communication, resistant to the deleterious effects of masking, reverberation, and many kinds of other signal distortions (Fletcher and Galt, 1950; Miller and Licklider, 1950). However, while normal-hearing listeners are able to cut through even the most severe interference, hearing-impaired listeners are generally less successful. The difference in performance between normal-hearing listeners and the hearing-impaired is especially pronounced in understanding speech in the presence of temporally and spectrally fluctuating background interference (e.g., Peters *et al.*, 1998).

Unfortunately, many modern acoustic backgrounds such as traffic noise and competing conversations are fluctuating in nature. Normal-hearing listeners have been shown to take advantage of the dips and valleys inherent in fluctuating interferers. For example, in the presence of a competing voice, both temporal and spectral dips exist. The temporal dips arise because there are moments when the overall level of the competing speech is low, for example during brief pauses in the speech or during production of low-energy sounds such as /m/, /n/, /k/, or /p/. During these temporal dips the target-to-masker ratio is high, and this allows brief ''glimpses'' to be obtained of the target speech. The spectral dips arise because the spectrum of the target speech is usually different from that of the background speech measured over any short interval. Although parts of the target spectrum may be completely masked by the background, other parts may be hardly masked at all. Thus, parts of the spectrum of the target speech may be "glimpsed." These "glimpses" of the target speech can often provide sufficient information to allow the listener to infer the entire message.

1.1 Glimpsing

Glimpsing, or "dip listening," has been proposed as an explanation for the finding that fluctuating interference produces less masking of speech than steady-state maskers for normal-hearing listeners. Similarly, a reduction in the ability to glimpse has been proposed as an explanation for the performance difference between normal-hearing and hearing-impaired listeners in the presence of fluctuating interference.

The idea of glimpsing was first forwarded by Miller and Licklider (1950). They observed that speech signals could be turned on and off periodically without substantial loss of intelligibility. They found that intelligibility was lowest for interruption rates below 2 Hz, where large fragments of each word are omitted. If the interruption rate was higher (between 10 and 100 Hz) listeners identified more than 80% of the monosyllabic words correctly. Miller and Licklider suggested that listeners were able to somehow "patch together" successive glimpses of speech to form linguistic percepts, and therefore reconstruct the intended message from glimpses of the speech signal.

Improvements in speech reception as a product of modulations in the masker are referred to as masking release. The amount of masking release in normal-hearing listeners ranges from less than 5 dB to as much as 20 dB, depending on the target stimuli and the temporal and spectral characteristics of the maskers (Bacon *et al.*, 1998). Listeners with hearing loss are less able than normal listeners to obtain release from modulated maskers (e.g., Festen and Plomp, 1990; Eisenberg *et al.*, 1995; Bacon *et al.*, 1998; Peters *et al.*, 1998).

Festen and Plomp (1990) measured the speech reception threshold (SRT) for sentences presented in fluctuating background interference for normal-hearing listeners and listeners with moderate hearing-impairment. The interfering sounds were steady-state noise, modulated noise, and a single competing voice. Their results showed that, for normal-hearing listeners, the SRT for sentences in modulated noise was 4-6 dB lower than for steady-state noise. For listeners with moderate sensorineural hearing loss, they obtained elevated SRTs, without appreciable effect of masker fluctuations. They suggested that the mechanisms contributing to the absence of masking release were abnormal forward masking, reduced temporal resolution, and a reduction in co-modulation masking release.

Eisenberg et al. (1995) tested listeners with normal hearing and listeners with hearing loss for their understanding of consonants in steady and fluctuating noise. Listeners with normal hearing were tested with shaped-noise designed to simulate the hearing sensitivity of the impaired listeners. Their results suggested that listeners with true hearing loss obtained far less release from modulated maskers than did normal-hearing listeners with or without simulated hearing losses. They found that amplification restored some, but not all, of the expected release from masking for impaired listeners. They concluded that audibility alone could not explain the additional masking experienced by listeners with sensorineural hearing loss.

A subsequent study by Bacon et al. (1998) evaluated listeners' understanding of sentences in speech-shaped noise that was modulated by the envelope of one of the following: steady-state noise, multi-talker babble, single-talker babble, and a 10-Hz square wave with 100% modulation depth. They observed that for normal-hearing listeners, the square-wave modulation provided the greatest release from masking. In addition, they found that the impaired listeners obtained significantly less release from masking than did their normal-hearing counterparts. Noise-masked normal-hearing listeners obtained somewhat less masking release than they had with full access to the signals. They suggested that while audibility may account for some loss of masking release, excessive forward masking in impaired ears might account for the additional loss.

Thus far, the consequences of hearing impairment have only been discussed in terms of increased masking due to reduced frequency selectivity and increased forward masking.

However, the ability to benefit from glimpsing may in fact depend on at least two separate processes. First, the targets must be audible (i.e. above the masked threshold or absolute threshold). Second, the listener must have some basis for distinguishing the target from the masker. In order to exploit the benefits of glimpsing the auditory system must solve the problem of detection and segregation. There is the possibility that even if the interference does not render the target inaudible through energetic masking, listeners may not be able to perceptually separate the target from interference, producing something akin to informational masking (e.g., Durlach *et al.*, 2003). This is a potentially important change in emphasis, as it suggests that it is not just the short-term target audibility, but also the inability to distinguish target from background interference that limits performance, particularly for the hearing-impaired.

Informational masking is thought of as a threshold elevation due to non-energetic factors, such as stimulus uncertainty or masker-target similarity. The presence of a cue that reduces the similarity between the target and masker can presumably reduce the effects of informational masking. Voice pitch, or the fundamental frequency (F0) of voicing, has long been thought to be a powerful primitive grouping cue, playing an important role in the perceptual segregation of speech sources (e.g., Bregman, 1990; Darwin and Carlyon, 1995). The aim of this dissertation is to examine the role of pitch perception, in our ability to perceive speech in the presence of complex background interference.

1.2 Pitch

1.2.1 Voice pitch

According to the source-filter model of speech production, the speech signal can be considered as the output of a linear system. Depending on the type of input excitation (source), two classes of speech sounds are produced, namely voiced and unvoiced. If the input excitation is noise, then unvoiced sounds like /s/, /t/, etc. are produced, and if the input excitation is periodic then voiced sounds like /a/, /i/, etc., are produced. In the unvoiced case, noise is generated either by forcing air through a narrow constriction (e.g., production of /f/) or by building air pressure behind an obstruction and then suddenly releasing that pressure (e.g., production of /t/). In contrast, the excitation used to produce voiced sounds is periodic and is generated by the vibrating vocal cords. Voiced speech signals, such as vowels, can be decomposed into a series of discrete sinusoids, called harmonics. The frequencies of these harmonics are integer multiples of a common fundamental frequency (F0). A single pitch corresponding to the F0 is generally heard.

1.2.2 Utility of voice pitch

Studies of normal-hearing listeners have found that in the presence of a competing voice, listeners generally find it easier to understand the target voice if the competing voice has a different F0 (see, Bregman, 1990; Darwin and Carlyon, 1995). One source of evidence for the contribution of F0 comes from studies of the perception of concurrent vowels. Scheffers (1983), studied the effects of a fundamental frequency difference on identification of simultaneous synthetic speech sounds using a pair of steady-state, synthetic vowels. The vowels were presented simultaneously, with identical onset and offset, at the same amplitude, monaurally or diotically. Subjects were required to identify both vowels. Scheffers reported that the identification performance improved when a ΔF0 was introduced. The improvement increased rapidly as ΔF0 increased from 0 to about 1 semitone, then asymptote (typically at between 60 - 80%) between 1 and 2 semitones. He noted that subjects commonly reported the subjective impression of a single talker when presented with vowels with the same F0, whereas stimuli with a ΔF0 give the impression of two talkers.

Brokx and Nooteboom (1982) provided another source of evidence with their experiment on concurrent sentences. They used a linear predictive coding vocoder to artificially modify the characteristics of the excitation source to produce synthesized, monotone sentences. They then varied the difference in F0 between the target sentence and a continuous speech masker. They found that identification accuracy was lowest when the target and masker had the same F0, and improved with increasing difference in F0. Compared to concurrent vowels, the identification function for concurrent sentences did not flatten out between 1 and 2 semitones, but instead showed a continued increase. These results were replicated and extended by Bird and Darwin (1998), using monotone versions of short declarative sentences consisting of mainly voiced sounds. They found little to no improvement when F0 differences were less than 1 semitone, but above that identification performance improved up to 8 semitones. Overall, the influence of F0 differences was greater for concurrent sentences than for concurrent vowels.

1.2.3 Pitch perception

Before proceeding further, it is worthwhile to briefly discuss several aspects of the peripheral auditory system that may directly affect the perception of pitch. Fig. 1-1 is an illustration of the analysis of a tone complex (F0 = 100 Hz) in the peripheral auditory system.

It is well known that the basilar membrane (BM) in the cochlea separates the different frequency components of the incoming signal along its length (Moore, 2003b). The low-frequency components excite the apex of the BM whereas high-frequency components excite the base. Each place on the BM is sensitive to a limited range of frequency components. The BM is often modeled as a bank of overlapping bandpass filters. Notice here that the tone complex contains a number of equal-amplitude harmonics. Notice also that the auditory filters have a bandwidth that is roughly a tenth of their center frequency (and so is roughly constant on a log scale), whereas harmonics are equally spaced in frequency on a linear scale. This leads to the lower-order harmonics being resolved on the BM, producing distinct peaks in the excitation pattern of activity. At a place tuned to the frequency of a lower-order harmonic, the waveform on the BM is approximately a sine wave at the harmonic frequency. In contrast, the higher-order harmonics are unresolved, and do not give rise to distinct peaks in the BM. Instead, these harmonics interact with each other in the filters to give rise to a complex vibration that shows beats at the F0.

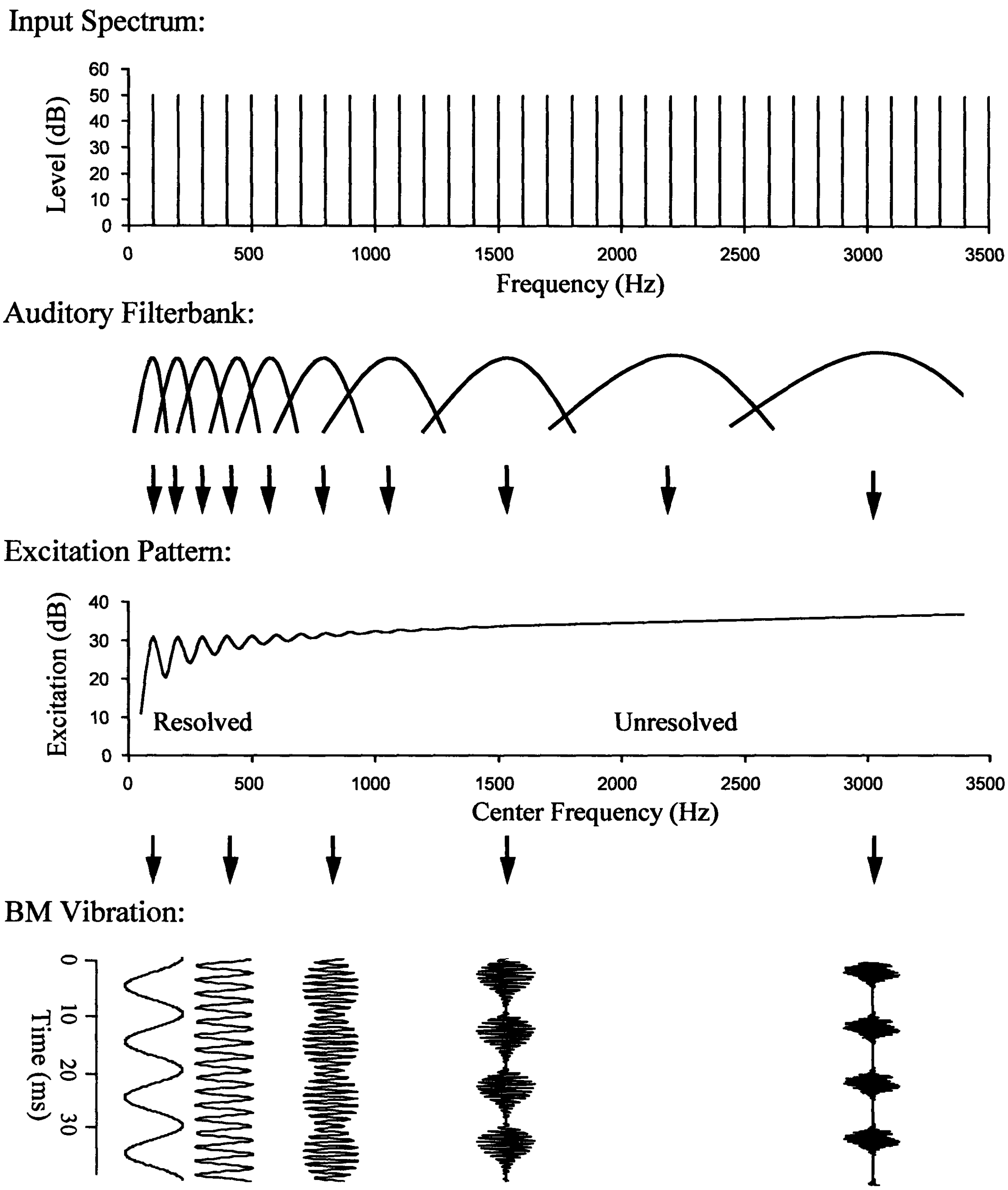

Figure 1-1: This is an illustration of the analysis of a tone complex (F0 = 100Hz) in the peripheral auditory system. (Figure adopted from Plack and Oxenham, "Psychophysical studies of pitch", Delmenhorst, Aug. 2002).

Auditory nerve fibers are also known to have the tendency to fire at a particular phase of the stimulus waveform on the basilar membrane (i.e. they are phase-locked to the input). Furthermore, it is known that phase locking decreases with increasing frequency. In fact, psychoacoustic estimates put the limit of phase-locking at around 4-5 kHz for humans (e.g., Moore, 1973). Therefore, while auditory nerve fibers may effectively phase-lock to resolved low-frequency components, they are generally less effective at phase locking to high-frequency components. However, if several high-frequency components fall into the same auditory filter (see Fig. 1-1), auditory nerve fibers can potentially phase-lock to the envelope modulations/beats created by the interaction of these components.

Although the term pitch is often used interchangeably with F0, it is important to keep in mind that pitch is a percept, whereas F0 is a description of the physical stimulus. It is also important to remember that F0 (the common fundamental frequency) is not the same as the fundamental component. In fact, the fundamental component does not have to be present for the pitch of a stimulus to be perceived. Schouten (1940) goes further to suggest that the harmonics, not the fundamental component, of a complex tone make the greatest contribution to the perception of the pitch. Studies by Ritsma (1967) and Plomp (1967) observed that for F0s in the range of human voicing (100 Hz – 400 Hz), the 3rd, 4th, and 5th harmonics tends to dominate the pitch sensation. These findings have been broadly confirmed, though the data from Moore et al. (1984; 1985b) show that some individual differences exist in precisely which harmonic is the most dominant. Although the general consensus is that resolved low-order harmonics dominate the pitch percept, pitch perception has also been demonstrated using only high-order unresolved harmonics (e.g., Houtsma and Smurzynski, 1990) as well as amplitude-modulated noise (Burns and Viemeister, 1976; Burns and Viemeister, 1981). However, the pitch evoked by unresolved harmonics is generally less salient than the pitch associated with resolved low-order harmonics.

While pitch perception remains a subject of continuing investigation and controversy, there is now consensus about some aspects of pitch perception of harmonic complex stimuli. For normal-hearing listeners, the perception of voice pitch and the ability to discriminate different F0s is believed to rely primarily on the information carried in peripherally resolved lower-order harmonics (e.g., Plomp, 1967; Ritsma, 1967; Moore *et al.*, 1985b; Houtsma and Smurzynski, 1990; Dai, 2000; Smith *et al.*, 2002). Under normal circumstances, the frequencies of these harmonics may be encoded by their place of excitation on the basilar membrane, by the temporal pattern of their auditory-nerve responses, or by some combination of the two.

1.2.4 Effects of peripheral impairment on pitch

The auditory filter shapes of cochlear-impaired listeners have been estimated in several studies (e.g., Glasberg and Moore, 1986; Peters and Moore, 1992; Stone *et al.*, 1992; Leek and Summers, 1993). The results generally agree in showing that auditory filters in hearing-impaired listeners are broader than normal. Despite some scatter in the degree of broadening of auditory filters (Stone *et al.*, 1992), on average, the degree of broadening increases with increasing hearing loss.

As mentioned above, resolved harmonics are believed to produce more salient pitch percepts than unresolved harmonics. The distinctions between pitch derived from resolved and unresolved harmonics may be of special relevance when considering the pitch perception of hearing-impaired listeners. Take, for example, an impaired ear with auditory filters three times broader than normal. The 4th and 5th harmonics of a tone complex with an F0 of 200 Hz, well resolved in a normal auditory system, would be poorly resolved in the impaired ear, which may translate into poorer pitch perception.

Pitch discrimination abilities for complex tones by hearing-impaired people has been the subject of several studies (e.g., Moore and Glasberg, 1988; Moore and Peters, 1992; Arehart, 1994). These studies have required subjects to identify which of two successive harmonic complex tones had the higher F0 (corresponding to a higher pitch). The thresholds determined in such a task are described as F0 difference limens (F0DLs). Overall, the results suggest that, relative to normal-hearing listeners, people with cochlear damage depend more on temporal information from unresolved harmonics than on spectral/temporal information from resolved harmonics. Moreover, these studies revealed that F0 discrimination performance was clearly worse for subjects in the hearing-impaired group than the normal-hearing group.

In a study to examine the ability of listeners to utilize F0 difference in the perceptual segregation of simultaneous speech, Summers and Leek (1998) measured both F0DLs for individual synthetic vowels and the ability to identify concurrent vowels. They found that, as a group, normal-hearing listeners benefited more from F0 differences between concurrent vowels than hearing-impaired listeners. They also found that hearing-impaired listeners with small F0DLs obtained benefit when the ΔF0 was increased up to four semitones, while listeners with large F0DLs showed no benefit of F0 separation. Their findings suggest that F0 discriminability may be predictive of performance in identifying concurrent vowels differing in F0.

1.3 Pitch processing by cochlear implants

For centuries, people believed that only a miracle could restore hearing to the deaf. Today, cochlear implants, prosthetic device implanted in the inner ear, can restore partial hearing to the sensorineural deaf. Unlike acoustic hearing aids that amplify sounds, cochlear implants operate by bypassing the outer, middle, and inner ear to directly stimulate the auditory nerve fibers. Cochlear implants are based on the premise that while an individual may have sensorineural deafness, there are still sufficient auditory nerve fibers remaining for stimulation. The aim of a cochlear implant is to generate, via electrical stimulation, patterns of neural activity that convey to a listener the information in the auditory environment.

Most cochlear implant recipients today are implanted with multi-channel devices (Fig. 1-2). In a multi-channel implant, an electrode array is inserted into the scala tympani of the cochlea to stimulate the nearby auditory neurons in the modiolus. By stimulating toward the apical part of the cochlea with low-frequency information and the basal part with high-frequency information, the electrode array is designed to take advantage of the natural tonotopic organization of the cochlea.

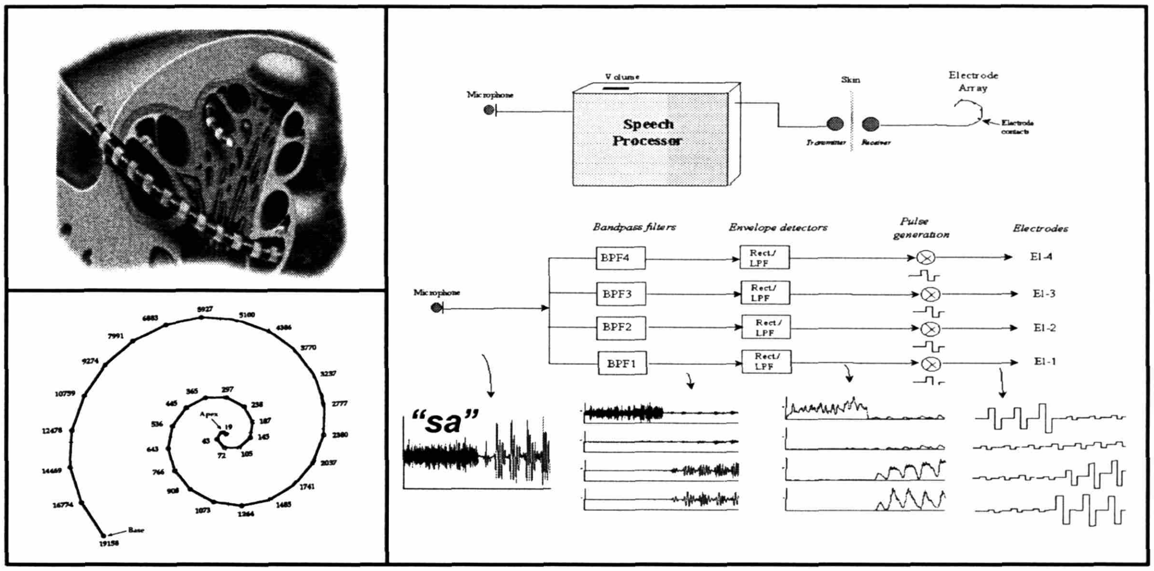

Figure 1-2: The multi-channel implants were developed to take advantage of the tonotopic organization in the cochlea, i.e. the apical part of the cochlea encodes low frequency and the basal part encodes high frequency. All of the multi-channels implants have some sort of bank of bandpass filters to divided incoming signal into different frequency bands. For instance, if the speech signal contains mostly high frequency information (e.g., /s/), then the fourth channel will be large relative to the amplitudes of channels 1-3. Similarly, if the speech signal contains mostly low frequency information (e.g., /a/) then the amplitude of the first and second channels will be large relative to the amplitudes of channels 3 and 4. The electrodes are stimulated according to the energy level of each frequency channel. (Figure adopted from (Loizou, 1999)).

Several sound-processing schemes have been developed over the years for multi-channel implants. One of the most widely used implants signal-processing schemes today is the envelope-vocoder processing scheme. With the envelope-vocoder scheme, sound is passed through a bank of (typically 6-8) bandpass filters, and the envelope of each filter output is obtained via rectification and lowpass filtering. The envelope from each band is then used to modulate a train of biphasic pulses on the appropriate electrode. The pulses are presented continuously, but are interleaved between the different electrodes. Two consequences of this processing scheme are worth noting at this point: 1) the lower-order harmonics of voice speech are not resolved by the coarse bandpass filtering; and 2) the lowpass filtering of the envelopes will eliminate any temporal fine-structure cues.

In the past decade, speech recognition by cochlear-implant users has improved significantly. Approximately half of the population of adult patients fit with the current generation of implants achieves scores of 80% – 100% correct on sentence recognition tests when the sentences are presented in isolation (Dorman, 2000). As cochlear implant technology matures, the number of cochlear implant users has grown exponentially to a total of 60,000 worldwide, including 20,000 children (Zeng, 2004). However, despite enormous advances made in the development of these hearing prostheses, the speech reception performance of most cochlear implant users is still not comparable to that of normal-hearing listeners. The differences in performance between normal-hearing listeners and cochlear-implant users are especially pronounced for speech reception in complex auditory environments. Fu et al. (1998a) and Friesen et al. (2001) reported that implant users required higher target-to-masker ratios to achieve levels of speech reception performance comparable to normal-hearing listeners. Given that these findings parallel those of cochlear-impaired listeners discussed above, it seems reasonable to consider cochlear-implant users as a special class of cochlear-impaired listeners.

While the underlying mechanisms of auditory perception differ between implant users and other cochlear-impaired listeners, both are likely to shift away from resolved harmonics dominating the pitch percept to an increased reliance on unresolved harmonics. For cochlear implant users, the reliance on unresolved harmonics is likely due to the relatively broad analysis filters used in speech processors, combined with the spread of electrical charge along and across the cochlea. However, in principle, voice pitch cues can still be preserved in the envelope periodicity of implant-processed speech (Green *et al.*, 2002; Moore, 2003a), provided that the cutoff frequency of the envelope-extraction filter is sufficiently high to allow the F0 to pass, akin to the pitch derived from unresolved harmonics in acoustic hearing.

McKay *et al.* (1994), investigating the pitch associated with sinusoid amplitude-modulated pulse trains, reported that implant users can potentially detect pitch differences as small as 2%, in the range of human voice pitch. However, most studies have shown poorer performance. For instance, Geurts and Wouters (2001) measured F0 difference limens using synthetic vowels. They found that implant users could detect F0 differences of between 4% and 13%. While these findings show that implant users have access to F0 information via envelope periodicity, they also highlight the weakness of the pitch percept compared to normal-hearing listeners, whose F0DLs are typically well below 1%.

By varying the envelope cutoff frequency of their implant sound-processing simulator, Faulkner *et al.* (2000) manipulated the amount of voicing information present in their processed stimuli. They found that voicing information had little or no effect on intelligibility in almost all their conditions. Their results could be interpreted as evidence against the importance of pitch or voicing information to speech reception. However, it is important to note that Faulkner *et al.* (2000) presented their stimuli in isolation; the major contribution of voicing and pitch information to understanding speech in the everyday setting may lie in the role it plays in segregating a target from interference. Currently, little is known about the effectiveness of envelope cues in conveying F0 information for the purpose of perceptual segregation.

1.4 Noise-excited vocoder as a research tool

While investigations using real cochlear-implant users are direct, they can be problematic. Many factors are known to affect the performance of implant users, such as: 1) the duration of deafness, 2) the age of onset of deafness, 3) the age at implantation, 4) the duration of implant use, 5) the number of surviving spiral ganglion cells, 6) the electrode placement and insertion depth, and 7) the electrical dynamic range. Any attempts at interpreting the findings will undoubtedly be clouded by these differences between implant users.

In an effort to minimize the variability inherent in testing implant users, an acoustic vocoder method (i.e., noise-excited vocoder) was developed to simulate the effects of implant signal processing (Shannon *et al.*, 1995). While the noise-excited vocoder is not an appropriate simulation for all implant processing schemes, most notably analog-based schemes such as Compress Analog (Eddington, 1980) and Simultaneous Analog Stimulation (Kessler, 1999), it does provide a simple and straightforward way of simulating many aspects of hearing impairment and implant envelope-vocoder processing (see Fig. 1-3). By using only a small number of frequency bands, this form of processing mimics the limited information in the spectral distribution of energy available through implant systems. Temporal information carried by these simulations also mimics that carried by envelope-vocoder processors, with temporal fine structure being eliminated by envelope smoothing, and lower-rate temporal envelope information being preserved. Provided that there is sufficient envelope bandwidth, a noise-excited vocoder is capable of conveying pitch information for modulation rates up to a few hundred Hz, as indicated by studies using amplitude-modulated noise (e.g., Burns and Viemeister, 1976; Burns and Viemeister, 1981).

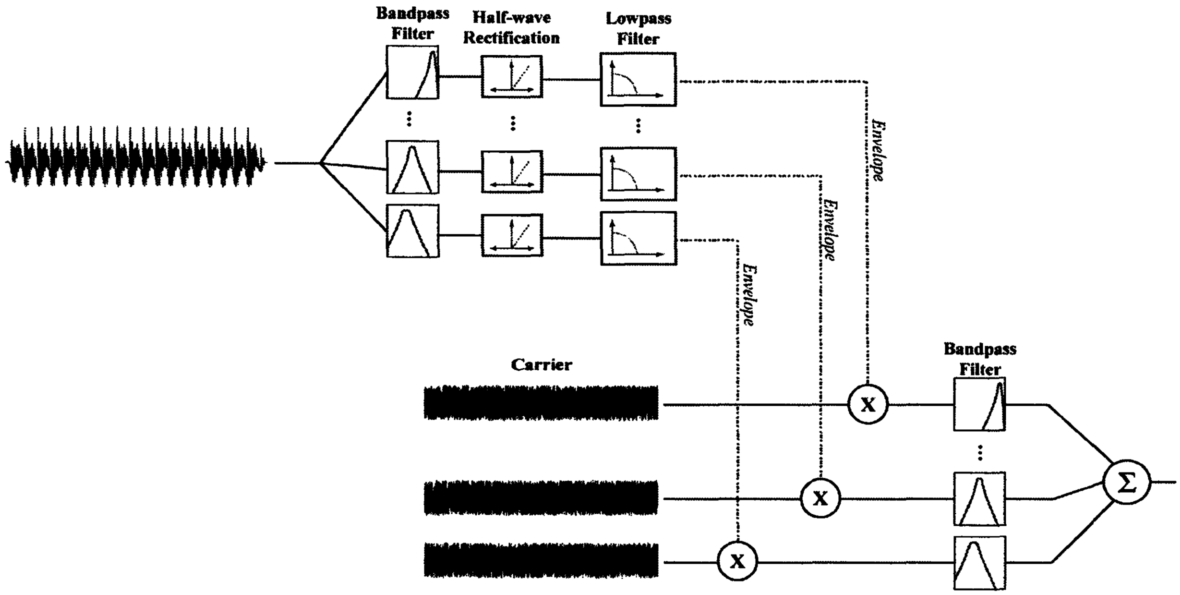

Figure 1-3: Schematic diagram of the noise-excited vocoder used to simulate the effects of implant sound processing. The unprocessed stimuli are first bandpass filtered into N contiguous frequency channels. The envelopes of the signals in each channel are extracted by half-wave rectification and lowpass filtering. The envelopes are then used to amplitude modulate independent white-noise carriers. The same bandpass filters that were used to filter the original stimuli are then used to filter the AM noises. Finally, the modulated narrowband noises were summed. The acoustic vocoder parallels implant signal processing in two important ways: 1) the signal spectrum is poorly resolved by the coarse bandpass filtering; and 2) the lowpass filtering eliminates any temporal fine-structure cues in the original stimuli.

Using normal-hearing listeners and noise-excited vocoders to study the effects of implant processing and hearing impairment has several advantages: 1) it allows for more control over the experimental variables; 2) it minimizes the between-subject variability associated with actual impaired listeners and implant users; and 3) it allows experimenters to examine the benefits of potential processing schemes beyond the limits of current implant technology. While studies with normal-hearing listeners and vocoder simulators have their advantages, interpretations of the results in terms of applicability to implant users must be approached with caution. Although the higher-order processing may be similar for implant users and normal-hearing listeners, auditory-nerve responses elicited by acoustic stimulation are inherently different from those elicited by electrical stimulation (Throckmorton and Collins, 2002). Therefore, results from acoustic vocoder simulation studies should be interpreted in terms of potential trends rather than quantitative estimates of implant user performance.

1.5 Thesis overview

The aim of this dissertation is to examine the effects of envelope-vocoder processing on pitch perception, and the consequences that this has on speech perception in the presence of complex background interference. While noise-excited vocoders have been used to examine the intelligibility of speech in isolation and in noise, the effects of more complex maskers (e.g. fluctuating interference and competing speech) have yet to be studied. If the intelligibility of envelope-vocoded speech depends on envelope fluctuations, then it is possible that adding spurious fluctuating maskers could be devastating to performance. In this dissertation, we shall examine the effects of envelope-vocoding on the perception of voice pitch, specifically the role of pitch as a segregation cue. We speculate that envelope-periodicity pitch, as encoded in current cochlear implants, will not provide a sufficiently salient cue to be utilized in the segregation of target speech from background interference.

In chapter 2, the effects of envelope-vocoding on speech reception in complex situations are examined. The experiments involved measuring the intelligibility of sentences in backgrounds ranging from steady-state noise to a single-talker interferer. Overall, the results were as expected. Higher channel numbers led to better performance. However, even with 24 channels, where the filter bandwidths approached those of normal-hearing listeners, performance was considerably poorer than in the unprocessed conditions. In particular, the advantage of fluctuating maskers (relative to steady-state maskers) found in normal hearing became a disadvantage under envelope-vocoder processing. The results may be due to the inability of envelope-vocoder processing to convey the temporal and/or spectral fine structure associated with low-order resolved harmonics.

In chapter 3, the effects of envelope-vocoding on F0 discriminability and utility are examined. As stated above, for normal-hearing listeners, the ability to discriminate F0 differences of complex sounds is thought to be dominated by the resolved lower-order harmonics. The limited spectral resolution of envelope-vocoding means that implant users must rely on the perceptually weaker envelope-periodicity cue for pitch. This suggests that implant users may be less sensitive to F0 differences than normal-hearing listeners. Furthermore, reverberation and competing speech have a smearing effect on the temporal envelope (Houtgast *et al.*, 1980). Given that implant users rely on temporal modulation to extract pitch, we suggest that reverberation and competing speech are likely be more detrimental to F0 discrimination for envelope-vocoder listeners. The experiments described in this chapter involved measuring F0 difference limens (F0DLs) for complex tones with and without added reverberation, as well as measuring the identification accuracy of concurrent vowels as a function of the F0 difference between constituent vowels. The results were consistent with our hypothesis. Reverberation was found to be detrimental to F0 discrimination in conditions with fewer vocoder channels. However, despite the F0DLs being less than 1 semitone with 24- and 8-channel vocoder processing in quiet, listeners were unable to benefit from F0 differences between the competing vowels in a concurrent-vowel paradigm.

In Chapter 4, we speculate on how pitch perception might be improved in implant users. As cochlear implant technology matures, the criteria for implant candidacy have become more lenient. Now patients with some residual hearing are regarded as good candidates for a cochlear implant. If robust pitch cues are carried by resolved low-frequency components of sounds, then augmenting existing cochlear-implant processing with low-frequency residual hearing (< 300 Hz or < 600 Hz) may help mitigate the susceptibility of cochlear-implant users to complex interference. The experiments described in this chapter were designed to examine whether adding well-resolved lower-order harmonics to envelope-vocoder processed speech could improve the utility of voice pitch for speech reception in complex auditory environments.

In chapter 5, a general summary is presented. This chapter summarizes the contributions of this dissertation to the general understanding of the importance of F0 information for speech perception and makes recommendations for future research.

Chapter 2: Effects of envelope-vocoder processing on speech reception in the presence of interference[1]

2.0 Abstract

This study investigated the effects of envelope-vocoder processing on speech reception in a variety of complex masking situations. Speech recognition was measured as a function of target-to-masker ratio, processing condition (4, 8, 24 channels and unprocessed) and masker type (speech-shaped noise, amplitude modulated speech-shaped noise, single male talker, and single female talker). The results showed that envelope-vocoder processing was more detrimental to speech reception in fluctuating interference than in steady-state noise. Performance in the 24-channel processing condition was substantially poorer than in the unprocessed condition, despite the comparable representation of the spectral envelope. The detrimental effects of envelope-vocoder processing in fluctuating maskers, even with large numbers of channels, may be due to the reduction in the pitch cues used in sound source segregation, which are normally carried by the peripherally resolved low-frequency harmonics and the temporal fine structure. The results suggest that using steady-state noise to test speech intelligibility may underestimate the difficulties experienced by cochlear-implant users in fluctuating acoustic backgrounds.

2.1 Introduction

Speech has been shown to be a very robust medium for communicating information (Fletcher and Galt, 1950; Miller and Licklider, 1950; Remez *et al.*, 1994; Stevens, 1998).

[1] A version of this chapter was published as Qin and Oxenham (2003).

Although the precise mechanisms underlying the apparent resilience to interference and distortion are still not well understood, the ability of speech to convey information under adverse conditions is generally attributed to the layers of acoustic, phonetic, and linguistic redundancies. Shannon *et al.* (1995), using a noise-excited vocoder, provided a dramatic demonstration of these redundancies at work. They found that despite a severe reduction in spectral cues and the elimination of temporal fine-structure information, sentences presented in the absence of interfering sounds could be recognized with as few as four frequency bands. Subsequent studies have shown that while more frequency bands are needed for speech reception in steady-state noise, good sentence recognition is still possible at relatively low signal-to-noise ratios (e.g., Dorman *et al.*, 1998).

The processing schemes used in these studies are designed to simulate the effects of cochlear-implant stimulation (Wilson *et al.*, 1991). They can therefore be used to provide insights into the relative efficacy of different processing algorithms without using valuable implantee testing time (Blamey *et al.*, 1984). Indeed, at least for low numbers of frequency bands, acoustic simulations of cochlear-implant processing using normal-hearing listeners have yielded results that are reasonably comparable to those of actual implant patients (Friesen *et al.*, 2001; Carlyon *et al.*, 2002). Another use for such schemes is to probe the acoustic features necessary for speech reception in normal-hearing listeners. A number of studies indicate that important information is carried in the envelopes of the stimulus after filtering into frequency sub-bands (Houtgast *et al.*, 1980; Drullman, 1995; Smith *et al.*, 2002). From the results obtained so far, it may be concluded that speech reception requires minimal frequency selectivity and no temporal fine-structure information. This conclusion seems at odds with the experiences of many impaired-hearing listeners.

While hearing-impaired listeners often perform well in quiet conditions (when audibility is corrected for with amplification), many experience great difficulty in noisy conditions. The difference in performance between normal-hearing and hearing-impaired listeners is especially pronounced in temporally fluctuating maskers and maskers with spectral notches (Gustafsson and Arlinger, 1994). In particular, while normal-hearing listeners show large improvements in speech reception when spectral and/or temporal fluctuations are introduced into a masker, hearing-impaired listeners often show much less benefit (Festen and Plomp, 1990; Peters *et al.*, 1998). It is thought that normal-hearing listeners are able to make use of the improved local target-to-masker ratio in the masker's spectral and temporal dips. In contrast, hearing-impaired listeners, with their poorer frequency selectivity (Patterson *et al.*, 1982; Glasberg and Moore, 1986) and poorer effective temporal resolution (Glasberg and Moore, 1992; Oxenham and Moore, 1997), may be less able to benefit from the improved local target-to-masker ratio found in the spectral and temporal dips of the masker.

In the case of cochlear implants and implant simulations, the finding that better frequency resolution (i.e., a greater number of frequency bands) is required for speech reception in noise than in quiet (Dorman *et al.*, 1998; Fu *et al.*, 1998a) parallels the finding that spectral smearing is more detrimental to speech reception in noise than in quiet (ter Keurs *et al.*, 1992; Baer and Moore, 1993). It is also consistent with the hypothesized effect of poorer frequency selectivity in impaired-hearing listeners. The perceptual effect of eliminating the temporal fine structure in cochlear-implant simulations is less clear. Pitch perception and the ability to discriminate different fundamental frequencies (F0s), is thought to rely primarily on fine-structure information, in particular the information carried in peripherally resolved, lower-order harmonics (e.g., Plomp, 1967; Houtsma and Smurzynski, 1990; Smith *et al.*, 2002). While the envelopes of implant-processed stimuli carry some periodicity information, the pitch salience associated with such envelope periodicity is rather weak (Burns and Viemeister, 1976; 1981; Shackleton and Carlyon, 1994).

Fundamental frequency information has long been thought to play an important role in perceptually segregating simultaneous and non-simultaneous sources (Brokx and Nooteboom, 1982; Assmann and Summerfield, 1990; 1994; see Darwin and Carlyon, 1995 for a review; Bird and Darwin, 1998; Vliegen and Oxenham, 1999). A reduction in F0 cues produced by cochlear-implant processing may lead to greater difficulties in segregating different sources. If the perception of implant-processed speech is based on envelope fluctuations, as suggested above, then listeners must accurately distinguish the envelope fluctuations of the target from those of the masker. Similarly, a listener can only take advantage of spectral and temporal dips in the masker if the listener can accurately identify the presence of the dips.

The aim of the present study was to investigate the effects of fluctuating maskers on the reception of envelope-vocoder processed speech. We hypothesized that the reduction in

F0 cues produced by the implant simulations would particularly affect conditions where the ability to discriminate the target from the masker is thought to play an important role in determining speech reception thresholds (e.g., speech in the presence of competing talkers or fluctuating backgrounds). Speech reception was measured in normal-hearing listeners as a function of target-to-masker ratio, processing condition (4, 8, or 24 channels, or unprocessed) and masker type (steady-state speech-shaped noise, speech-shaped noise modulated with a speech envelope, single male talker, and single female talker).

2.2 Methods

2.2.1 Participants

Thirty-two normal-hearing listeners (fifteen females) with audiometric thresholds of less than 20 dB HL at octave frequencies between 125 and 8000 Hz, participated in this study. Their ages ranged from 18 to 46 (median age 22). They were all native speakers of American English.

2.2.2 Stimuli

All stimuli in this study were composed of a target sentence presented in the presence of a masker. The stimulus tokens were processed prior to each experiment. The targets and maskers were combined at the desired target-to-masker ratios (TMRs) prior to any processing. TMRs were computed based on the token-length root-mean-square (RMS) amplitudes of the signals. Maskers were gated on and off with 250-ms raised-cosine ramps 250 ms prior to and 250 ms after the end of each target sentence.

The targets were H.I.N.T. sentences (Nilsson *et al.*, 1994) spoken by a male talker. The H.I.N.T sentence corpus consists of 260 phonetically balanced high-context sentences of easy-to-moderate difficulty. Each sentence is composed of four to seven keywords.

Since differences in the F0 of voicing are thought to contribute to speaker segregation (Brokx and Nooteboom, 1982; Assmann and Summerfield, 1990; 1994; Darwin and Carlyon, 1995; Bird and Darwin, 1998), we chose a male single-talker masker with a mean F0 (111.4 Hz) similar to that of the target talker (110.8 Hz) and a female single-talker masker with a mean F0 (129.4 Hz) almost 3 semitones higher. The motivation for using different gender single-talker interferers came from the observation that normal-hearing listeners benefit from F0 differences between target and interfering talkers (Brokx and Nooteboom, 1982; Assmann and Summerfield, 1990; 1994; Bird and Darwin, 1998). Talker F0s were estimated using the YIN program provided by de Cheveigné and Kawahara (2002). The male single-talker maskers were excerpts from the audio book "Timeline" (novel by M. Crichton) read by Stephen Lang. The female single-talker maskers were excerpts from the audio book "Violin" (novel by A. Rice) read by Maria Tucci. To avoid long silent intervals in the masking speech, such as sentence-level pauses, both single-talker maskers were automatically preprocessed to remove silent intervals greater than 100 ms. The maskers were then subdivided into non-overlapping segments to be presented at each trial.

The single-talker maskers and speech-shaped-noise masker were spectrally shaped to match the long-term power spectrum of the H.I.N.T. sentences. The amplitude-modulated speech-shaped noise masker was generated by amplitude modulating the steady-state speech-shaped noise with the broadband speech envelope of the male single-talker masker (lowpass filtered at 50 Hz; 1st-order Butterworth filter).

For a given listener, the target sentence lists were chosen at random, without replacement, from among the 25 lists of H.I.N.T. sentences. This was done to ensure that no target sentence was presented more than once to any given listener. Data were collected using one list (i.e., 10 sentences) for each TMR.

2.2.3 Stimulus processing

All stimulus tokens were processed prior to each experiment. The cochlear-implant simulator was implemented using Matlab (Mathworks, Natick MA) in the following manner. The stimuli (target plus masker) were first bandpass filtered (6th order Butterworth filters) into 4, 8, or 24 contiguous frequency bands (or channels) between 80 and 6000 Hz. The entire frequency range was divided equally in terms of the Cam scale[2] (Glasberg and Moore, 1990). The 3-dB channel bandwidths were approximately 6.98 Cams, 3.49 Cams, and 1.16 Cams for the 4-, 8-, and 24-channel conditions, respectively. The envelopes of the signals were extracted by half-wave rectification and lowpass

[2] This is more frequently referred to as the ERB scale. However, as pointed out by Hartmann (1997), ERB simply refers to equivalent rectangular bandwidth, which could be used to define all estimates of auditory filter bandwidths. We, therefore, follow Hartmann's convention of referring to the scale proposed by Glasberg and Moore as the Cam scale, in recognition of its origins in the Cambridge laboratories. Described in Glasberg and Moore (1990), $Cam = 21.4\log_{10}(0.00437f + 1)$, where f is frequency in Hz.

filtering (using a 2nd-order Butterworth filter) at 300 Hz, or half the bandpass filter bandwidth, whichever was lower. The 300-Hz cutoff frequency was chosen to preserve as far as possible F0 cues in the envelope. The envelopes were then used to modulate narrowband noises, filtered by the same bandpass filters that were used to filter the original stimuli. Finally, the modulated narrowband noises were summed and scaled to have the same level as the original stimuli.

2.2.4 Procedure

The 32 listeners were divided into four groups of eight. Each group was tested on only one of the four processing conditions (i.e. 4, 8, 24 channels, or unprocessed). The speech reception of each listener was measured in the presence of all four masker types (single male and female talkers, modulated and steady-state speech-shaped noise), at six TMRs (see Table 2-1). The TMRs for each processing condition and masker type were determined in an earlier pilot study, using two to three listeners. The TMRs for each experimental condition were set to avoid floor and ceiling effects in the psychometric function.

The target and masker were combined at the appropriate TMR, processed, and stored on disk prior to the experiments. The processed stimuli were converted to the analog domain using a soundcard (LynxStudio, LynxOne) at 16-bit resolution with a sampling rate of 22050 Hz. The stimuli were then passed through a headphone buffer (TDT HB6) and presented diotically at 60 dB SPL via Sennheiser HD580 headphones to the listener seated in a double-walled sound-insulation booth. Listeners typed their responses into a computer via the keyboard.

Table 2-1: The values in the table represent the minimum, maximum, and step size of the Target-to-masker ratios (in dB). The step sizes are in parentheses.

Processing condition	Masker type	Target-to-masker ratio (dB)
Unprocessed	Male interference	-20 to 5 (5)
	Female interference	-20 to 5 (5)
	Modulated noise	-25 to 0 (5)
	Steady-state noise	-15 to 0 (3)
24 channels	Male interference	-15 to 10 (5)
	Female interference	-15 to 10 (5)
	Modulated noise	-20 to 5 (5)
	Steady-state noise	-10 to 10 (4)
8 channels	Male interference	-5 to 20 (5)
	Female interference	-5 to 20 (5)
	Modulated noise	-10 to 15 (5)
	Steady-state noise	-5 to 20 (5)
4 channels	Male interference	5 to 30 (5)
	Female interference	5 to 30 (5)
	Modulated noise	5 to 30 (5)
	Steady-state noise	5 to 30 (5)

For practice, the listeners were presented with 20 stimuli, five from each of the four masking conditions. In each practice masking condition, the target sentences were presented at four different TMRs. The target sentences used in the practice session were from the Harvard-Sentence database (IEEE, 1969). The practice sessions were designed to acclimate the listeners to the processed stimuli. Feedback was given during the practice sessions, but not during the experimental sessions.

2.2.5 Analysis

Listener responses were scored offline by the experimenter. Each listener's responses for a given TMR, under a given masker condition, were grouped together to produce a percent correct score. Keywords were used to calculate the percent correct. Obvious misspellings of the correct word were considered correct.

2.3 Results

2.3.1 Fits to the psychometric functions

The percent correct scores as a function of TMR under a given masker condition for each listener were fitted to a two-parameter sigmoid model (a cumulative Gaussian function):

$$\text{Percent Correct} = \frac{100}{\sqrt{2\pi\sigma}} \int_{-\infty}^{\text{TMR}} \exp\left(\frac{-(\text{x}-\text{SRT})^2}{2\sigma^2}\right) dx \qquad \text{(Eq. 1-1)}$$

,

where x is the integration variable, SRT is the speech-reception-threshold[3] (dB), σ is related to the slope of the function, and TMR is the target-to-masker ratio (dB).

Fig. 2-1 shows sample data from one listener, along with the best-fitting curve (heavy) according to Eq. (1-1). The other, lighter curves in the figure are the fits to the data from the other seven listeners in that experimental condition. The two-parameter model assumes that listeners' peak reception performance is 100%. This assumption may be valid for the 24- and 8-channel conditions, but it is probably not valid for the 4-channel condition. Therefore, the initial model had a third parameter, associated with the peak

[3] Speech-reception-threshold is the target-to-masker ratio (dB TMR) at which 50% of the words were correctly identified (see Fig. 2-1).

performance. However, the goodness of fit and the estimated SRTs of the three-parameter model were very similar to those of the two-parameter model, leading us to select the model with fewer parameters.

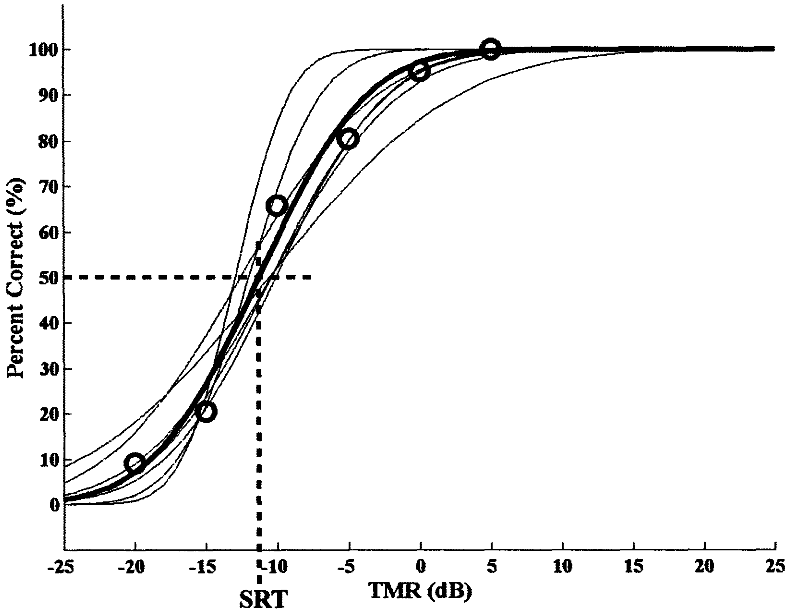

Figure 2-1: An example of the two-parameter sigmoid model fitting procedure. The two-parameter sigmoid model (heavy line) is fitted to the speech reception performance data of an individual listener (open circles). The light lines are the functions fitted to the data of the other listeners in the same experimental condition (data not shown). The speech reception threshold (SRT) is the target-to-masker ratio (TMR) where 50% of the words were correctly identified.

The two-parameter model provided generally good fits to the curves of performance as a function of TMR. Presented in Table 2-2 are the mean values of SRT, σ, and standard error of fit, averaged across listeners. The standard deviations are shown in parentheses. The individual standard errors of fit[4] had a mean of 7.25% with a standard deviation of

[4] The standard error of fit is the square root of the summed square of error divided by the residual degrees of freedom, $\sqrt{\frac{SSE}{v}}$, where $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$. The residual degrees-of-freedom term (v) is defined as the number of response values (n) minus the number of fitted coefficients (m) estimated for the response values, $v = n - m$.

3.7% (median of 7.01% and a worst case of 20.93%). Combined according to experimental conditions, the average standard errors of fit (Table 2-2) were generally less than 10%.

Table 2-2: Mean sigmoidal model parameter values (Eq. 2.1), averaged across listeners. The standard deviations are in parentheses. The standard error of fit provides a numerical indicator for how well the model fits the data. SRT is the speech reception threshold, and σ is related to the slope of the function.

Processing condition	**Masker type**	**SRT (dB TMR)**	**σ**	**Standard error of fit (%)**
Unprocessed	Male interference	-10.3 (2.4)	7.5 (1.9)	8.3 (3.5)
	Female interference	-11.3 (1.1)	6.3 (2.3)	6.8 (3.2)
	Modulated noise	-9.1 (0.6)	4.1 (1.5)	5.6 (2.3)
	Steady-state noise	-6.7 (0.8)	3.4 (0.7)	5.5 (3.4)
24 channels	Male interference	0.7 (1.8)	5.1 (1.1)	6.4 (3.4)
	Female interference	0.6 (1.8)	5.0 (1.0)	4.8 (2.1)
	Modulated noise	-3.3 (0.6)	5.0 (1.1)	4.0 (2.3)
	Steady-state noise	-1.2 (1.2)	3.3 (0.6)	3.8 (1.8)
8 channels	Male interference	6.4 (2.2)	6.0 (2.0)	9.1 (2.0)
	Female interference	6.7 (1.5)	5.2 (1.4)	7.2 (1.7)
	Modulated noise	4.6 (2.1)	7.7 (1.5)	6.8 (3.4)
	Steady-state noise	4.2 (0.8)	4.1 (1.2)	5.4 (1.9)
4 channels	Male interference	18.1 (3.1)	14.1 (1.8)	9.9 (3.1)
	Female interference	18.3 (4.3)	15.3 (3.3)	12.5 (3.8)
	Modulated noise	15.6 (4.4)	18.7 (4.8)	9.3 (2.2)
	Steady-state noise	14.9 (5.4)	19.3 (9.2)	10.6 (3.4)

2.3.2 Speech-reception thresholds

In general, performance across listeners was reasonably consistent, so only the mean SRT values as a function of masker condition and processing condition are plotted in Fig. 2-2. The mean SRT values and standard errors of means (Table 2-2) were derived from the

SRT values of individual model fits. Since an SRT value is the TMR where 50% of the keywords are correctly identified, a higher SRT value implies a condition more detrimental to speech reception.

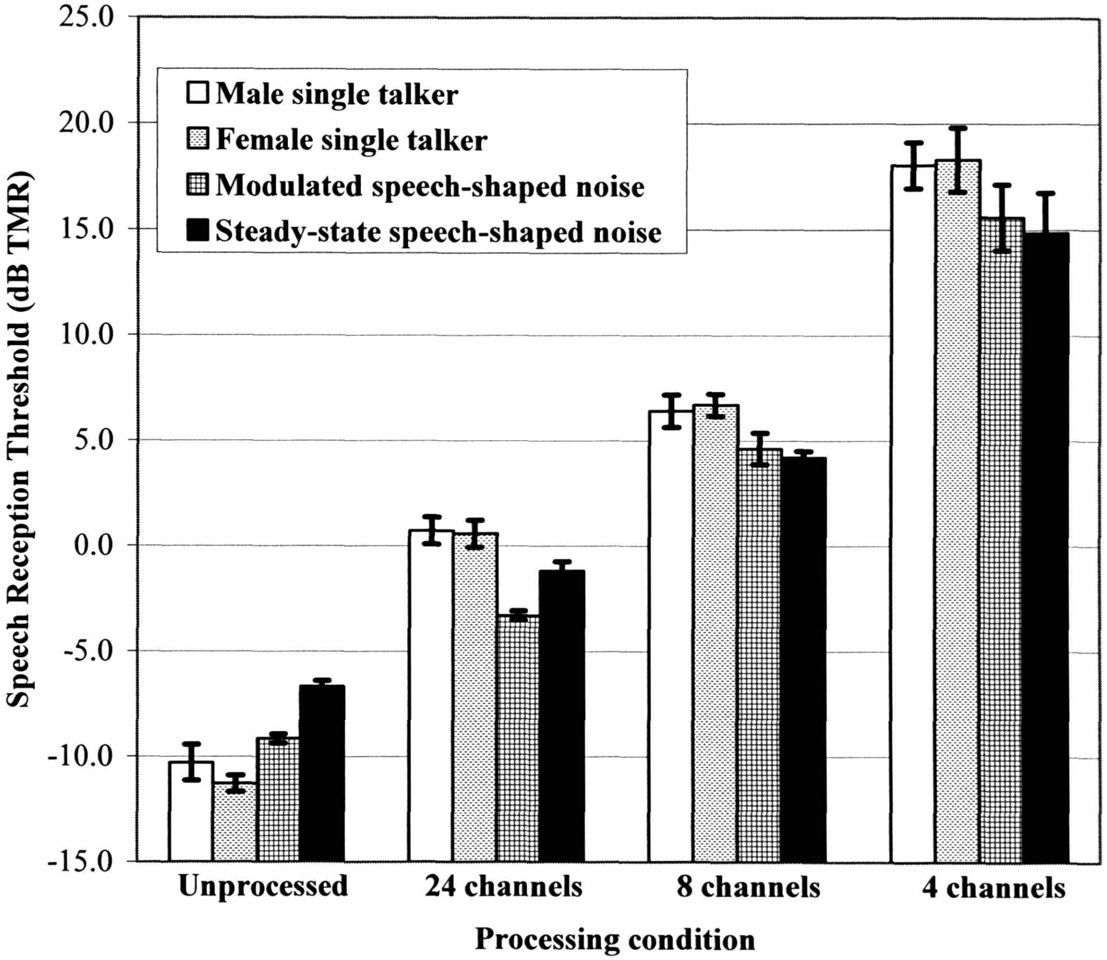

Figure 2-2: Speech reception threshold, in terms of target-to-masker ratio, as a function of processing condition in the presence of a male single talker (unshaded), a female single talker (dotted), modulated noise (grid), and steady-state noise (solid). The plotted values and their respective standard deviations can be found in Table 2-2.

Fig. 2-2 shows that SRT values in all masker conditions were strongly affected by implant processing. As the number of spectral channels deceased, the SRT values under all masker types increased. However, the rate of increase differed between masker types. As the number of spectral channels decreased, SRT values increased faster in the presence of fluctuating maskers than in the presence of steady-state noise, particularly for the single-talker maskers.

A two-way mixed-design analysis of variance (ANOVA) was performed using Statistica (StatSoft, Tulsa OK) to determine the statistical significance of the findings, with SRT as the dependent variable, processing condition as the between-subjects factor, and masker type as the within-subjects factor. The ANOVA indicated that both main factors and their interaction were statistically significant (processing condition: $F_{3,28} = 218.1$; masker type: $F_{3,84} = 7.5$; interaction: $F_{9,84} = 8.7$; $p < 0.001$ in all cases). A *post hoc* test according to Fisher's LSD (alpha = 0.05) indicated several significant differences between the different experimental conditions, as outlined below.

In the unprocessed conditions, the steady-state noise masker was significantly more effective than any of the modulated maskers. However, under implant processing the reverse was true, with the exception of the 24-channel processed modulated speech-shaped noise condition. These differential effects are illustrated in Fig. 2-3, which treats the steady-state masker as the baseline condition and plots the differences in SRT values between the steady-state noise and the other maskers as a function of processing condition. Significant differences between SRTs in steady-state noise and those in the other conditions are labeled with asterisks in Fig. 2-3.

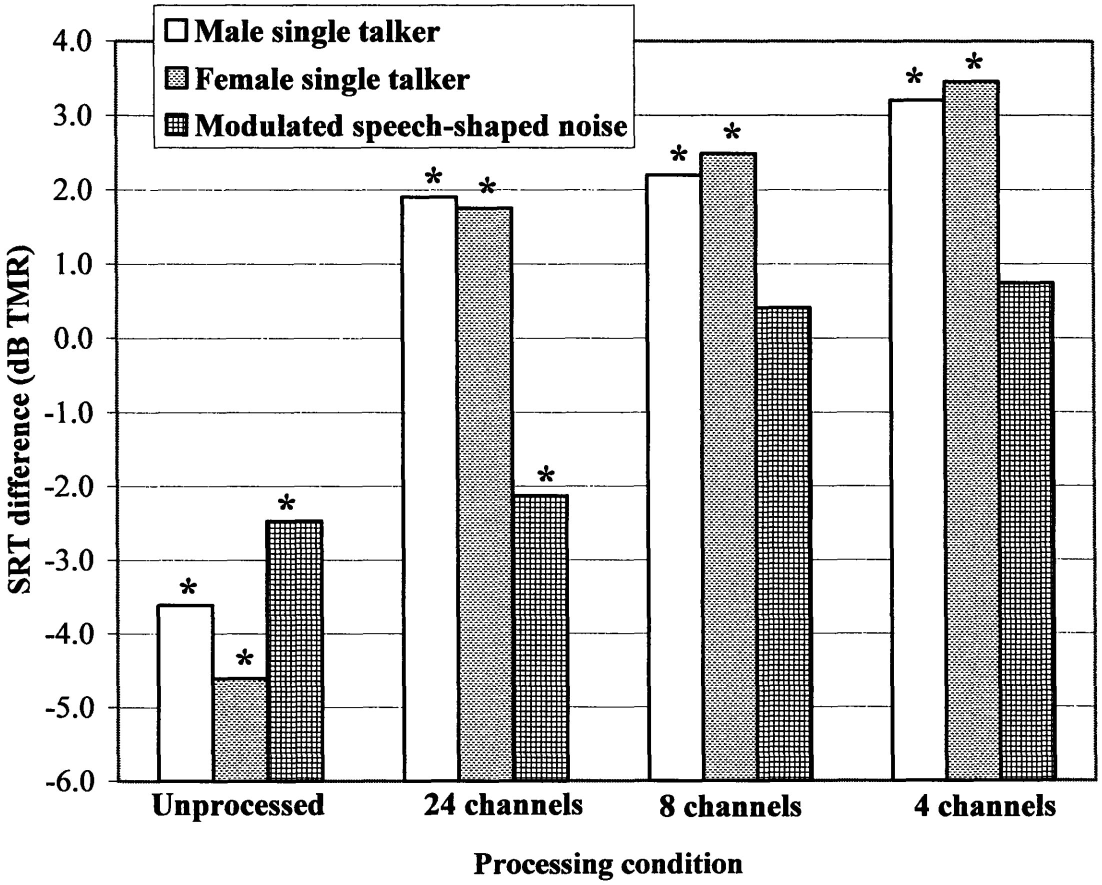

Figure 2-3: SRT differences between the steady-state noise masker and the male single-talker (unshaded), female single-talker (dotted), and modulated noise (grid) maskers are shown as a function of processing condition. Masked thresholds significantly different from those in the steady-state noise, according to Fisher's LSD test (alpha = 0.05), are labeled by an asterisk.

The single-talker interferers produced significantly higher SRT values than steady-state noise in all processed conditions (i.e. 24, 8, and 4 channels). In contrast, the modulated noise produced lower SRT values than the steady-state noise in the 24-channel condition, and was not significantly different from the steady-state noise in the 8- and 4-channel conditions. As illustrated in Fig. 2-2, in all conditions the transition from unprocessed to 24-channel processing resulted in a large increase in SRT value, despite the fact that the

24-channel condition represents frequency resolution approaching that found in normal-hearing listeners. This finding is explored further in the discussion section (section 2.4.3).

There was no significant difference in the SRT values between the male and female single-talker maskers in any processing condition (Fig. 2-2). Given our hypothesized effect of F0 differences in source segregation, this may seem unexpected. This finding is explored further in the discussion section.

2.4 Discussion

2.4.1 Single-talker interference vs. steady-state noise

Our results in the unprocessed conditions are consistent with previous studies in showing that SRT values are lower for single-talker interferers than for steady-state noise (e.g., Festen and Plomp, 1990; Peissig and Kollmeier, 1997; Peters *et al.*, 1998). The improved performance found with single-talker interference, relative to steady-state noise, has been ascribed to listeners' ability to gain information from temporal or spectral minima in the maskers. However, to make use of local masker minima, the listener must have cues to distinguish the target from masker. Voice F0 is a generally accepted segregation cue for normal-hearing listeners (Brokx and Nooteboom, 1982; Bird and Darwin, 1998; Freyman *et al.*, 1999; Brungart, 2001). Our hypothesis was that the reduction in F0 cues, produced by envelope-vocoder processing, would particularly affect speech reception where the ability to discriminate the target from the masker is thought to play an important role. The results from the processed conditions are consistent with the hypothesis: not only are the benefits of spectral and temporal masker dips unseen, but the single-talker interferers go from being the least effective maskers in the unprocessed conditions to being the most effective maskers in all the processed conditions (see Figs. 2-2 and 2-3).

2.4.2 Modulated noise vs. steady-state noise

If envelope-vocoder processing led to a global inability to use temporal minima in fluctuating maskers, the same deterioration in performance would be expected in the modulated-noise masking conditions as was found for the single-talker interferers. In fact, the difference in performance between the modulated-noise and the steady-state-noise conditions remains roughly constant for the unprocessed and 24-channel conditions. For 8 and 4 channels, SRT values in the modulated-noise conditions are not significantly higher than in the steady-state noise conditions.

Without F0 cues, listeners may still maintain high levels of speech reception in the presence of interference by utilizing different cues. For example, when speech is presented in the presence of steady-state noise, a listener may be able to use slow-varying envelope modulation as a cue for segregating the target from the noise, as most slow-varying envelope modulations will belong to the target. In the case of the amplitude-modulated speech-shaped noise masker, the noise is always modulated coherently across frequency. Speech, on the other hand, does not always modulate coherently across all frequencies. Listeners could use the more consistent comodulation of the amplitude-modulated noise as a cue for source segregation. However, to use comodulation as a segregation cue, spectral resolution must be sufficiently fine to distinguish the time-varying spectral changes of the target speech from the comodulated noise masker. This could account for the SRT difference between modulated speech-shaped noise and steady-state noise in the unprocessed and 24-channel processing conditions. If the spectral resolution is too coarse (e.g., in the 8- and 4-channel conditions) the stimulus representation of target speech will also exhibit very strong comodulation. This may eliminate differences in comodulation as a valid cue for distinguishing between masker and target, and may account for the lack of an SRT difference between the modulated speech-shaped noise and steady-state noise in the 8- and 4-channel processing conditions.

2.4.3 Unprocessed vs. 24-channel processing

As shown in Fig. 2-2, performance with 24-channel processing was considerably worse than with no processing, for all masker types. This may seem surprising, given that the spectral resolution in the 24-channel processing condition was chosen to be similar to that found for normal-hearing listeners, with 3-dB filter bandwidths of 1.16 Cams. In Fig. 2-4, the excitation patterns (Moore *et al.*, 1997) for the vowel */a/* with and without 24-channel processing are plotted. It can be seen that the spectral peaks of the vowel are comparably well represented in both the processed (dashed) and unprocessed (solid) conditions. This may suggest that the temporal fine structure, discarded by the processing, while not necessary for speech recognition in quiet, may play an important role in segregating speech from interfering sounds. Similarly, it can be seen from Fig. 2-4 that the spectral resolution of the first few harmonics (below 500 Hz) is degraded in the 24-channel processing condition. While that information may not be important for speech reception *per se*, the first few harmonics carry important information about the stimulus F0. It is therefore possible that the loss of F0 information, due to a reduction in harmonic resolution and/or a loss of temporal fine-structure information, is responsible for the large difference in performance between the unprocessed and 24-channel conditions.

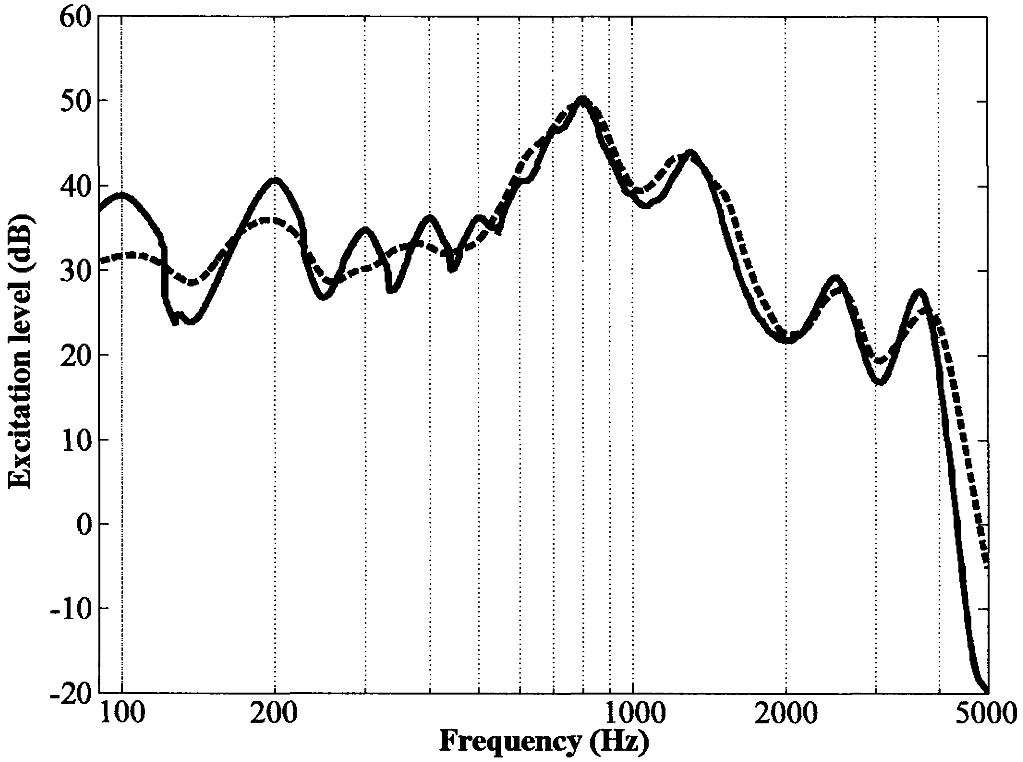

Figure 2-4: An illustration of the difference in effective spectral resolution between unprocessed (solid) and 24-channel processed (dashed). This figure is the output of the excitation pattern model (Moore et al., 1997) in response to a 500-ms Klatt synthesized vowel /a/, with a F0 at 100 Hz (Klatt, 1980).

2.4.4 Male vs. female single-talker interference

As mentioned in the methods section, the motivation for using different gender single-talker interferers came from the observation that normal-hearing listeners benefit from F0 differences between target and interfering talkers. This benefit generally increases with increases in F0 differences (Assmann and Summerfield, 1990; 1994; Bird and Darwin, 1998). Our finding of no significant difference between the male and female interferers may therefore seem surprising. There are at least three possible explanations for this null effect.

The first possible explanation lies in the instantaneous F0 values. In many past studies of single-talker interference (Brokx and Nooteboom, 1982; Assmann and Summerfield, 1990; 1994; Bird and Darwin, 1998), the F0s of the targets and maskers were held constant, either through the use of short-duration stimuli or synthesized speech with a fixed F0. In the present study, the single-talker interferers were taken from recorded books, where exaggerated prosody is common. Although the mean F0 of the male single-talker interference (111.4 Hz) was approximately equal to that of the target (110.8 Hz) and the mean F0 of the female single-talker interference (129.4 Hz) was about three semitones higher, the natural variations in F0 were left unaltered. As a result, the F0 differences between the target and single-talker interference were distributed such that the probability[5] of a two-semitone difference in F0 between the target and male single-talker interference was 0.69, and between the target and the female single-talker interference was 0.76. The lack of difference in SRT values between the male and female masker may therefore be due to the large differences in instantaneous F0 between the target and both maskers. However, contrary to the hypothesis, previous studies showed little or no improvements in identification as a result of time-varying F0s, as compared to constant F0s (Darwin and Culling, 1990; Summerfield and Culling, 1992; Assmann, 1999). Their findings suggest that the instantaneous difference in F0 between the

[5] The probability of a 2-semitone difference in F0 was computed by integrating the F0 joint probability distribution function of the target and male or female single-talker interference.

competing talkers in this study may not be the main factor behind the lack of difference between the two single-talker maskers.

The second possible explanation lies in the atypical F0 range of the female voice used in this experiment. Adult female voices have an average F0 of around 220 Hz (Hillenbrand *et al.*, 1995). The finding of no significant difference between the male and female interferers in this study may be due to unusually low mean F0 (129.4 Hz) of the female interferer. The lack of a gender effect in this study may, therefore, not generalize to everyday situations.

The third possible explanation lies in the individual vocal characteristics of the talkers (e.g. vocal tract length, accent, speaking style, sentence level stress etc.) The differences in vocal characteristics between the target and interfering talkers may have been sufficiently large to render any further improvement due to mean F0 difference negligible.

2.4.5 Importance of frequency selectivity and temporal fine structure

Speech is an ecologically important stimulus for humans. If speech reception in quiet can be achieved with minimal spectral resolution and no temporal fine-structure information, then is the exquisite frequency selectivity and sensitivity to fine structure of the human auditory system necessary for speech communication? One important function of frequency selectivity may be found in earlier studies of spectral smearing (ter Keurs *et al.*, 1992; Baer and Moore, 1993) and of cochlear-implant simulations in steady-state noise (Dorman *et al.*, 1998; Fu *et al.*, 1998a). From these studies, and from the present study, it can be seen that greater frequency selectivity results in lower signal-to-noise ratios necessary for speech reception. The present results suggest that sensitivity to low-frequency temporal fine structure (Meddis and O'Mard, 1997), and/or the spectral resolution of the lower harmonics (Terhardt, 1974) is critical to good speech reception in complex backgrounds. Performance is greatly affected by envelope-vocoder processing, even with 24-channel resolution. The effect of stimulus processing is especially dramatic for the single-talker interferers, where an SRT benefit with respect to steady-state noise in the unprocessed condition is transformed into a deficit in all processed conditions. We hypothesize that this dramatic deterioration in performance is related to an inability to perceptually segregate the target from the masker, and that successful segregation relies on good frequency selectivity and F0 sensitivity.

The present results have possible implications for cochlear-implant design. As in previous studies (Dorman *et al.*, 1998; Fu *et al.*, 1998a), the results support separating the spectrum into as many channels as possible, given the technical constraints of ensuring channel independence (Friesen *et al.*, 2001). However, the results also suggest that simply increasing the number of channels (at least to 24) may not assist in providing sufficient F0 information to successfully segregate a target from an acoustically complex background. The problem of presenting usable fine-structure information to implant users is current topic of research (e.g., Litvak *et al.*, 2001), and the present results provide further support for such endeavors. Finally, the large differences between performance in steady-state noise and performance in fluctuating backgrounds, particularly single-talker interferers, suggest that testing cochlear-implant patients in steady-state noise alone may underestimate the difficulties faced by such listeners in everyday acoustic environments.

2.5 Summary

- Envelope-vocoder processing leads to a large deterioration in speech reception in the presence of a masker, even when the spectral resolution approaches that of normal hearing.
- Under envelope-vocoder processing, single-talker interference is more detrimental to speech reception than steady-state noise. This is the converse of the situation found without processing, and it highlights the potential importance of frequency selectivity and temporal fine-structure information in segregating complex acoustic sources.
- In the presence of steady-state noise, the amplitude modulations associated with the target speech may provide useful source segregation cues, even under simulated implant processing, provided that the spectral resolution is sufficiently fine.
- Using steady-state noise to test speech intelligibility may underestimate the difficulties experienced by cochlear-implant patients in everyday acoustic backgrounds.

Chapter 3: Effects of vocoder processing on F0 discrimination and concurrent-vowel identification[6]

3.0 Abstract

The aim of this study was to examine the effects of envelope-vocoder sound processing on listeners' ability to discriminate changes in fundamental frequency (F0) in anechoic and reverberant conditions, and on their ability to identify concurrent vowels based on differences in F0. In the first experiment, F0 difference limens (F0DLs) were measured as a function of number of envelope-vocoder frequency channels (1, 4, 8, 24, and 40 channels, and unprocessed) in four normal-hearing listeners, with degree of simulated reverberation (no, mild, and severe reverberation) as a parameter. In the second experiment, vowel identification was measured as a function of the F0 difference between two simultaneous vowels in six normal-hearing listeners, with the number of vocoder channels (8 and 24 channels, and unprocessed) as a parameter. Reverberation was detrimental to F0 discrimination in conditions with fewer numbers of vocoder channels. Despite the reasonable F0DLs (< 1 semitone) with 24- and 8-channel vocoder processing, listeners were unable to benefit from F0 differences between the competing vowels in the concurrent-vowel paradigm. The overall detrimental effects of vocoder processing are likely due to the poor spectral representation of the lower-order harmonics. The F0 information carried in the temporal envelope is weak, susceptible to reverberation, and may not suffice for source segregation. To the extent that vocoder

[6] A version of this chapter is accepted for publication as Qin and Oxenham (2005).

processing simulates cochlear-implant processing, users of current implant processing schemes are unlikely to benefit from F0 differences between competing talkers when listening to speech in complex environments. The results provide further incentive for finding a way to make the information from low-order, resolved harmonics available to cochlear-implant users.

3.1 Introduction

Fundamental frequency (F0) information has long been thought to play an important role in perceptually segregating simultaneous and non-simultaneous sources (for reviews see Bregman, 1990; Darwin and Carlyon, 1995). Studies with normal-hearing listeners have found that when a competing voice is present, listeners generally find it easier to understand the target voice if the competing voice has a different F0 (Brokx and Nooteboom, 1982; Assmann and Summerfield, 1990; Assmann and Summerfield, 1994; Darwin and Carlyon, 1995; de Cheveigné *et al.*, 1997; Bird and Darwin, 1998). Most models that use F0 differences to separate concurrent speech require explicit estimation of the F0 of either one or both sources (Assmann and Summerfield, 1990; Meddis and Hewitt, 1992). These models would predict that ambiguous F0 information leads to a deterioration in speech reception performance.

For normal-hearing listeners, the perception of voice pitch and the ability to discriminate different F0s are thought to rely primarily on temporal fine-structure information, in particular the information carried in peripherally resolved lower-order harmonics (e.g., Plomp, 1967; Houtsma and Smurzynski, 1990; Smith *et al.*, 2002). Under normal circumstances, the frequencies of these harmonics are believed to be encoded by their place of excitation on the basilar membrane, by the temporal pattern of their auditory-nerve (AN) responses, or by some combination of the two.

Cochlear-implant users are unlikely to use the same F0 cues as normal-hearing listeners. The reasons for this are related to various properties of cochlear implants, which operate by bypassing the outer, middle and inner ear to directly stimulate the auditory nerve. Most cochlear-implant users today are implanted with multi-channel devices (Clark *et al.*, 1990; Loizou, 1999). In continuous interleaved sampling (CIS), a widely used processing strategy for cochlear implants (Wilson *et al.*, 1991), the electrical stimulation delivered to the auditory nerve represents amplitude envelopes extracted from a small number of contiguous frequency bands or channels. The amplitude envelopes from each channel are low-pass filtered, typically at 400 Hz, and imposed on biphasic pulse carriers. The limited spectral resolution of current implant systems means that the lower harmonics of speech that give normal-hearing listeners spectral cues to pitch are not resolved. Furthermore, the lowpass filtering of the envelopes eliminates most temporal fine-structure cues. However, voice pitch is in principle available in implant-processed speech via the periodicity in the temporal envelope (Green *et al.*, 2002; Moore, 2003a), so long as the cutoff frequency of the envelope-extraction filter is sufficiently high to pass the voice F0. McKay *et al.* (1994) showed that some implant users can detect differences, in the range of human voice pitch, as small as 2%, although most users have considerably higher difference limens. For instance, Geurts and Wouters (2001), using synthetic vowels to measure the F0 difference limens (F0DLs), found that CIS implant users could discriminate differences of between 4% and 13%. Such difference limens (see also Busby *et al.*, 1993; Wilson, 1997) are an order of magnitude higher than those found in normal-hearing listeners when low-order (resolved) harmonics are present, but are only slightly higher than those found when only temporal envelope cues are presented to normal-hearing listeners (e.g., Burns and Viemeister, 1976; 1981; Shackleton and Carlyon, 1994; Kaernbach and Bering, 2001; Bernstein and Oxenham, 2003). Thus, normal-hearing listeners and cochlear-implant users may share the same inability to efficiently code periodicity from information in the temporal envelope (Carlyon *et al.*, 2002).

Another possible difficulty with envelope-periodicity F0 cues is their susceptibility to reverberation. The daily acoustic environments of implant users are often reverberant (e.g. living rooms, classrooms, music halls, and houses of worship). Previous studies examining the effects of cochlear-implant processing on F0 discrimination were conducted under anechoic conditions (Fu *et al.*, 1998b; Faulkner *et al.*, 2000; Green *et al.*, 2002), and the potential influence of reverberation has not been systematically examined. Reverberation has a "smearing" effect on envelope modulation (Houtgast *et al.*, 1980; Steeneken and Houtgast, 1980) and is therefore likely to be detrimental to envelope-based F0 perception, even for steady-state sounds.

Cochlear-implant users invariably exhibit poorer speech reception than normal. While poor speech reception in implant users can be due to many factors, poor F0 information may be one reason why performance is particularly poor in complex, fluctuating backgrounds, in which listeners must perceptually segregate the target from the masker (Nelson *et al.*, 2003; Qin and Oxenham, 2003; Nelson and Jin, 2004; Stickney *et al.*, 2004). Specifically, poor F0 coding may result in the loss of F0 as a segregation cue. While a link between F0 coding and speech segregation ability has been hypothesized before (Qin and Oxenham, 2003), no direct test has yet been undertaken.

The aim of this study was to address the question of a link between F0 coding and source segregation by examining the effects of using primarily envelope periodicity cues on F0 discrimination in anechoic and reverberant conditions (Experiment 1) and on the ability to use F0 differences in segregating and identifying competing sound sources (Experiment 2). We tested normal-hearing listeners using noise-excited envelope vocoder processing, as used in many previous studies (e.g., Shannon *et al.*, 1995; Dorman *et al.*, 1997; Fu *et al.*, 1998a; Shannon *et al.*, 1998; Rosen *et al.*, 1999). This technique simulates certain aspects of cochlear-implant processing, such as the loss of frequency resolution, by filtering the stimulus into a small number of broadly tuned frequency channels, and the loss of temporal fine-structure information by using only the temporal envelope in each frequency channel to modulate noise carriers. While such simulations clearly do not capture all aspects of cochlear implant perception (such as the limited dynamic range), they have certain advantages in that they avoid the large inter-subject variability in the performance of actual implant users, and that they can also be used to probe certain aspects of normal hearing.

3.2 F0 difference limens

The focus of this experiment was to examine the effects of noise-excited vocoder processing and reverberation on the ability of listeners to discriminate the small changes in F0 between two sequentially presented harmonic tone complexes.

3.2.1 Methods

3.2.1.1 Participants

Four normal-hearing listeners participated in this experiment (audiometric thresholds between 125 and 8000 Hz were <20 dB HL). They were undergraduate and graduate students with ages ranging from 19 to 28 years.

3.2.1.2 Stimuli

All stimuli were digitally generated, treated, processed, and stored on computer disk using Matlab (Mathworks, Natick MA). The original stimuli were harmonic tone complexes, composed of equal-amplitude harmonics between 80 and 6000 Hz. The stimuli were first treated to simulate various reverberation conditions, and then processed to simulate the effects of cochlear-implant sound processing.

Three conditions were tested. The first condition used sine-phase tone complexes, which simulate a pulsatile source, such as the human vocal folds, in an anechoic environment. The second condition convolved the sine-phase tone complexes with the recorded impulse response of a classroom ($RT_{60} = 0.5$ s) to simulate the effects of mild reverberation. The third condition used random-phase harmonic tone complexes, which were designed to simulate the phase relationships present in a highly reverberant environment. Randomizing the phase does not simulate all aspects of reverberation, but it does result in greatly reduced average temporal envelope modulations, which are probably most important for F0 discrimination based on temporal envelope properties.

Noise-excited envelope vocoder (EV) processing was used to simulate the effects of cochlear-implant sound processing (see Fig. 1-3). The unprocessed stimuli were first bandpass filtered (6th order Butterworth filters) into 1, 4, 8, 24, or 40 contiguous frequency channels between 80 and 6000 Hz. The entire frequency range was divided equally in terms of the Cam scale[7] (Glasberg and Moore, 1990). For instance, in the 24-channel condition, the filter with the lowest center frequency had lower and upper cutoff frequencies of 80 and 121.18 Hz, respectively. The bandwidths of the filters in the 24-channel condition are 1.16 Cams, which is only somewhat wider than the estimated bandwidths of human auditory filters (1 Cam, by the definition of Glasberg and Moore, 1990). To avoid differences in group delay between filters, zero-phase digital filtering

[7] This is more frequently referred to as the ERB scale. However, as pointed out by Hartmann (1997), ERB simply refers to equivalent rectangular bandwidth, which could be used to define all estimates of auditory filter bandwidths. We, therefore, follow Hartmann's convention of referring to the scale proposed by Glasberg and Moore as the Cam scale, in recognition of its origins in the Cambridge laboratories. Described in Glasberg and Moore (1990), $Cam = 21.4\log_{10}(0.00437f + 1)$, where f is frequency in Hz.

was performed.[8] The envelopes of the signals were extracted by half-wave rectification and lowpass filtering (using a 2nd-order Butterworth filter) at 300 Hz, or half the bandpass filter bandwidth, whichever was lower. The 300-Hz cutoff frequency was chosen to preserve F0 cues in the envelope as far as possible. The envelopes were then used to amplitude modulate independent white-noise carriers. The same bandpass filters that were used to filter the original stimuli were then used to filter the amplitude-modulated noises. Finally, the modulated narrowband noises were summed and scaled to have the same level as the unprocessed stimuli.

On each trial, the listener was presented with two successive stimulus tokens, separated by 200-ms pauses. Each stimulus token had a total duration of 200 ms and was gated on and off with 50-ms raised-cosine ramps. In the mild-reverberation condition, the stimuli were gated after reverberation had been added, so the total duration remained 200 ms. During each trial, one of the intervals contained the stimulus token with the nominal F0 ($F0_{ref}$), while the other interval contained the stimulus token with the comparison F0 ($F0_{ref} + \Delta F0$). The order of presentation of the two intervals was selected randomly with equal probability from trial to trial. The $F0_{ref}$ was roved by ±10% from trial to trial to encourage listeners to compare the F0 of the two stimuli presented within each trial, rather than relying on an internal reference. Two nominal $F0_{ref}$ of 130 and 220 Hz were tested. These values were selected as they represent the mean F0 of male and female speech respectively. The stimulus levels were roved by ±3 dB from interval-to-interval,

[8] Zero-phase forward and reverse digital filtering was implemented using the Matlab 'filtfilt' command.

around the mean overall level of 70 dB SPL, to minimize the effects of intensity on listener judgments.

3.2.1.3 Procedure

The F0DLs were measured using a two-alternative forced-choice paradigm. The one-up two-down adaptive procedure was used to track the 70.7% correct point (Levitt, 1971). At the beginning of a run, ΔF0 was set to 20%. The value of ΔF0 was reduced after two consecutive correct responses and increased following an incorrect response. The factor of variation of ΔF0 was initially 1.58. It was reduced to 1.25 after the first reversal, and then to 1.12 after next two reversals. Thresholds were calculated as the geometric mean of the ΔF0 values at the last six reversals. A threshold measurement was considered out of range if a listener was repeatedly (in at least 3 of the 5 runs) unable to identify the higher-F0 interval at ΔF0 values of 50%.

The experiment was conducted with the participant seated inside a double-walled soundproof booth. The pre-processed stimuli were played out via a soundcard (LynxStudio LynxOne) with 24-bit resolution at a sampling frequency of 22.05 kHz. The stimuli were then passed through a programmable attenuator (TDT PA4) and headphone buffer (TDT HB6) before being presented diotically via a pair of Sennheiser HD580 headphones. The listeners were instructed to indicate the interval that contained the stimulus with the higher pitch. The two intervals were marked visually, and visual feedback was provided after each trial. The response on each trial was collected via a computer keyboard inside the sound booth.

All participants went through a training session, of approximately 2 hours, to familiarize them with the stimuli and experimental tasks. Each participant took part in five experimental sessions of approximately 2 hours each. The five sessions took place over the span of 2-3 weeks, depending on the availability of the participants. Each experimental session consisted of 36 runs, measuring the F0DL for each experimental condition (2 F0 conditions x 3 reverberation conditions x 6 vocoder processing conditions). Conditions were presented in a random order that varied across subjects and across repetitions. Individual thresholds were calculated as the geometric mean from the five repetitions of each condition.

3.2.2 Results

The patterns of results from the individual subjects were very similar and so only the mean data are shown. Figure 3-1 shows the estimated F0DLs (expressed as a percentage of the $F0_{ref}$) as a function of number of vocoder channels. Each point represents the geometric mean across subjects and the error bars denote ±1 standard error of the mean on the logarithmic scale. The results from $F0_{ref} = 130$ Hz and $F0_{ref} = 220$ Hz are shown in the left and right panels, respectively. The different reverberation conditions (i.e. non-reverberant, mild reverberation, and random phase) are shown as different symbols. The long-dashed line shows the measurement limit. The up-pointing triangles represent F0DLs outside the range of measurement[9]. Overall, F0DLs decreased with increasing

[9] For the purposes of statistical analysis, out-of-range thresholds were set to the maximum allowable percentage (50%).

number of vocoder channels. Furthermore, the detrimental effects of reverberation on F0DLs increased with decreasing number of channels.

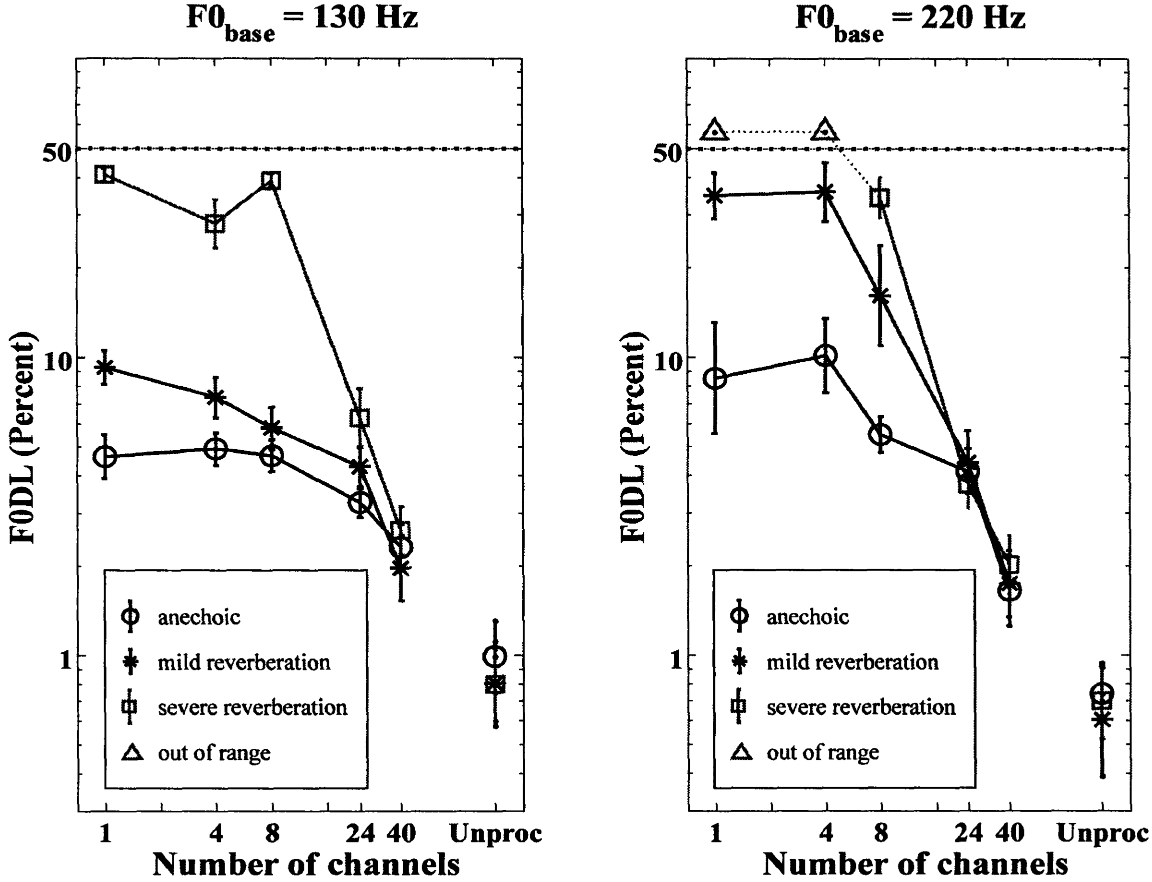

Figure 3-1: Mean results from experiment 1. Each data point represents the mean F0DL (%) across four subjects; error bars denote ±1 standard error of the mean. The results from the different $F0_{ref}$ are shown on separate plots, $F0_{ref}$ = 130 Hz on the left and $F0_{ref}$ = 220 Hz on the right. The different reverberation conditions (i.e. no reverberation, mild reverberation, and severe reverberation) are shown as different symbols. The long-dashed horizontal line shows the measurement limit. The up-pointing symbols represent F0DLs outside the range of measurement.

A three-factors (F0, reverberation, and processing condition) within-subject analysis of variance (ANOVA) was performed on the log-transformed data. The following main effects and interactions were found to be significant. There was a significant main effect of processing condition [$F_{5,15} = 116.725, p<0.001$], confirming the clear deterioration in performance with decreasing number of channels. Similarly, there was a main effect of reverberation [$F_{2,6} = 63.978, p<0.001$]. Significant interactions were found between processing condition and reverberation [$F_{10,30} = 32.259, p<0.001$], as well as between F0 and processing condition [$F_{5,15} = 18.186, p<0.001$]. These interactions reflect the trends in the data for reverberation to be more detrimental at small channel numbers than at large (or unprocessed), and for thresholds to be poorer at 220 Hz than at 130 Hz, especially in the reverberant conditions with a small number of channels.

Our findings can be understood in the context of envelope, spectral, and temporal fine-structure cues. When spectral and temporal fine-structure cues are more representative of the original stimuli (i.e. with large numbers of channels) the F0 percept is probably driven by spectral and temporal fine-structure cues. When spectral and temporal fine-structure cues are weak (i.e. with small numbers of channels) the F0 percept is probably derived from envelope cues. The salience associated with envelope cues is known to be weaker than that associated with spectral or fine-structure cues (Burns and Viemeister, 1976; 1981; Shackleton and Carlyon, 1994), therefore, the more listeners rely on envelope cues the less they are able to discriminate fine F0 differences. Furthermore, when listeners are forced to rely on envelope cues (i.e. with small numbers of channels) to perform F0 discrimination, the effects of reverberation are likely to be more detrimental. Reverberation can be characterized as a lowpass-filtering of the envelope modulation (e.g., Houtgast *et al.*, 1980; Steeneken and Houtgast, 1980). Thus, reverberation is likely to smear out envelope-based F0 information, particularly at the higher (220-Hz) F0, where the envelope fluctuations are more rapid. The auditory system itself also seems to act as a lowpass filter in the modulation domain (Kohlrausch *et al.*, 2000), which may explain why (with low channel numbers) thresholds were still somewhat poorer in the 220-Hz condition than in the 130-Hz condition, even in the absence of reverberation.

Overall, the detrimental effects of vocoder processing may be attributed to the poor spectral and temporal fine-structure representation of the lower-order harmonics and to the disruption to envelope F0 cues caused by reverberation.

3.3 Concurrent-vowel identification

As stated in the introduction, F0 information is believed to play an important role in perceptually segregating sound sources. While, in principle, temporal envelope cues to voice pitch are available in implant processed speech (e.g., Busby *et al.*, 1993; McKay *et al.*, 1994; Wilson, 1997; Geurts and Wouters, 2001), less is known about the ability of listeners to use such cues for segregation. Scheffers (1983) showed that two vowels played simultaneously with different F0s were easier to understand than two vowels with the same F0. Though Scheffers' concurrent-vowel paradigm is not an accurate representation of the everyday situation, it does offer a well-controlled means of examining the contribution of F0 information to the segregation of two speech sounds. By presenting two synthetic vowels simultaneously at equal level, the effects of semantic, grammatical, and other perceptual grouping cues, such as onset asynchronies, can be eliminated. While the concurrent-vowels paradigm has been used to examine the utility of F0 information for source segregation with normal-hearing listeners (Brokx and

Nooteboom, 1982; Assmann and Summerfield, 1990; Summerfield and Assmann, 1991; Culling and Darwin, 1993; Assmann and Summerfield, 1994; Culling and Darwin, 1994; Darwin and Carlyon, 1995; de Cheveigné *et al.*, 1997; Bird and Darwin, 1998) and hearing-impaired listeners (Arehart *et al.*, 1997), to our knowledge, the paradigm has not been used to examine the utility of F0 information with either actual or simulated cochlear-implant users. This experiment examined the effects of vocoder processing on normal-hearing listeners' ability to use F0 information for source segregation. Vowel identification was measured as a function of the difference in F0 (ΔF0) between the two vowels (0, 1, 2, 4, 8, 12, and 14 semitones) and processing conditions (unprocessed, 24-channel EV, and 8-channel EV).

3.3.1 Methods

3.3.1.1 Participants

Six native speakers of American English (audiometric thresholds between 125 and 8000 Hz were <20 dB HL) took part in this experiment. Their ages ranged from 19 to 22.

3.3.1.2 Stimuli

Five American English vowels (/i/ as in *heed*, /a/ as in *hod*, /u/ as in *hood*, /ɛ/ as in *head*, /ɝ/ as in *herd*) were synthesized using an implementation of Klatt's cascade synthesizer (Klatt, 1980). They were generated at a sampling frequency of 20 kHz, with 16-bit quantization. The formant frequencies and bandwidths (see Table 3-1) used to synthesize the vowels were based on the estimates of Hillenbrand *et al.* (1995) for an average male talker. The vowels were chosen for their positions in the F1-F2 space and because their natural duration characteristics (House, 1960; House, 1961) are similar to the stimulus durations used in this experiment (200 ms).

Table 3-1: Formant frequencies (Hz) for vowels. Values enclosed in parentheses represent formant bandwidths in Hz.

	Formant			
Vowel	**F1** (60)	**F2** (90)	**F3** (150)	**F4** (200)
/i/	342	2322	3000	3657
/a/	768	1333	2522	3687
/u/	378	997	2343	3357
/ε/	580	1799	2605	3677
/ɝ/	474	1379	1710	3334

Each vowel was generated with seven different fundamental frequencies (see Table 3-2). The synthesized F0 difference is more akin to an intra-talker F0 difference than an inter-talker F0 difference; this was intended to eliminate potential confounding sources of speaker cues (e.g. other glottal source differences and vocal tract length).

Table 3-2: Difference in fundamental frequency between constituent vowels in a concurrent-vowel pair, described in terms of the semitone difference between ΔF0 and 100 Hz.

ΔF0 (semitone)	**0**	**1**	**2**	**4**	**8**	**12**	**14**
$F0_A$ (Hz)	100.0	100.0	100.0	100.0	100.0	100.0	100.0
$F0_B$ (Hz)	100.0	105.9	112.2	126.0	158.7	200.0	224.5

The concurrent-vowel pairs were constructed by summing two single vowels with equal levels, with their onsets and offsets aligned, and with their pitch periods in phase at the onset of the stimulus. No vowel was paired with itself to generate the concurrent-vowel pairs. Each concurrent-vowel token was constructed using one vowel with an F0 of 100 Hz and the other with an F0 of 100 Hz + ΔF0, where the ΔF0 ranged from 0 to 14 semitones (see Table 2). This yielded a total of 140 concurrent-vowel stimuli (10 vowel-pairs x 2 F0 combinations x 7 ΔF0s). Each stimulus had a total duration of 200 ms and was gated on and off with 25-ms raised-cosine ramps. The stimuli were presented at an overall level of 70 dB SPL.

All stimulus tokens were digitally generated, processed, and stored on computer disk prior to the experiments. The noise-excited vocoder processing used to simulate the effects of cochlear implant sound processing is the same as that used in the previous experiment. All vowels and concurrent vowel pairs were processed under the 24-channel and the 8-channel processing conditions.

3.3.1.3 Procedure

Both single-vowel and concurrent-vowel identification performance was measured using a forced-choice paradigm. The listeners were instructed to identify the vowels heard by selecting visual icons associated with the vowels. In the single-vowel identification task, listeners were instructed to identify the vowel heard by selecting from five different choices. In the concurrent-vowel identification task, listeners were instructed to identify both the constituent vowels. Listener performance was scored as the percentage of correct responses. In the double-vowel identification task, a response was considered correct only if both vowels were correctly identified.

Each experimental session was broken into six blocks comprising the three vocoder-processing conditions with either the single vowels or concurrent vowels. Within each block, the presentation orders of the F0s (for single vowel identification) or F0 differences between the constituent vowels (for concurrent-vowel identification) were randomized. The single-vowel blocks contained 70 stimuli each (5 vowels x 7 F0 x 2 repetitions). The concurrent-vowel blocks contained 140 stimuli each (20 vowel-pairs x 7 ΔF0). During each trial, the responses were entered via a computer keyboard and mouse inside the booth. No feedback was provided.

Listeners were tested in a double-walled soundproof booth. The stimuli were played out via a soundcard (LynxStudio LynxOne), passed through a programmable attenuator (TDT PA4) and headphone buffer (TDT HB6), before being presented diotically via a pair of Sennheiser HD580 headphones.

Every participant took part in two training sessions and three experimental sessions. The training sessions were design to familiarize the participants with the experimental stimuli and the identification tasks. They were asked to perform the same task in the training sessions as in the experiment sessions. Prior to the experimental sessions, all participants were required to achieve at least 90% identification accuracy in the single-vowel task. The five sessions took place over the span of 1-2 weeks, depending on the availability of the participants. In this way, each subject heard each single-vowel condition a total of 1050 times, and each double-vowel condition a total of 2100 times.

3.3.2 Results

Figure 3-2 shows the identification accuracy as a function of the F0 in the single vowel identification task (dotted lines; shown as semitones above 100 Hz) and difference between the F0 of the constituent vowels in the concurrent-vowel identification task (solid lines). The unprocessed conditions are shown in the left panel, the 24-channel conditions in the center panel, and the 8-channel conditions in the right panel. For presentation purposes, the results are pooled across subjects, across vowels (or vowel-pairs) and across vowel order (i.e., which vowel had the higher F0). Some more fine-grained analysis is provided later.

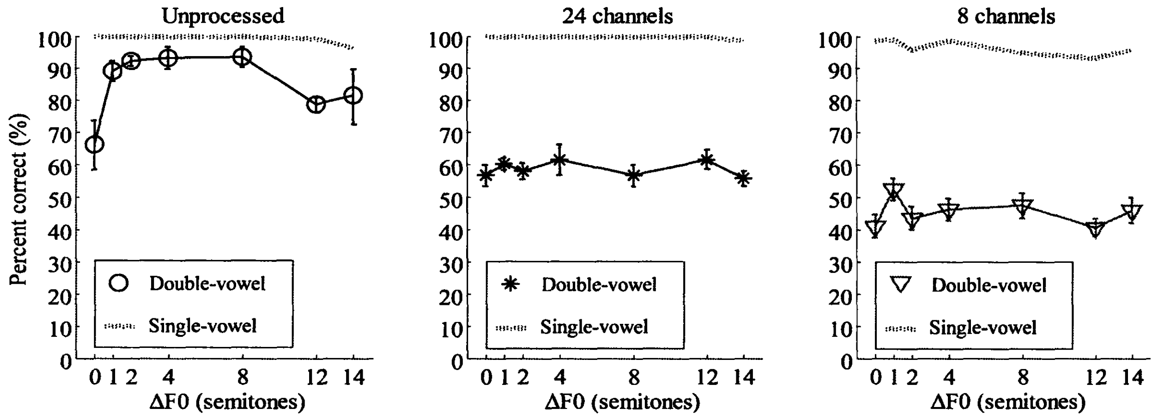

Figure 3-2: Dotted lines show the percent of correct responses as a function of the F0 in the single vowel identification task (dotted lines), where F0 is described in terms of the number of semitones from 100Hz. Solid lines show the percent of correct responses as a function of the ΔF0 (in semitones) between constituent vowels in the concurrent-vowel identification task, where the lower F0 was always 100 Hz. The error bars denote ±1 standard error of the mean. The unprocessed conditions are shown in the left panel the 24-channel condition in the center panel, and the 8-channel condition in the right panel.

To investigate trends in the concurrent-vowel data, a within-subject ANOVA with two factors (ΔF0 and processing condition) was conducted. All scores were arcsine transformed[10] (Studebaker, 1985) prior to analysis. The ANOVA analysis showed significant main effects of ΔF0 [$F_{6,30} = 10.343$, $p<0.001$] and processing condition [$F_{2,10} = 42.158$, $p<0.001$], and an interaction between ΔF0 and processing condition [$F_{12,60} = 10.150$, $p<0.001$]. Post-hoc comparison, using Bonferroni correction, showed that the mean score difference between the 24-channel condition (58.7%) and the 8-channel condition (45.3%) was significant ($p<0.05$), as was the difference between the unprocessed condition and the two processed conditions ($p<0.05$ in both cases).

In the unprocessed conditions (Fig. 3-2, left panel), listeners show an average 26 percentage points improvement in performance as ΔF0 increases from 0 to 2 semitones, after which their performance plateaus until the ΔF0 equals 12 semitones (one octave), consistent with Brokx and Nooteboom (1982). As expected, when the ΔF0 equals one octave the harmonics of the two constituent vowels becomes inseparable, leading to a drop in identification performance. At ΔF0 of 14 semitones, the identification performance seems to improve somewhat, although the performance difference between 12 and 14 semitones was not statistically significant ($p>0.05$).

In contrast to the unprocessed conditions, ΔF0 had no effect on performance in the concurrent-vowel identification task in either of the vocoder-processed conditions. A possible explanation for the lack of ΔF0 effect in the vocoder-processed conditions may

[10] The rationale for arcsine transformation is that the data of percent correct have non-uniform variance, whereas the transformed data have the property of stabilized variance of binomial data and thus are more suitable for analysis of variance (ANOVA) and other statistical analysis.

lie in the inability of the auditory system to use envelope cues to extract the F0s of two periodic stimuli that excite the same region of the cochlea. Carlyon (1996) and Carlyon *et al.* (2002) have shown that listeners fail to hear two underlying pitches when mixtures of two periodic pulse trains were either applied to a single implant electrode or presented to normal-hearing listeners after filtering out the lower harmonics. Our findings extend those of Carlyon and colleagues by showing that even with substantial spectral differences between the two sources (presumably leading to spectral formant regions where one or the other F0 dominates) listeners were not able to use these envelope-based F0 cues for segregation.

In the unprocessed conditions, the benefits of a difference in F0 were evident for all vowel pairs. However, levels of performance differed among vowel pairs. The largest ΔF0 benefit was seen when constituent vowels had similar first formant (F1) or second formant (F2) values (e.g., /herd, head/ and /herd, hod/). For example, /herd, head/ went from 38% at ΔF0 of 0 semitone to 85% at ΔF0 of 2 semitone; /herd, hod/ went from 41% at ΔF0 of 0 semitone to 75% at ΔF0 of 2 semitone. The least ΔF0 benefit was seen when constituent vowels have dissimilar F1 and F2 values (e.g., /heed, hod/). This is likely due to the already high performance (80%) at ΔF0 of 0 semitones.

With vocoder processing, while listeners had no trouble identifying any of the vowels presented in isolation (mean percent correct > 90%), they experienced difficulties with concurrent vowel identification. In the concurrent-vowel identification task, while listeners on average performed above chance (>5%), had a great deal of trouble when constituent vowels had similar F1 or F2 values. For example, in the 24-channel condition, the mean correct identification of the vowel pair /herd, head/ was ~40%, whereas /herd, hod/ was ~20 %. In the 8-channel condition, the mean correct identification of the vowel pair /herd, head/ was ~30%, whereas /herd, hod/ was ~10 %.

3.4 Discussion

In the first experiment, F0DLs were measured as a function of number of envelope-vocoder frequency channels, with degree of simulated reverberation as a parameter. It was found that sensitivity to small F0 differences decreased with decreasing number of vocoder channels, and the detrimental effects of reverberation on F0DLs increased with decreasing number of channels (Fig. 3-1). In the second experiment, vowel identification was measured as a function of the F0 difference in a concurrent-vowel paradigm, with the number of vocoder channels as a parameter. Under vocoder processing, listeners were unable to benefit from F0 differences, even when the ΔF0 between concurrent vowels was as large as 8 semitones (Fig. 3-2). In contrast, under the same vocoder processing conditions, F0DLs of less than 1 semitone (or < 6%) were found (Fig. 3-1).

A possible explanation for the apparent contrast between the findings of the two experiments may lie in the limits of the auditory system's ability to extract the F0 of two periodic stimuli presented simultaneously. In Experiment 1, listeners were asked to discriminate F0 difference of sequential presented tone complexes. In Experiment 2, listeners were asked to identify simultaneously presented vowels. It is conceivable that while the auditory system may be able to use envelope periodicity cues to extract the F0 of one periodic stimulus (as in Experiment 1), it may be incapable of extracting the F0s of two simultaneously presented stimuli (as is required in Experiment 2). In agreement with the present study, Deeks and Carlyon (2004) found no clear advantage for having their masker and target speech on different carrier rates, when the masker and target speech were passed through the same channels. They went further to show that the different carrier rates were not a useful cue for segregation, even when the masker and target speech excited separate channels. If the ability of listeners to extract the F0 from sequentially and simultaneously presented stimuli are indeed different, then the F0DLs found with sequentially presented stimuli would not be an appropriate indicator of a listener's ability to use F0 cues for speech segregation.

An alternative explanation is that it is not the mode of presentation (sequential vs. concurrent) but rather the nature of the stimuli (equal-amplitude sine-phase harmonics vs. vowel-shaped harmonics with phase shifts dependent on the vowel filtering) that explains the different outcomes of Experiments 1and 2. However, informal examinations of the temporal envelopes after vocoder processing suggest that both types of stimuli give similarly well-defined temporal envelopes.

It is worth reiterating here that while studies using normal-hearing listeners and noise-excited envelope vocoders have their advantages, the inherent differences between acoustic and electrical stimulation (Rubenstein *et al.*, 1999; Litvak *et al.*, 2001; Throckmorton and Collins, 2002) mean that the results from simulation studies should be interpreted in terms of trends rather than quantitative estimates of implant user performance. Taken in this light, the current results nevertheless suggest severe limits in the perceptual use of the envelope F0 information for cochlear-implant users. In the absence of other (auditory and non-auditory) cues, implant users are unlikely to benefit from differences in F0 between competing sources. While our current experiments do not address the question of whether F0 can be used as an additional cue in conjunction with others, our results are in line with predictions from earlier studies that poor F0 representation may underlie some of the difficulties experienced by implant users in complex environments (e.g., Qin and Oxenham, 2003).

3.5 Summary

1) Simulated cochlear-implant processing, using a noise-excited vocoder, had a detrimental effect on listeners' F0 discrimination abilities. Under processed conditions, performance worsened with decreasing numbers of vocoder frequency channels.
2) Reverberation did not affect F0 discrimination in unprocessed conditions, but was detrimental to F0 discrimination in processed conditions with fewer numbers of vocoder channels. The effect of reverberation was particularly marked at the higher $F0_{ref}$ (220 Hz), in line with expectations based on temporal-envelope processing.
3) Despite the reasonable F0DLs (< 1 semitone) with 24- and 8-channel vocoder processing in a sequential paradigm, listeners were unable to benefit from F0 differences between the competing vowels in a concurrent-vowel paradigm.
4) The overall detrimental effects of vocoder processing are likely due to the poor representation of the lower-order harmonics. The present study provides further

incentive for finding ways of making the information from low-order, resolved harmonics available to cochlear-implant users.

5) It remains possible that F0 information may provide additional help in real-world situations, where other (auditory, visual, and linguistic) cues are present. However, to the extent that vocoder processing simulates cochlear-implant processing, results from this and other studies provide no evidence that users of current implant processing schemes can benefit from F0 differences between competing talkers.

Chapter 4: Effects of introducing unprocessed low-frequency information on the reception of envelope-vocoder processed speech

4.0 Abstract

As cochlear implant technology matures, the criteria for implantation have become more lenient. Now, individuals with some residual hearing can be regarded as good candidates for cochlear implants. While residual acoustic hearing alone is unlikely to provide substantial speech intelligibility, low-frequency acoustic information added to implant-processed high-frequency information might provide significant benefits. The aims of the present study are to investigate the frequency extent of residual hearing necessary for a tangible benefit and to examine the extent to which any benefits of residual hearing are due to the increase salience of voice pitch. Normal-hearing listeners were presented with sound processed by noise-excited vocoders, designed to simulate aspects of envelope-vocoder implant processing. Residual hearing was simulated by introducing unprocessed information at the lowest frequencies. Experiment 1 measured sentence-level speech reception as a function of target-to-masker ratio, with the amount of residual hearing and masker type as parameters. Experiment 2 examined the effects of introducing unprocessed low-frequency information on listeners' ability to identify vowels in a concurrent-vowel paradigm. In experiment 1, adding information below 600 Hz to envelope-vocoded high-frequency information improved speech reception thresholds by 5.9 dB in the presence of a competing talker and by 4.0 dB in the presence of speech-

shaped noise, as compared with the 8-channel vocoder-only condition. Although the addition of unprocessed low-frequency information did not return speech reception performance to normal levels, it nevertheless produced a significant improvement compared to performance with vocoder-alone. In experiment 2, significant ΔF0 benefits were observed when unprocessed low-frequency speech information was added to the vocoder processed speech, even when the unprocessed information was extremely limited in range (<300 Hz). Findings of the present study suggest that the speech-reception benefits associated with the addition of residual hearing can be, at least in part, attributed to an increase in F0 representation. The current findings, taken together, lead us to be cautiously optimistic about the ability of combined electric and acoustic stimulation to enhance the perceptual segregation of speech.

4.1 Introduction

Despite many significant advances made in the development of cochlear implants (e.g., Dorman, 2000; Zeng, 2004), even the most successful cochlear-implant users do not hear as well as normal-hearing listeners. The differences in performance between normal-hearing listeners and cochlear-implant users are especially pronounced in understanding speech in complex auditory environments. Fu *et al.* (1998a) and Friesen *et al.* (2001) reported that implant users require higher target-to-masker ratios in broadband noise to achieve levels of speech reception performance comparable to normal-hearing listeners. More recent findings suggest that the differences in performance between normal-hearing listeners and implant users (real and simulated) are especially pronounced for speech in the presence of fluctuating backgrounds (Nelson *et al.*, 2003; Qin and Oxenham, 2003; Stickney *et al.*, 2004).

Nelson *et al.* (2003) evaluated cochlear-implant users' ability to understand sentences in quiet, steady-state noise, and modulated speech-shaped noise, with normal-hearing listeners serving as the comparison group. Their results showed that while normal-hearing listeners obtained significantly greater release from masking from a modulated masker than from steady-state noise, implant users obtained very little masking release. They attributed the lack of masking release to the nature of implant processing strategies, including the lack of spectral detail in the processed stimuli.

Qin and Oxenham (2003) examined the speech reception performance of normal-hearing listeners under noise-excited vocoder processing (designed to simulate the effects of envelope-vocoder implant processing) in speech-shaped noise, amplitude-modulated speech-shaped noise, and single-talker speech interference. They found that while increasing the number of spectral channels improved listeners' performance in fluctuating interference, even performance in the 24-channel processing condition was substantially poorer than in the unprocessed condition, despite comparable representations of the spectral envelope. The results also showed that vocoder processing was more detrimental to speech reception in fluctuating interference than in steady-state noise, so that the release from masking in normal-hearing listeners became an increase in masking under vocoder processing.

Stickney *et al.* (2004), using both real implant users and vocoder simulations, measured speech reception thresholds for target sentences from a male talker presented in the presence of one of three competing talkers (same male, different male, or female) or speech-shaped noise. They found no masking release in the presence of competing talkers for either implant users or normal-hearing subjects listening through the implant simulation. They suggested that the lack of masking release originated from the increased perceptual similarity of the target and masker due to reduced spectral resolution, a consequence of envelope-vocoder processing.

"Glimpsing" or "dip-listening" has been proposed as an explanation for the finding that fluctuating interference produces less masking of speech than steady-state maskers for normal-hearing listeners (e.g., Peters *et al.*, 1998). It is believed that normal-hearing listeners have the ability to take advantage of the spectral and temporal valleys inherent in fluctuating interferers to glimpse parts of the target speech. These glimpses of the target speech can then provide sufficient information to allow the listener to infer the entire message. It is important to remember here that the ability to benefit from glimpsing depends on both the audibility of the target (with respect to absolute threshold and masked threshold) and the ability to distinguish the target from masker. It is, therefore, possible that even if the interference does not render the target inaudible through energetic masking, listeners may not be able to perceptually separate the target from interference, producing something akin to informational masking. Informational masking is thought of as a threshold elevation due to non-energetic factors, such as target-masker similarity (e.g., Durlach *et al.*, 2003). The presence of a cue that reduces the similarity between the target and masker can presumably reduce the effects of informational masking.

Voice pitch, or the fundamental frequency (F0) of voicing, has long been thought to be a powerful grouping cue, playing an important role in the perceptual segregation of speech sources (e.g., Bregman, 1990; Darwin and Carlyon, 1995). While, in principle, voice pitch information is available to implant users via envelope modulations (McKay *et al.*, 1994; Wilson, 1997; Geurts and Wouters, 2001; Green *et al.*, 2002; Moore, 2003a), the pitch salience associated with envelope periodicity is known to be less robust than that associated with resolved lower-order harmonics in normal-hearing listeners (e.g., Burns and Viemeister, 1976; 1981; Shackleton and Carlyon, 1994). Qin and Oxenham (2003) suggested that the difference in speech reception performance between implant users and normal-hearing listeners in fluctuating maskers can be attributed in part to the loss of pitch as a segregation cue on the part of the implant users.

This idea was tested more directly in a recent study examining the effects of envelope-vocoder processing on F0 discrimination, and the use of F0 as a segregation cue (Qin and Oxenham, 2005). In that study, normal-hearing listeners using acoustic envelope-vocoders were asked to discriminate F0 differences between successive tone complexes and to identify vowels in a concurrent-vowel paradigm. F0 difference limens (F0DLs) were measured as a function of number of vocoder channels, corresponding to the degree of spectral resolution. Vowel identification performance was measured as a function of the F0 difference between the concurrent vowels, with the number of vocoder channels as a parameter. The results showed that vocoder processing had a detrimental effect on listeners' F0 discrimination abilities, but that the F0DLs were still less that 1 semitone (~6 %) in many conditions. However, despite reasonable F0DLs with 24- and 8-channel vocoder processing, listeners were unable to benefit at all from F0 differences in the concurrent-vowel paradigm. The results also show that simply increasing the number of channels (up to 24) will probably not provide sufficient F0 cues to aid segregation in cochlear implants, at least in a concurrent-vowel paradigm.

A possible explanation for this deficit, as suggested by the work of Carlyon and colleagues (Carlyon, 1996; Deeks and Carlyon, 2004), may be that the auditory system is incapable of using envelope-periodicity cues to extract the F0s of two concurrent voices, and instead relies on low-frequency harmonics as the segregation cue. Unfortunately, current implant systems are incapable of delivering usable resolved fine-structure information to implant users. However, as cochlear implant technology has matured, the criteria for implantation have become more lenient. Now, individuals with some residual hearing are being considered for cochlear implantation. Some of these individuals are able to use an implant in one ear while using an acoustic hearing aid in the other ear to form a hybrid electric plus acoustic system. While the residual acoustic hearing present in these implant users is unlikely to contribute directly to speech intelligibility, the additional low-frequency temporal fine-structure cue in acoustic hearing may provide sufficient information to aid in source segregation. A recent study by Kong et al. (in press) showed that speech reception in the presence of interference improved with combined electric and acoustic hearing compared to either alone.

Another recent development in cochlear implantation has been to implant electrodes only partially into the cochlea (short-insertion implant), in order to preserve the low-frequency residual acoustic hearing of these patients (von Ilberg *et al.*, 1999; Gantz and Turner, 2003). A study by Turner *et al.* (2004) showed the potential advantages of preserving low-frequency acoustic hearing in cochlear implant patients. In their study, three participants with an implanted "short-electrode" cochlear implant and preserved low-frequency acoustic hearing were tested on speech recognition in competing backgrounds. Performance of these participants was compared to that of a larger group of traditional cochlear implant users. Each of the three short-electrode subjects performed better than any of the traditional long-electrode implant subjects for speech recognition in a background of competing speech, but not in steady noise. When the short-electrode implant users were compared to a group of traditional implant users matched for speech recognition ability in quiet, the short-electrode implant users showed a 9-dB advantage in the multi-talker background. While both Kong *et al.* (in press) and Turner *et al.* (2004) suggest that the speech reception improvements may be due to the availability of salient F0 cues from the combined low-frequency acoustic and electric hearing, neither study was designed to explicitly examine this hypothesis.

The aims of the present study are to examine the frequency extent of residual hearing necessary for a tangible benefit and to investigate the extent to which the benefits of residual hearing are due to the increased salience of voice pitch. To achieve these aims the study was conducted using normal-hearing listeners presented with sound processed by noise-excited envelope-vocoders (EV), designed to simulate aspects of cochlear-implant processing. Residual hearing was simulated by re-introducing non-vocoder processed information at the low frequencies (LF). Experiment 1 measured speech reception accuracy as a function of target-to-masker ratio, with masker type (speech-shaped noise or competing talker) and processing condition [Unprocessed, 8-channel envelope-vocoder only (EV_{1-8}), unprocessed information below 300 Hz plus envelope-vocoder processed high-frequency information (LF_{300}+EV_{3-8}), and unprocessed information below 600 Hz plus envelope-vocoder processed high-frequency information (LF_{600}+EV_{4-8})] as parameters. Experiment 2 measured vowel identification accuracy as a function of F0 in both single-vowel and concurrent-vowel identification tasks, with the amount of simulated residual hearing (<300 Hz or < 600 Hz) and the type of processing [Unprocessed, unprocessed low-frequencies only (LF), envelope-vocoder processed high frequency only (EV), and LF+EV simulating the hybrid electric plus acoustic stimulation (EAS)] as the parameters.

4.2 Experiment 1: Speech reception in the presence of interference

4.2.1 Methods

4.2.1.1 Participants

18 native speakers of American English (audiometric thresholds between 125 and 8000 Hz were <20 dB HL) participated in this study. Their ages ranged from 18 to 28.

4.2.1.2 Stimuli

All stimuli in this study were composed of a target sentence presented in the presence of a masker. The stimulus tokens were processed prior to each experiment. The targets were H.I.N.T. sentences (Nilsson *et al.*, 1994) spoken by a male talker. The maskers were either a male competing talker or speech-shaped noise. The targets and maskers were combined at the desired target-to-masker ratios (TMRs) prior to any processing. TMRs were computed based on the token-length root-mean-square amplitudes of the signals. The maskers were gated on and off with 250-ms raised-cosine ramps.

The H.I.N.T sentence corpus consists of 260 phonetically balanced high-context sentences of easy-to-moderate difficulty. Each sentence is contains of four to seven keywords. The competing-talker maskers were excerpts from the audio book "Timeline" (novel by M. Crichton) read by Stephen Lang (as used in Qin and Oxenham, 2003). The competing-talker masker had a mean F0 (111.4 Hz) similar to that of the target talker (110.8 Hz), as estimated by the YIN program (de Cheveigné and Kawahara, 2002). To avoid long silent intervals in the masking speech (e.g. sentence-level pauses) the competing-talker maskers were automatically preprocessed to remove silent intervals greater than 100 ms. The competing-talker maskers and speech-shaped-noise maskers were spectrally shaped to match the long-term power spectrum of the H.I.N.T. sentences.

The maskers were then subdivided into non-overlapping segments to be presented at each trial. For a given listener, the target sentence lists were chosen at random, without replacement, from among the 25 lists of H.I.N.T. sentences. This was done to ensure that no target sentence was presented more than once to any given listener. Data were collected using one list (i.e., 10 sentences) for each TMR.

4.2.1.3 Stimulus Processing

The experimental stimuli for each listener were pre-processed and stored on disk prior to each experiment. All stimulus tokens were digitally generated, processed, and stored on computer disk. Stimulus processing was performed using Matlab (Mathworks, Natick MA) in the following manner (see Fig. 4-1).

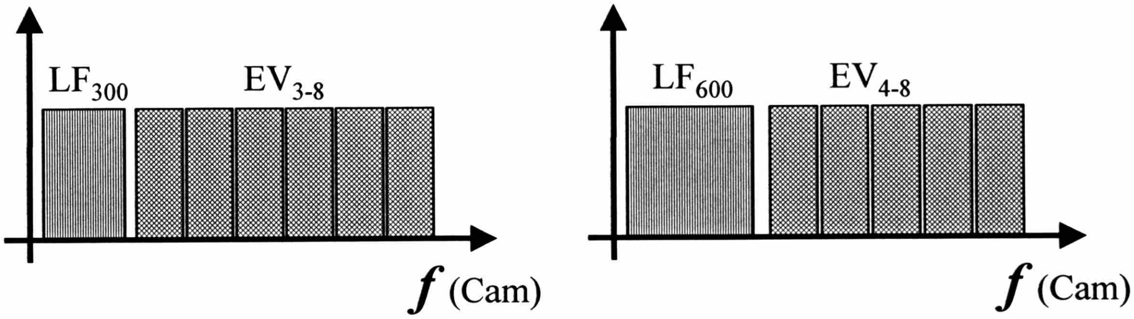

Figure 4-1: Schematic diagram of the processing conditions.

The experimental stimuli were presented in four processing conditions. In all conditions, the target levels were fixed at 65 dB SPL and the masker levels were varied to meet the desired TMR. The first processing condition (unprocessed), the stimuli were filtered between 80 Hz and 6 kHz, but were otherwise left unchanged. The second processing condition (EV_{1-8}), designed to simulate the effects of envelope-vocoder implant processing, used an 8 channel noise-excited vocoder (Qin and Oxenham, 2005). The input stimulus was first bandpass filtered (6th order Butterworth filters) into 8 contiguous frequency channels between 80 and 6000 Hz (see Table 4-1). The entire frequency range was divided equally in terms of the Cam scale (Glasberg and Moore, 1990). The envelopes of the signals were extracted by half-wave rectification and lowpass filtering (using a 2nd-order Butterworth filter) at 300 Hz, or half the bandpass filter bandwidth, whichever was lower. The 300-Hz cutoff frequency was chosen to preserve as far as possible F0 cues in the envelope. The envelopes were then used to amplitude modulate independent white-noise carriers. The same bandpass filters that were used to filter the original stimuli were then used to filter the amplitude-modulated noises. Finally, the modulated narrowband noises were summed and scaled to have the same level as the original stimuli.

Table 4-1: Filter cutoffs for the noise-excited vocoders.

	$EV_{1\text{-}8}$		**$LF_{300} + EV_{3\text{-}8}$**		**$LF_{600} + EV_{4\text{-}8}$**	
Channel number	**Low** (kHz)	**High** (kHz)	**Low** (kHz)	**High** (kHz)	**Low** (kHz)	**High** (kHz)
1	0.080	0.221	Unprocessed (0.080 – 0.300)		Unprocessed (0.080 – 0.600)	
2	0.221	0.426				
3	0.426	0.724	0.426	0.724		
4	0.724	1.158	0.724	1.158	0.724	1.158
5	1.158	1.790	1.158	1.790	1.158	1.790
6	1.790	2.710	1.790	2.710	1.790	2.710
7	2.710	4.050	2.710	4.050	2.710	4.050
8	4.050	6.000	4.050	6.000	4.050	6.000

The last two processing conditions (LF_{300}+$EV_{3\text{-}8}$ and LF_{600}+$EV_{4\text{-}8}$) were designed to simulate "electric plus acoustic" systems (EAS). To simulate the differing frequency extent of low-frequency residual hearing, the unprocessed stimuli were low-pass filtered at 300 Hz and 600 Hz (i.e., LF_{300} and LF_{600}), using a 3rd-order Butterworth filter. To simulate the effects of an EAS with residual hearing below 300 Hz (LF_{300}+$EV_{3\text{-}8}$), LF_{300} was paired together with $EV_{3\text{-}8}$, consisted of the upper 6 channels of the 8-channel vocoder. To simulate the effects of an EAS with residual hearing below 600 Hz (LF_{600}+$EV_{4\text{-}8}$), LF_{600} was paired with $EV_{4\text{-}8}$, consisted of the upper 5 channels of the 8-channel vocoder simulation (see Table III for vocoder channel frequency cutoffs).

4.2.1.4 Procedure

The 18 listeners were divided into two groups of nine. The speech reception of each group was measured under only one of the masker types (i.e. competing-talker or speech-shaped noise). Listeners were instructed to type their responses into the computer via a keyboard. No feedback was given.

The stimuli were converted to the analog domain using a soundcard (LynxStudio, LynxOne) at 16-bit resolution with a sampling rate of 22050 Hz. They were then passed through a headphone buffer (TDT HB6) and presented diotically via Sennheiser HD580 headphones to the listener seated in a double-walled sound-insulation booth.

Prior to the experiment session, listeners were given practice performing the experimental tasks as well as given exposure to the processed stimuli. However, the target sentences used in the training sessions came from the IEEE corpus (IEEE, 1969), whereas the target sentences in the experiment sessions came from the H.I.N.T. corpus. While the maskers in the training and experiment session came from the same corpus, care was taken to ensure that the same masker token was never repeated. During the training session, the listener was exposed to a total of 35 stimulus tokens (five lists, with 7 sentences per list), at each of the 4 processing conditions. At each processing condition, the target sentences were presented at a TMR in the mid-range of the experimental TMRs (see Table 4-2). The listeners were instructed to enter their responses, as in the experiment. No feedback was given.

Table 4-2: The values in the table represent the Target-to-masker ratio (in dB).

Processing condition	Masker type	Target-to-masker ratio (dB)
Unprocessed	Competing talker	[-20, -15, -10, -5, 0]
	Steady-state noise	[-10, -7, -5, -3, 0]
NEV$_{1-8}$	Competing talker	[-5, 0, 5, 10, 15]
	Steady-state noise	[-5, -1, 2, 6, 10]
LPF$_{300}$ + NEV$_{3-8}$	Competing talker	[-10, -5, 0, 5, 10]
	Steady-state noise	[-5, -1, 2, 6, 10]
LPF$_{600}$ + NEV$_{4-8}$	Competing talker	[-10, -6, -3, 1, 5]
	Steady-state noise	[-10, -6, -3, 1, 5]

In the experiment session, speech reception of each listener was measured under all four processing conditions (Unprocessed, EV$_{1-8}$, LF$_{300}$+EV$_{3-8}$, LF$_{600}$_EV$_{4-8}$), at 5 TMRs (see Table 4-2), in the presence of one masker type. The TMRs for each processing condition and masker type were determined in an earlier pilot study, using two to three listeners. The TMRs were chosen to minimize floor and ceiling effects in the psychometric function.

4.2.1.5 Analysis

Listener responses were scored offline by the experimenter. Obvious misspellings of the correct word were considered correct. Each listener's responses for a given TMR, under a given masker condition, were grouped together to produce a percent correct score. Keywords were used to calculate the percent correct.

4.2.2 Results and discussion

4.2.2.1 Fits to the psychometric functions

The percent correct scores as a function of TMR under a given masker condition for each listener were fitted to a two-parameter sigmoid model (a cumulative Gaussian function):

$$\text{Percent Correct} = \frac{100}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\text{TMR}} \exp\left(\frac{-(\text{x}-\text{SRT})^2}{2\sigma^2}\right)dx \quad \text{(Eq. 4-1)}$$

,

where x is the integration variable, SRT is the speech reception threshold in dB at which 50% of words were correctly identified, σ is related to the slope of the function, and TMR is the target-to-masker ratio (dB). The two-parameter model assumes that listeners' peak reception performance is 100%. Presented in Table 4-3 are the mean values of SRT, σ, and standard error of fit, averaged across listeners. The standard deviations are shown in parentheses. The two-parameter model provided generally good fits to performance as a function of TMR. The individual standard-errors-of-fit had a mean of 2.65% with a standard deviation of 2.52% (median of 1.87% and a worst case of 12.93%).

Table 4-3: Mean sigmoidal model parameter values (Eq. 4.1), averaged across listeners. SRT is the speech reception threshold, and σ is related to the slope of the function. The standard error of fit provides a numerical indicator for how well the model fits the data. The standard errors of mean are in parentheses.

Processing condition	Masker type	SRT (dB TMR)	σ	Standard error of fit (%)
Unprocessed	Competing talker	-13.58 (1.1)	10.43 (1.9)	5.0 (1.5)
	Steady-state noise	-6.04 (0.2)	3.93 (0.2)	1.2 (0.2)
EV_{1-8}	Competing talker	3.59 (0.4)	5.66 (0.8)	2.7 (0.9)
	Steady-state noise	1.59 (0.4))	5.60 (0.6)	1.7 (0.5)
LF_{300} + EV_{3-8}	Competing talker	1.34 (0.8)	6.87 (0.3)	2.5 (0.6)
	Steady-state noise	-0.94 (0.3)	4.59 (0.2)	2.2 (0.5)
LF_{600} + EV_{4-8}	Competing talker	-2.33 (1.0)	9.71 (1.2)	4.2 (0.9)
	Steady-state noise	-2.37 (0.3)	4.50 (0.4)	1.2 (0.5)

4.2.2.2 Speech reception thresholds (SRT)

The mean SRT values and standard errors of means (Table 4-3) were derived from the SRT values of individual model fits. In general, performance across listeners was consistent, so only the mean SRT values as a function of masker condition and processing condition are plotted in Fig. 4-2. A higher SRT value implies a condition more detrimental to speech reception. Figure 4-2 shows that in the unprocessed conditions, the steady-state noise masker was a more effective masker than the competing-talker masker. However, with 8-channel envelope-vocoder processing (EV_{1-8}) the reverse was true (i.e. competing talker was the more effective masker), consistent with the findings of Qin and Oxenham (2003). When unprocessed low-frequency information (LF) was added to the envelope-vocoder processed speech, the SRT associated with both maskers decreased, indicating a speech reception benefit.

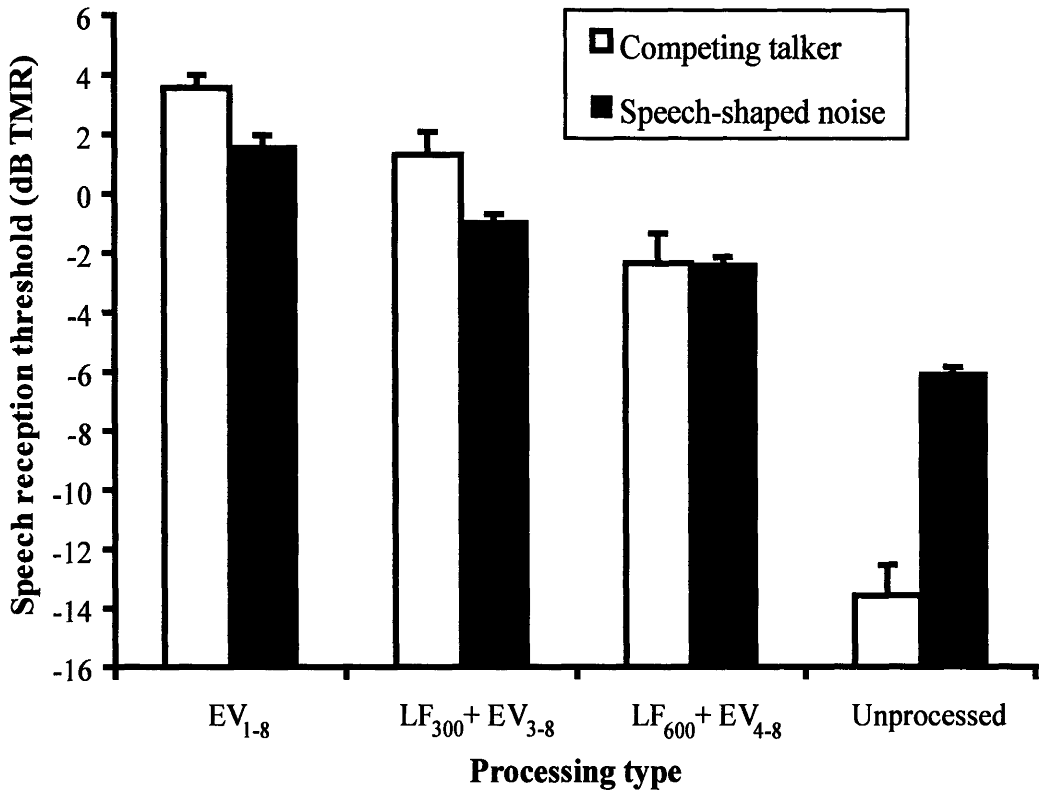

Figure 4-2: Group mean speech reception threshold (SRT) values for the two types of background interference. The error bars denote ±1 standard error of the mean.

A two-way mixed-design analysis of variance (ANOVA) was performed on the data for the EV_{1-8}, LF_{300}+EV_{3-8}, and LF_{600}+EV_{4-8} conditions and for both SSN and CT maskers. The statistical significance of the findings was determined with SRT as the dependent variable, masker type as the between-subjects factor, and processing condition as the within-subjects factor. The ANOVA analysis showed significant main effects of both processing condition ($F_{2,32}$ = 57.16, p<0.05) and masker type ($F_{1,16}$ = 5.18, p<0.05). Of particular interest is the significant interaction between masker type and processing condition ($F_{2,32}$ = 3.49, p<0.05), which indicates that an advantage was seen in competing talkers over speech-shaped noise for the LF+EV conditions as compared to the EV alone condition. Fisher's Least Significance Difference test (α=0.05) indicated several significant differences between the different experimental conditions, outlined below.

In the presence of a speech-shaped noise, when unprocessed low-frequency information (LF) was added to the envelope-vocoder (EV) processed speech, improved SRTs were observed in both $LF_{300}+EV_{3-8}$ and $LF_{600}+EV_{4-8}$ conditions.

In the presence of a competing talker, when unprocessed low-frequency information (LF) was added to the envelope-vocoder (EV) processed speech, both $LF_{300}+EV_{3-8}$ and $LF_{600}+EV_{4-8}$ seem to show an improvement in SRT, although only the $LF_{600}+EV_{4-8}$ was significantly different ($p<0.05$) when compared with the 8-channel envelope-vocoder alone condition (EV_{1-8}).

The present result differs slightly from that of Turner *et al.* (2004) in the noise condition. In their study, no significant difference was observed between processing conditions in the presence of noise, whereas in the current study improved SRTs were observed between both LF+EV conditions and the EV alone condition. Differences in the experimental paradigm and specifics of the speech materials may account for this discrepancy. The general finding of decreased SRT when unprocessed low-frequency information (LF) is added to the envelope-vocoder processed speech agrees with their results (Turner *et al.*, 2004).

In summary, the addition of unprocessed low-frequency information improved performance in speech-shaped noise, regardless of the low-frequency cutoff. The LF_{600}+EV_{4-8} condition performance produced a significant improvement compared to the EV alone condition in the presence of competing talkers. Adding unprocessed information below 600 Hz to envelope-vocoded high-frequency information improved speech reception threshold by 5.9 dB in the presence of a competing talker and by 4.0 dB in the presence of speech-shaped noise compared to the 8-channel envelope-vocoder only condition. Although the addition of low-frequency information did not return speech reception performance to normal levels, it was nevertheless a significant improvement when compared with envelope-vocoder alone.

4.3 Experiment 2: Concurrent-vowel identification

4.3.1 Rationale

This experiment examined the effects of reintroducing low-frequency information on listeners' ability to identify vowels in a concurrent-vowel paradigm. As stated in the Introduction, while previous studies (Turner *et al.*, 2004; Kong *et al.*, in press) have suggested that the speech reception improvements from the combined electric hearing and low-frequency acoustic hearing may be due to the availability of salient F0 cues, those studies were not designed to explicitly examine this hypothesis. Though the concurrent-vowel paradigm is not an accurate representation of the everyday situation, it does offer a well-controlled means of examining the contribution of F0 information to the segregation of two speech sounds. By presenting two synthetic vowels simultaneously at equal level, the effects of semantic, grammatical, and other grouping cues, such as onset asynchronies, can be eliminated. Vowel identification accuracy was measured as a function of F0 in both single-vowel and concurrent-vowel identification tasks, with the amount of simulated residual hearing (<300 Hz or < 600 Hz) and processing condition as the parameters.

4.3.2 Methods

4.3.2.1 Participants

Six native speakers of American English (audiometric thresholds between 125 and 8000 Hz were <20 dB HL) were paid for their participation in this experiment. Their ages ranged from 19 to 26.

4.3.2.2 Stimuli

Five American English vowels (/i/ as in *heed*, /a/ as in *hod*, /u/ as in *hood*, /ɛ/ as in *head*, /ɝ/ as in *herd*) were synthesized using an implementation of Klatt's cascade synthesizer (Klatt, 1980). They were generated at a sampling frequency of 20 kHz, with 16-bit quantization. The formant frequencies (see Table 4-4) used to synthesize the vowels were based on the estimates of Hillenbrand *et al.* (1995) for an average male talker. The vowels were chosen for their positions in the F1-F2 space and because their natural duration characteristics (House, 1960; 1961) are similar to the stimulus durations used in this experiment (i.e. 200 ms). Each vowel was generated with seven different fundamental frequencies ranging from 0 to 14 semitones above 100 Hz (100, 105.9, 112.2, 126.0, 158.7, 200.0, and 224.5 Hz).

Table 4-4: Formant frequencies (Hz) for vowels. Values enclosed in parentheses represent formant bandwidths in Hz.

	Formant			
Vowel	**F1** (60)	**F2** (90)	**F3** (150)	**F4** (200)
/i/	342	2322	3000	3657
/a/	768	1333	2522	3687
/u/	378	997	2343	3357
/ɛ/	580	1799	2605	3677
/ɝ/	474	1379	1710	3334

The concurrent-vowel pairs were constructed by summing two single vowels with equal levels, with their onsets and offsets aligned, and with their pitch periods in phase at the onset of the stimulus. No vowel was paired with itself to generate the concurrent-vowel pairs. Each concurrent-vowel token was constructed using one vowel with an F0 of 100 Hz and the other with an F0 of 100 Hz + ΔF0, where the ΔF0 ranged from 0 to 14 semitones. This yielded a total of 140 concurrent-vowel stimuli (20 vowel-pairs x 7 ΔF0s). Each stimulus had a total duration of 200 ms and was gated on and off with 25-ms raised-cosine ramps. The stimuli were presented at an overall level of 70 dB SPL.

4.3.2.3 Stimulus processing

All stimulus tokens were digitally generated, processed, and stored on computer disk prior to the experiments. The experimental stimuli were presented in five conditions (Unprocessed, LF_{300}, LF_{600}, $EV_{3\text{-}8}$, $EV_{4\text{-}8}$, LF_{300}+$EV_{3\text{-}8}$, and LF_{600}+$EV_{4\text{-}8}$). In the unprocessed condition, the stimuli were filtered between 80 Hz and 6 kHz, but were otherwise left unchanged. To simulate the differing amounts of residual hearing (LF_{300} and LF_{600}) the unprocessed stimuli were low-pass filtered at 300 Hz and 600 Hz respectively, using a 3rd-order Butterworth filter. To simulate the effects of envelope-vocoder implant processing, an 8-channel noise-excited vocoder (Qin and Oxenham, 2005) was used. The condition $EV_{3\text{-}8}$ consisted of the upper 6 channels of the 8-channel vocoder. The condition $EV_{4\text{-}8}$ consisted of the upper 5 channels of the 8-channel vocoder simulation. To simulate the effects of residual hearing below 300 Hz plus cochlear implant (LF_{300}+$EV_{3\text{-}8}$), LF_{300} was paired together with $EV_{3\text{-}8}$. To simulate the effects of residual hearing below 600 Hz plus cochlear implant (LF_{600}+$EV_{4\text{-}8}$), LF_{600} was paired with $EV_{4\text{-}8}$. In all cases, the processing was identical to that described in the previous experiment.

4.3.2.4 Procedure

Listeners were individually tested in a double-walled soundproof booth. The stimuli were played out via a soundcard at 16-bit resolution and a sampling rate of 20 kHz (LynxStudio LynxOne), and passed through a programmable attenuator (TDT PA4) and headphone buffer (TDT HB6), before being presented diotically to listeners via a pair of Sennheiser HD580 headphones.

Performance on single-vowel and concurrent-vowel identification was measured using a forced-choice paradigm. The listeners were instructed to identify the vowels heard by selecting visual icons associated with the vowels. In the single-vowel identification task, listeners were instructed to identify the vowel heard by selecting from five different choices. In the concurrent-vowel identification task, listeners were instructed to identify both of the constituent vowels. The responses were entered via a computer keyboard and mouse inside the booth. No feedback was provided.

Each listener took part in six 2-hour sessions. Three sessions incorporated conditions simulating below 300 Hz residual hearing (Unprocessed, LF_{300}, $EV_{3\text{-}8}$, and LF_{300}+$EV_{3\text{-}8}$), and the 3 other sessions incorporated conditions simulating below 600 Hz residual hearing (Unprocessed, LF_{600}, $EV_{4\text{-}8}$, and LF_{600}+$EV_{4\text{-}8}$). The 3 sessions simulating below 300 Hz residual hearing and 3 sessions simulating below 600 Hz residual hearings were interleaved, with the order randomized across subjects.

Each experiment session was sub-divided into 8 blocks, in accordance with processing condition (Unprocessed, LF, EV, and LF+EV), with the first 4 blocks measuring the single-vowel identification and the next 4 blocks measuring concurrent-vowel identification. The orders of the blocks were randomized from session to session.

Within a given block, the stimulus tokens were presented at random. Within each single-vowel identification block, a total of 70 stimulus tokens were presented (7 F0 x 5 vowels x 2 repetitions). For each listener, this translates to a total of 30 trials (5 vowels x 2 repetitions x 3 sessions) at each F0 under each processing condition. Within each concurrent-vowel identification block, a total of 140 stimulus tokens were presented (7 ΔF0 x 20 vowel-pairs). For each listener, this translates to a total of 60 trials (20 vowel-pairs x 3 sessions) at each F0 under each processing condition. In the unprocessed condition listeners were expose to twice as many trials at each F0, because the processed condition was presented in each of the 6 sessions. The 6 sessions took place over the span of 2-3 weeks, depending on the availability of the participants.

Prior to each experiment session, every listener was given practice performing the experimental tasks as well as given exposure to the stimuli. Listeners were instructed to enter their responses as in the experiment, with no feedback provided. Data gathering did not commence until the experimenter was satisfied that the listener understood the tasks. On average, listeners were exposed to 40-80 stimulus tokens (5-10 stimulus tokens x 8 blocks), prior to data gathering.

4.3.3 Results and discussion

Fig. 4-3 shows the identification accuracy as a function of the F0 in the single-vowel identification task (dotted lines) and the F0 of the upper vowel in the concurrent-vowel identification task (solid lines), in terms of semitones above 100 Hz. The unprocessed conditions are shown in Fig. 4-3A, the LF conditions in Fig. 4-3B, the EV conditions in Fig. 4-3C, and the LF+EV conditions in Fig. 4-3D. To investigate trends in the data, repeated-measures ANOVAs were conducted. All scores were arcsine transformed prior to statistical analysis.

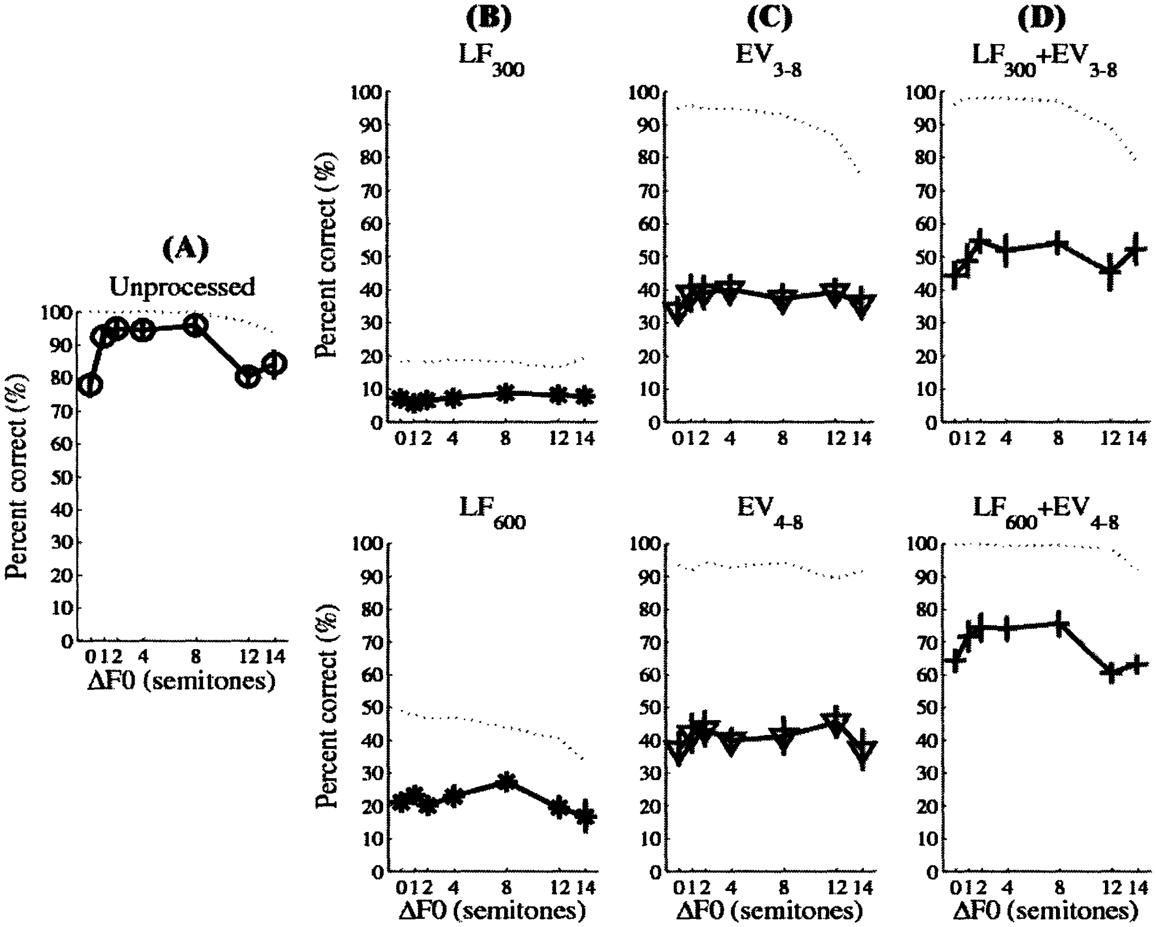

Figure 4-3: Dotted lines show the percent of correct responses as a function of the F0 in the single vowel identification task (dotted lines), where F0 is described in terms of the number of semitones from 100 Hz. Solid lines show the percent of correct responses as a function of the ΔF0 (in semitones) between constituent vowels in the concurrent-vowel identification task, where the lower F0 was always 100 Hz. The error bars denote ±1 standard error of the mean. The unprocessed conditions are shown in the left panel (A), next the lowpass filtered (LF) conditions (B), then the envelope-vocoder (EV) conditions (C), finally the LF+EV conditions in the right panels (D). The top row of figures is associated with the 300 Hz lowpass sessions, and the bottom row of figures is associated with the 600 Hz lowpass sessions.

In the unprocessed conditions (Fig. 4-3A), a single factor repeated-measures ANOVA was conducted with ΔF0 as the within-subject factor. The analysis revealed that the effect of ΔF0 differences was statistically significant ($F_{6,30} = 11.87$, $p<0.05$). In fact, listeners show an improvement in performance from 78% to 95% as ΔF0 increases from 0 to 2 semitones. Above this ΔF0, performance plateaus until the ΔF0 equals 12 semitones (one octave), consistent with previous studies (Brokx and Nooteboom, 1982; Culling and Darwin, 1993; Qin and Oxenham, 2005). When the ΔF0 equals one octave, the harmonics of the two constituent vowels becomes inseparable, leading to a drop in identification performance. At ΔF0 of 14 semitones, the identification performance seems to improve somewhat, although the performance difference between 12 and 14 semitones is not statistically significant (Fisher's LSD, $p>0.05$).

In the LF conditions (Fig. 4-3B), identification performance was greatly reduced, as predicted (ANSI, 1997). When a repeated-measures ANOVA with two within-subject factors (ΔF0 and lowpass cutoff) was conducted, a statistically significant difference was found between LF_{300} and LF_{600} ($F_{1,5} = 11.03$, $p<0.05$), but there was no statistically significant effect of ΔF0 ($F_{6,30} = 1.67$, $p>0.05$). In the LF_{300} condition, performance was reduced to near chance (5%) and no benefits of F0 differences were seen. In the LF_{600} condition, although identification of both single and concurrent vowels improved, no benefit of F0 differences was observed.

In the EV conditions (Fig. 4-3C), while single vowel identification was generally high, concurrent vowel identification was modest (~40%). When a repeated-measures ANOVA with two within-subject factors (ΔF0 and lowpass cutoff) was conducted, no statistically significant difference was found between $EV_{3\text{-}8}$ and $EV_{4\text{-}8}$ ($p>0.05$). In addition, no effect of ΔF0 ($F_{6,30} = 1.67$, $p>0.05$) was seen. This suggests that vowel identification performance does not improvement as a function of F0 difference in either of the EV conditions, consistent with observations from our previous study (Qin and Oxenham, 2005).

In the LF+EV conditions (Fig. 4-3D), identification performance improved compared to both LF-only and EV-only conditions; in addition, overall performance in the $LF_{600}+EV_{4-8}$ condition was better than that in the $LF_{300}+EV_{3-8}$ condition. A repeated-measures ANOVA with two within-subject factors (ΔF0 and lowpass cutoff) showed significant main effects of cutoff ($F_{1,5} = 23.71$, $p<0.05$) and ΔF0 ($F_{6,30} = 6.73$, $p<0.05$). More interestingly, the interaction between cutoff and ΔF0 was found to be significant ($F_{6,30} = 2.59$, $p<0.05$), suggesting that somehow the effects of ΔF0 were different depending on the lowpass cutoff. However, when a repeated-measures ANOVA was performed, excluding those data associated with the 12- and 14-semitones condition, the analysis showed no interaction between cutoff and ΔF0 ($F_{4,20} < 1$, n.s.). This indicates that the ΔF0 benefits $LF_{300}+EV_{3-8}$ and $LF_{600}+EV_{4-8}$ are not statistically different, for moderate ΔF0 between 0 semitones to 8 semitones, independent of lowpass cutoff frequencies.

To further investigate trends in the LF+EV conditions, single factor repeated-measures ANOVAs were conducted on each the LF+EV conditions, with ΔF0 as the within-subject factor. The analysis indicates that the main effect of ΔF0 differences was statistically significant in both the $LF_{300}+EV_{3-8}$ condition ($F_{6,30} = 2.81$, $p<0.05$) and the $LF_{600}+EV_{4-8}$ condition ($F_{6,30} = 7.84$, $p<0.05$). In both LF+EV conditions (Fig. 4-3D), the findings parallel those seen in the unprocessed condition (Fig. 4-3A). Listeners show an average improvement of 10 percentage points in performance as ΔF0 increases from 0 to 2 semitones (Fisher's LSD, $p<0.05$), after which their performance plateaus until the ΔF0 equals 12 semitones, where a drop in identification performance is observed. At ΔF0 of 14 semitones, the identification performance seems to improve somewhat, although the performance difference between 12 and 14 semitones is not statistically significant (Fisher's LSD, $p>0.05$).

4.4 Discussion

As the criteria for implant candidacy become more lenient, there are new issues that must be considered in managing treatment for patients who retain some residual hearing. Recent developments in cochlear implantation, i.e. short-insertion implants and EAS systems, make it critical to gather evidence measuring the benefits of retaining residual hearing in conjunction with cochlear implant inputs. The results of the current study came from acoustic simulations of EAS users. Residual hearing plus implant processing was simulated by adding unprocessed low-frequency information to noise-excited vocoder processed high-frequency information. Throughout the study, we assumed perfect residual hearing (e.g., no broadening of the auditory filter or threshold shift) as well as perfect alignment between the analysis channels and the place of electrical stimulation. It is important to note that these assumptions are unlikely to be met by real EAS users. Therefore, care should be taken when interpreting the current findings and discussing their implications.

4.4.1 Frequency extent of residual hearing necessary to see tangible benefits in speech reception

Given the potential variability in the amount of residual hearing available across the population of cochlear implant candidates, one aim of the current study was to investigate the frequency extent of residual hearing necessary to show tangible speech reception benefits. Our findings suggest that even an extremely limited range (<300 Hz) of residual hearing may be beneficial to the reception of speech in the presence of interference. Both experiment 1 and 2 showed that when unprocessed information below 300 Hz was added to envelope-vocoder processing, significant speech identification improvement could be observed. In experiment 1, the SRT decreased by 2.5 dB in steady-state noise. This 2.5 dB improvement in SRT translates to an improvement in intelligibility of ~20%. In experiment 2, concurrent vowel identification improved beyond that of the vocoder only condition. In addition, listeners exhibited ΔF0 benefits that were not observed in the vocoder-only condition. While the addition of low-frequency information did not return speech reception performance to normal levels in either experiment, the improvement was nevertheless significant when compared with envelope-vocoder alone. The current findings, taken together with the positive results from real EAS users (Tyler *et al.*, 2002; Ching *et al.*, 2004; Turner *et al.*, 2004; Kong *et al.*, in press), lead us to be cautiously optimistic about the ability of combined electric and acoustic stimulation to enhance the perceptual segregation of speech.

4.4.2 Extent to which benefits can be attributed to improved F0 representation

As stated in the Introduction, previous researchers (Turner *et al.*, 2004; Kong *et al.*, in press) have suggested that speech reception benefits of residual hearing may be attributable to an improvement in F0 representation. However, the subjects in their studies had residual hearing up to frequencies as high as 1 kHz. With this level of residual acoustic hearing, speech-reception benefits could be attributed to increased spectral resolution, yielding more accurate formant frequency information, rather than to improvements in F0 representation. The current simulation experiment examined the effect of adding unprocessed information below 300 Hz. While this frequency range contains very little speech information (ANSI, 1997), it provides information regarding the fundamental frequency of voicing.

In experiment 2 of this study, five of the six vowels had first formant frequencies below 600 Hz (see Table 4-4) but none had first formant frequencies below 300 Hz. Thus spectral resolution available with simulated residual hearing could have provided more accurate formant frequency information when unprocessed information below 600 Hz was present, but not when only unprocessed information below 300 Hz was added. This was confirmed by the results from experiment 1, where single-vowel identification performance in the below 300 Hz only condition was at chance, but above chance in the below 600 Hz only condition. In contrast, the ΔF0 benefits in concurrent-vowel identification results were similar in the <300 Hz and <600 Hz conditions, supporting the notion that the addition of unprocessed low-frequency information improved F0 representation and thus improved speech segregation abilities.

It has also been speculated that the F0 benefit observed in concurrent-vowel identification arises from "beating" between adjacent components with concurrent vowels (Culling and Darwin, 1994; Darwin and Carlyon, 1995; Bird and Darwin, 1998). Culling and Darwin (1994) suggested that beating between closely spaced harmonic components introduces a temporal modulation in the concurrent-vowel stimulus, whereby one constituent vowel may be more identifiable than the other at particular moments within the modulation cycle. However, this theory primarily deals with improvements in identification of concurrent vowel at very small (up to 1 semitone) F0 differences, whereas the current study deals with semitones differences greater than 1. For the larger semitone-differences, the beat frequency is likely to be too high to allow "glimpsing" of either of the vowels at different points in the modulation cycle. Therefore, it is not obvious that the "beating" explanation is appropriate for the present results. Besides, our use of noise carriers, with their own inherent fluctuations, may limit the audibility of whatever beats remains in the unprocessed stimuli.

In the current study, it is more likely that listeners were able to use the F0 information derived from the well-resolved unprocessed region to aid in the grouping of channels in the envelope-vocoder processed region. With competing speech at the sentence level, it is likely that at any given time one voice will dominate the other. Consequently, accurate F0 information may guide the listeners to attend during times when the target signal is the strongest. Experiment 1 showed that SRT decreased more in competing-talker than in speech-shaped noise when unprocessed low-frequency information was added, consistent with this hypothesis. Taken together, the findings of experiment 1 and 2 suggest that reception benefits of residual hearing can be, at least in part, attributed to an increase in F0 representation.

4.4.3 Implication for long-term exposure

The EAS simulation used in this study (particularly LF_{600}+EV_{4-8}, with 5 vocoder channels) was intended to simulate the sound processing received by AES recipients (receiving only 6 channels of electrical stimulation) in the Turner *et al.* (2004) study. Given that we did not simulate any hearing loss in the low frequency region, common to most implant recipients, we anticipated comparable, if not greater, benefits associated with low-frequency hearing than those observed in the Turner *et al.* study. We were, therefore, surprised that while Turner *et al.* showed that with the addition of residual hearing, competing-talker was less detrimental to speech reception than noise, we found that adding unprocessed low-frequency information only reduced the detrimental effect of competing-talker to be comparable to that of speech-shaped noise. A possible explanation for this discrepancy may be long-term adaptation. Implant users have reported that over time the speech delivered through their implant processors sounds more 'natural' than when initially using the implant. It has been shown that speech reception performance of implant users improves over months of use (Dorman and Loizou, 1997; Fu *et al.*, 2002). The design of our current study does not capture the potential long-term adaptation to unusual patterns of stimulation (i.e. the combination of "tonal" low-frequency information with "raspy" high-frequency information). The EAS participants in Turner *et al.* (2004) study wore their devices for at least 12 month before data collection. It is reasonable to suppose that with long-term practice, listeners would perform at levels higher than reported here. The portable hearing-loss and prosthesis simulator being developed by Sensimetrics (Somerville, MA) may be an ideal tool for such a long-term adaptation study.

4.5 Summary

- While adding unprocessed information below 600 Hz did not return speech reception performance to the levels seen in the unprocessed conditions, SRTs did improve by 5.9 dB in the presence of a competing talker and by 4.0 dB in the presence of speech-shaped noise, as compared with the 8-channel vocoder.
- In the concurrent vowel experiment, significant ΔF0 benefits are observed even when an extremely limited range (<300 Hz) of low-frequency information was added.
- While speech-reception benefits could be attributed to increased spectral resolution, the findings of the current study suggest that it can be, at least in part, attributed to an increase in F0 representation.
- Results of the current study came from acoustic simulations of EAS users, where perfect residual hearing was assumed. Care should be taken when discussing the implications of the current findings in terms of implications for implant users.
- However, the findings of the present study lead us to be cautiously optimistic about the combined use of electric and acoustic stimulation (EAS) to enhance the perceptual segregation of speech.

Chapter 5: Summary and perspective

5.1 Thesis summary

The aim of this dissertation was to examine the role of voice pitch in speech perception in the presence of complex interference, specifically as it relates to effects of envelope-vocoder style implant processing. While acoustic envelope-vocoders stimulators have been used to examine the speech reception potential of implant users without the presence of interference and in the presence of steady-state noise, the effect of more complex maskers (e.g. fluctuating interference and competing speech) had not yet been investigated. Therefore, the first study (**Chapter 2**) examined the effects of envelope-vocoding on speech reception in complex situations. The central finding of this study was that under envelope-vocoder processing single-talker interference was more detrimental to speech reception than steady-state noise, while the reverse is true without processing. It was suggested that the difficulties experienced under vocoder processing in the presence of complex maskers might be related to the relatively weak pitch percept elicited by temporal envelope cues, and to the inability of listeners to use those cues to segregate the target from maskers.

To quantify the effects of envelope-vocoder sound processing on pitch perception (**Chapter 3**) listeners' F0 difference limens (F0DLs) and their ability to make use of F0 differences in a concurrent-vowel paradigm were measured. The key finding of this study was that despite the reasonable F0DLs (< 1 semitone or < 6% ΔF0) with 24- and 8-channel vocoder processing, listeners were unable to benefit from F0 differences between the competing vowels in a concurrent-vowel paradigm. This suggests that the mechanisms involved in using ΔF0 to segregate concurrent voices may be different from those involved in discriminating sequential differences in F0. More importantly, the findings suggest that implant users are unlikely to be able to use the F0 cues presented via envelope periodicity for speech segregation, and thus support the conclusion drawn in Chapter 2.

As cochlear implant technology improves, individuals with low-frequency residual hearing are electing to be implanted. These individuals are potentially able to use an implant and an acoustic hearing aid to form a hybrid electric-plus-acoustic system (EAS). While the residual acoustic hearing present in these individuals is likely to be too low in frequency to contribute directly to speech intelligibility, the additional low-frequency information may provide sufficient fine-structure cues to aid in source segregation. The aims of the third study (**Chapter 4**) were to examine the extent to which the benefits of residual hearing may be due to the increased salience of voice pitch and to investigate the amount of residual hearing necessary to see tangible benefits. The chief finding of this study was that significant ΔF0 benefits can be observed when unprocessed low-frequency information is added to the envelope-vocoder processed speech, even with an extremely limited range (<300 Hz) of low-frequency information. While adding low-frequency information to envelope-vocoder processed speech did not return speech reception to normal-hearing levels, it was able to improve the speech reception substantially beyond that of standard envelope-vocoder processing.

In summary, this dissertation shows that the poor speech performance experienced by implant users in complex backgrounds is likely to be, at least in part, due to the lack of salient voice pitch cues. In addition, F0 information encoded in the temporal envelope may not provide a sufficiently salient cue for the segregation of speech. Finally, adding low-frequency fine-structure information to envelope-vocoder processed high-frequency information returns some F0 segregation benefits and improves the reception of speech in complex backgrounds. Taken as a whole, the dissertation suggests that fine-structure F0 information is important to the task of speech segregation, and that every effort should be made to present such information to cochlear implant users.

5.2 Potential future research

Over the years, the criterion for implantation has become more lenient, expanding the implant candidate population pool to include people with more and more residual hearing. This dissertation points out important issues regarding the importance of preserving residual hearing and raises a number of interesting questions.

5.2.1 Compare EAS versus bilateral implantation

As mentioned in the conclusion of Chapter 4, a number of researchers have begun to experiment with bilateral implantation in the hope of gaining additional advantages (e.g., sound localization, increased channel numbers, and reduced electrode interaction) to improve speech reception performance of implant users in noise. Given the findings (Chapter 4) that unprocessed low-frequency information, when combined with envelope-vocoder processed high-frequency information, can yield greater speech reception with background interference, a systematic comparison between the reception performance of bilateral implant users and users of electric plus acoustic systems (EAS) may be warranted.

Bilateral implants are thought to have the potential of providing users with head-shadow advantages, binaural squelch, binaural summation effects, and (if implant processors are synchronized) sound localization as well. How might the performance of bilateral implant user compare to those of EAS users in a competing talker environment and in varying degrees of reverberation? How do users of synchronized bilateral implants compare with EAS users?

5.2.2 Preoperative predictors for the usefulness of residual hearing

Naturally, decisions concerning the implantation of an ear with residual hearing should take into consideration the potential usefulness of any residual acoustic hearing. As mentioned in Chapter 4, the EAS simulation used was designed to explore the maximum potential benefit of adding residual hearing to implant processing. It, therefore, assumed not only perfect residual hearing (i.e. no broadening of the auditory filters or threshold shift), but also no misalignment between electrode place and analysis band frequency. This, of course, is unlikely to be true with real implant candidates. Simulations of psychophysical aspects of hearing loss [e.g., threshold shift, broadening of auditory filters (Glasberg and Moore, 1986; Moore and Glasberg, 1986), abnormal growth in loudness (Moore *et al.*, 1985a), and dead regions (Moore *et al.*, 2000; Moore and Alcantara, 2001; Vickers *et al.*, 2001)] in combination with simulated implant processing, may be instructional in determining means of predicting the potential usefulness of varying degrees of residual hearing.

References

ANSI. **(1997)**. "Methods for calculation of the speech intelligibility index," **ANSI S3.5-1997**,

Arehart, K. H. **(1994)**. "Effects of harmonic content on complex-tone fundamental-frequency discrimination in hearing-impaired listeners," J. Acoust. Soc. Am. **95**, 3574-3585.

Arehart, K. H., King, C. A., and McLean-Mudgett, K. S. **(1997)**. "Role of fundamental frequency differences in the perceptual separation of competing vowel sounds by listeners with normal hearing and listeners with hearing loss," J. Speech Lang. Hear. Res. **40**, 1434-1444.

Assmann, P. (1999). Fundamental frequency and the intelligibility of competing voices. Proc.14th Int. Cong. of Phonetic Sci. 179-182.

Assmann, P. F., and Summerfield, Q. **(1990)**. "Modeling the perception of concurrent vowels: Vowels with different fundamental frequencies," J. Acoust. Soc. Am. **88**, 680-697.

Assmann, P. F., and Summerfield, Q. **(1994)**. "The contribution of waveform interactions to the perception of concurrent vowels," J. Acoust. Soc. Am. **95**, 471-484.

Bacon, S. P., Opie, J. M., and Montoya, D. Y. **(1998)**. "The effects of hearing loss and noise masking on the masking release for speech in temporally complex backgrounds," J. Speech. Lang. Hear. Res. **41**, 549-563.

Baer, T., and Moore, B. C. J. **(1993)**. "Effects of spectral smearing on the intelligibility of sentences in the presence of noise," J. Acoust. Soc. Am. **94**, 1229-1241.

Bernstein, J. G., and Oxenham, A. J. (**2003**). "Pitch discrimination of diotic and dichotic tone complexes: harmonic resolvability or harmonic number?," J Acoust Soc Am **113**, 3323-3334.

Bird, J., and Darwin, C. J. (**1998**). "Effects of a difference in fundamental frequency in separating two sentences," in Psychophysical and Physiological Advances in Hearing, edited by A. R. Palmer, A. Rees, A. Q. Summerfield and R. Meddis (Whurr, London).

Blamey, P. J., Dowell, R. C., Tong, Y. C., Brown, A. M., Luscombe, S. M., and Clark, G. M. (**1984**). "Speech processing strategies using an acoustic model of a multiple-channel cochlear implant," J. Acoust. Soc. Am. **76**, 104-110.

Bregman, A. S. (**1990**). Auditory Scene Analysis: The Perceptual Organisation of Sound (Bradford Books, MIT Press, Cambridge, Mass.).

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

Brungart, D. S. (**2001**). "Informational and energetic masking effects in the perception of two simultaneous talkers," J. Acoust. Soc. Am. **109**, 1101-1109.

Burns, E. M., and Viemeister, N. F. (**1976**). "Nonspectral pitch," J. Acoust. Soc. Am. **60**, 863-869.

Burns, E. M., and Viemeister, N. F. (**1981**). "Played again SAM: Further observations on the pitch of amplitude-modulated noise," J. Acoust. Soc. Am. **70**, 1655-1660.

Busby, P. A., Tong, Y. C., and Clark, G. M. (**1993**). "The Perception of Temporal Modulations by Cochlear Implant Patients," J. Acoust. Soc. Am. **94**, 124-131.

Carlyon, R. P. (**1996**). "Encoding the fundamental frequency of a complex tone in the presence of a spectrally overlapping masker," J. Acoust. Soc. Am. **99**, 517-524.

Carlyon, R. P., van Wieringen, A., Long, C. J., Deeks, J. M., and Wouters, J. (**2002**). "Temporal pitch mechanisms in acoustic and electric hearing," J. Acoust. Soc. Am. **112**, 621-633.

Ching, T. Y. C., Incerti, P., and Hill, M. (**2004**). "Binaural benefits for adults who use hearing aids and cochlear implants in opposite ears," Ear and Hearing **25**, 9-21.

Clark, G. M., Tong, Y. C., and Patrick, J. F. (1990). Cochlear Prostheses.

Culling, J. F., and Darwin, C. J. (**1993**). "Perceptual separation of simultaneous vowels: Within and across-formant grouping by F0," J. Acoust. Soc. Am. **93**, 3454-3467.

Culling, J. F., and Darwin, C. J. (**1994**). "Perceptual and computational separation of simultaneous vowels: Cues arising from low-frequency beating," J. Acoust. Soc. Am. **95**, 1559-1569.

Dai, H. P. (**2000**). "On the relative influence of individual harmonics on pitch judgment," Journal of the Acoustical Society of America **107**, 953-959.

Darwin, C. J., and Carlyon, R. P. (**1995**). "Auditory Grouping," in Hearing, edited by B. C. J. Moore (Academic Press, San Diego).

Darwin, C. J., and Culling, J. F. (**1990**). "Speech perception seen through the ear," Speech Comm. **9**, 469-475.

de Cheveigné, A., and Kawahara, H. (**2002**). "YIN, a fundamental frequency estimator for speech and music," J Acoust Soc Am **111**, 1917-1930.

de Cheveigné, A., Kawahara, H., Tsuzaki, M., and Aikawa, K. (**1997**). "Concurrent vowel identification. I. Effects of relative amplitude and Fo difference," J. Acoust. Soc. Am. **101**, 2839-2847.

Deeks, J. M., and Carlyon, R. P. (**2004**). "Simulations of cochlear implant hearing using filtered harmonic complexes: implications for concurrent sound segregation," J Acoust Soc Am **115**, 1736-1746.

Dorman, M. F. (**2000**). "Speech perception by adults," in Cochlear implants, edited by S. Waltzman and N. Cohen (Thieme New York, New York).

Dorman, M. F., and Loizou, P. C. (**1997**). "Changes in speech intelligibility as a function of time and signal processing strategy for an Ineraid patient fitted with continuous interleaved sampling (CIS) processors," Ear Hear **18**, 147-155.

Dorman, M. F., Loizou, P. C., Fitzke, J., and Tu, Z. (**1998**). "The recognition of sentences in noise by normal-hearing listeners using simulations of cochlear-implant signal processors with 6-20 channels," J. Acoust. Soc. Am. **104**, 3583-3585.

Dorman, M. F., Loizou, P. C., and Rainey, D. (**1997**). "Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs," J. Acoust. Soc. Am. **102**, 2403-2411.

Drullman, R. (**1995**). "Speech intelligibility in noise: Relative contribution of speech elements above and below the noise level," J. Acoust. Soc. Am. **98**, 1796-1798.

Durlach, N. I., Mason, C. R., Kidd, G., Arbogast, T. L., Colburn, H. S., and Shinn-Cunningham, B. G. (**2003**). "Note on informational masking (L)," Journal of the Acoustical Society of America **113**, 2984-2987.

Eddington, D. K. (**1980**). "Speech discrimination in deaf subjects with cochlear implants," J Acoust Soc Am **68**, 885-891.

Eisenberg, L. S., Dirks, D. D., and Bell, T. S. (**1995**). "Speech recognition in amplitude-modulated noise of listeners with normal and listeners with impaired hearing," Journal of Speech and Hearing Research **38**, 222-233.

Faulkner, A., Rosen, S., and Smith, C. (**2000**). "Effects of the salience of pitch and periodicity information on the intelligibility of four-channel vocoded speech: Implications for cochlear implants," J. Acoust. Soc. Am. **108**, 1877-1887.

Festen, J. M., and Plomp, R. (**1990**). "Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing," J. Acoust. Soc. Am. **88**, 1725-1736.

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

Freyman, R. L., Helfer, K. S., McCall, D. D., and Clifton, R. K. (**1999**). "The role of perceived spatial separation in the unmasking of speech," J. Acoust. Soc. Am. **106**, 3578-3588.

Friesen, L. M., Shannon, R. V., Baskent, D., and Wang, X. (**2001**). "Speech recognition in noise as a function of the number of spectral channels: Comparison of acoustic hearing and cochlear implants," J. Acoust. Soc. Am. **110**, 1150-1163.

Fu, Q. J., Shannon, R. V., and Galvin, J. J., 3rd. (**2002**). "Perceptual learning following changes in the frequency-to-electrode assignment with the Nucleus-22 cochlear implant," J Acoust Soc Am **112**, 1664-1674.

Fu, Q. J., Shannon, R. V., and Wang, X. S. (**1998a**). "Effects of noise and spectral resolution on vowel and consonant recognition: Acoustic and electric hearing," J. Acoust. Soc. Am. **104**, 3586-3596.

Fu, Q. J., Zeng, F. G., Shannon, R. V., and Soli, S. D. (**1998b**). "Importance of tonal envelope cues in Chinese speech recognition," J. Acoust. Soc. Am. **104**, 505-510.

Gantz, B. J., and Turner, C. W. (**2003**). "Combining acoustic and electrical hearing," Laryngoscope **113**, 1726-1730.

Geurts, L., and Wouters, J. (**2001**). "Coding of the fundamental frequency in continuous interleaved sampling processors for cochlear implants," J. Acoust. Soc. Am. **109**, 713-726.

Glasberg, B. R., and Moore, B. C. J. (**1986**). "Auditory filter shapes in subjects with unilateral and bilateral cochlear impairments," J. Acoust. Soc. Am. **79**, 1020-1033.

Glasberg, B. R., and Moore, B. C. J. (**1990**). "Derivation of auditory filter shapes from notched-noise data," Hear. Res. **47**, 103-138.

Glasberg, B. R., and Moore, B. C. J. (**1992**). "Effects of envelope fluctuations on gap detection," Hear. Res. **64**, 81-92.

Green, T., Faulkner, A., and Rosen, S. (**2002**). "Spectral and temporal cues to pitch in noise-excited vocoder simulations of continuous-interleaved-sampling cochlear implants," J. Acoust. Soc. Am. **112**, 2155-2164.

Gustafsson, H. A., and Arlinger, S. D. (**1994**). "Masking of speech by amplitude-modulated noise," J. Acoust. Soc. Am. **95**, 518-529.

Hartmann, W. M. (**1997**). <u>Signals, Sound, and Sensation</u> (Springer-Verlag, New York).

Hillenbrand, J., Getty, L. A., Clark, M. J., and Wheeler, K. (**1995**). "Acoustic characteristics of American English vowels," J. Acoust. Soc. Am. **97**, 3099-3111.

House, A. S. (**1960**). "Formant band widths and vowel preference," J. Speech. Hear. Res. **3**, 3-8.

House, A. S. (**1961**). "On vowel duration in English," J. Acoust. Soc. Am. **33**, 1174-1178.

Houtgast, T., Steeneken, H. J. M., and Plomp, R. (**1980**). "Predicting speech intelligibility in rooms from the modulation transfer function. I. General room acoustics," Acustica **46**, 60-72.

Houtsma, A. J. M., and Smurzynski, J. (**1990**). "Pitch identification and discrimination for complex tones with many harmonics," J. Acoust. Soc. Am. **87**, 304-310.

IEEE. (**1969**). "IEEE recommended practice for speech quality measurements," IEEE Trans. Aud. Electroacoust. **AU-17(3)**, 225-246.

Kaernbach, C., and Bering, C. (**2001**). "Exploring the temporal mechanism involved in the pitch of unresolved harmonics," J Acoust Soc Am **110**, 1039-1048.

Kessler, D. K. (**1999**). "The CLARION Multi-Strategy Cochlear Implant," Ann Otol Rhinol Laryngol Suppl **177**, 8-16.

Klatt, D. H. (**1980**). "Software for a cascade/parallel formant synthesizer," J. Acoust. Soc. Am. **67**, 971-995.

Kohlrausch, A., Fassel, R., and Dau, T. (**2000**). "The influence of carrier level and frequency on modulation and beat- detection thresholds for sinusoidal carriers," J. Acoust. Soc. Am. **108**, 723-734.

Kong, Y. Y., Stickney, G., and Zeng, F.-G. (**in press**). "Speech and melody recognition in binaurally combined acoustic and electric hearing," J. Acoust. Soc. Am.

Leek, M. R., and Summers, V. (**1993**). "Auditory filter shapes of normal-hearing and hearing-impaired listeners in continuous broadband noise," J. Acoust. Soc. Am. **94**, 3127-3137.

Levitt, H. (**1971**). "Transformed up-down methods in psychoacoustics," J. Acoust. Soc. Am. **49**, 467-477.

Litvak, L., Delgutte, B., and Eddington, D. (**2001**). "Auditory nerve fiber responses to electrical stimulation: Modulated and unmodulated pulse trains," J. Acoust. Soc. Am. **110**, 368-379.

Loizou, P. C. (**1999**). "Introduction to cochlear implants," IEEE Engineering in Medicine and Biology Magazine **18**, 32-42.

McKay, C. M., McDermott, H. J., and Clark, G. M. (**1994**). "Pitch percepts associated with amplitude-modulated current pulse trains in cochlear implantees," J. Acoust. Soc. Am. **96**, 2664-2673.

Meddis, R., and Hewitt, M. (**1992**). "Modeling the identification of concurrent vowels with different fundamental frequencies," J. Acoust. Soc. Am. **91**, 233-245.

Meddis, R., and O'Mard, L. (**1997**). "A unitary model of pitch perception," J. Acoust. Soc. Am. **102**, 1811-1820.

Miller, G. A., and Licklider, J. C. R. (**1950**). "The intelligibility of interrupted speech," J. Acoust. Soc. Am. **22**, 167-173.

Moore, B. C. (**2003a**). "Coding of sounds in the auditory system and its relevance to signal processing and coding in cochlear implants," Otol. Neurotol. **24**, 243-254.

Moore, B. C. J. (**1973**). "Frequency difference limens for short-duration tones," J. Acoust. Soc. Am. **54**, 610-619.

Moore, B. C. J. (**2003b**). An Introduction to the Psychology of Hearing, 5th Ed. (Academic, London).

Moore, B. C. J., and Alcantara, J. I. (**2001**). "The use of psychophysical tuning curves to explore dead regions in the cochlea," Ear and Hearing **22**, 268-278.

Moore, B. C. J., and Glasberg, B. R. (**1986**). "The role of frequency selectivity in the perception of loudness, pitch and time," in Frequency Selectivity in Hearing, edited by B. C. J. Moore (Academic, London).

Moore, B. C. J., and Glasberg, B. R. (**1988**). "Pitch perception and phase sensitivity for subjects with unilateral and bilateral cochlear hearing impairments," in Clinical audiology, edited by A. Quaranta (Laterza, Bari, Italy).

Moore, B. C. J., Glasberg, B. R., and Baer, T. (**1997**). "A model for the prediction of thresholds, loudness, and partial loudness," J. Aud. Eng. Soc. **45**, 224-240.

Moore, B. C. J., Glasberg, B. R., Hess, R. F., and Birchall, J. P. (**1985a**). "Effects of flanking noise bands on the rate of growth of loudness of tones in normal and recruiting ears," J. Acoust. Soc. Am. **77**, 1505-1515.

Moore, B. C. J., Glasberg, B. R., and Peters, R. W. (**1985b**). "Relative dominance of individual partials in determining the pitch of complex tones," J. Acoust. Soc. Am. **77**, 1853-1860.

Moore, B. C. J., Glasberg, B. R., and Shailer, M. J. (**1984**). "Frequency and intensity difference limens for harmonics within complex tones," J. Acoust. Soc. Am. **75**, 550-561.

Moore, B. C. J., Huss, M., Vickers, D. A., Glasberg, B. R., and Alcantara, J. I. (**2000**). "A test for the diagnosis of dead regions in the cochlea," British Journal of Audiology **34**, 205-224.

Moore, B. C. J., and Peters, R. W. (**1992**). "Pitch discrimination and phase sensitivity in young and elderly subjects and its relationship to frequency selectivity," J. Acoust. Soc. Am. **91**, 2881-2893.

Nelson, P. B., and Jin, S. H. (**2004**). "Factors affecting speech understanding in gated interference: Cochlear implant users and normal-hearing listeners," Journal of the Acoustical Society of America **115**, 2286-2294.

Nelson, P. B., Jin, S. H., Carney, A. E., and Nelson, D. A. (**2003**). "Understanding speech in modulated interference: Cochlear implant users and normal-hearing listeners," J Acoust Soc Am **113**, 961-968.

Nilsson, M., Soli, S., and Sullivan, J. (**1994**). "Development of the Hearing In Noise Test for the measurement of speech reception thresholds in quiet and in noise," J. Acoust. Soc. Am. **95**, 1085-1099.

Oxenham, A. J., and Moore, B. C. J. (**1997**). "Modeling the effects of peripheral nonlinearity in normal and impaired hearing," in Modeling Sensorineural Hearing Loss, edited by W. Jesteadt (Erlbaum, Hillsdale, NJ).

Patterson, R. D., Nimmo-Smith, I., Weber, D. L., and Milroy, R. (**1982**). "The deterioration of hearing with age: frequency selectivity, the critical ratio, the audiogram, and speech threshold," J. Acoust. Soc. Am. **72**, 1788-1803.

Peissig, J., and Kollmeier, B. (**1997**). "Directivity of binaural noise reduction in spatial multiple noise-source arrangements for normal and impaired listeners," J. Acoust. Soc. Am. **101**, 1660-1670.

Peters, R., Moore, B., and Baer, T. (**1998**). "Speech reception thresholds in noise with and without spectral and temporal dips for hearing-impaired and normally hearing people," J. Acoust. Soc. Am. **103**, 577-587.

Peters, R. W., and Moore, B. C. J. (**1992**). "Auditory filter shapes at low center frequencies in young and elderly hearing-impaired subjects," J. Acoust. Soc. Am. **91**, 256-266.

Plomp, R. (**1967**). "Pitch of complex tones," J. Acoust. Soc. Am. **41**, 1526-1533.

Qin, M. K., and Oxenham, A. (**2005**). "Effects of envelope-vocoder processing on F0 discrimination and concurrent-vowel identification," Ear & Hearing (in press).

Qin, M. K., and Oxenham, A. J. (**2003**). "Effects of simulated cochlear-implant processing on speech reception in fluctuating maskers," J Acoust Soc Am **114**, 446-454.

Remez, R. E., Rubin, P. E., Berns, S. M., Pardo, J. S., and Lang, J. M. (**1994**). "On the perceptual organization of speech," Psych. Rev. **101**, 129-156.

Ritsma, R. J. (**1967**). "Frequencies dominant in the perception of the pitch of complex sounds," J. Acoust. Soc. Am. **42**, 191-198.

Rosen, S., Faulkner, A., and Wilkinson, L. (**1999**). "Adaptation by normal listeners to upward spectral shifts of speech: Implications for cochlear implants," J. Acoust. Soc. Am. **106**, 3629-3636.

Rubenstein, J. T., Wilson, B. S., Finley, C. C., and Abbas, P. J. (**1999**). "Pseudospontaneous activity: Stochastic independence of auditory nerve fibers with electrical stimulation," Hear. Res. **127**, 108-118.

Scheffers, M. T. M. (**1983**). "Sifting vowels: Auditory pitch analysis and sound segregation," Ph.D. Thesis, Groningen University, The Netherlands.

Schouten, J. F. (**1940**). "The residue and the mechanism of hearing," Proc. Kon. Akad. Wetenschap. **43**, 991-999.

Shackleton, T. M., and Carlyon, R. P. (**1994**). "The role of resolved and unresolved harmonics in pitch perception and frequency-modulation discrimination," J. Acoust. Soc. Am. **95**, 3529-3540.

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

Shannon, R. V., Zeng, F. G., and Wygonski, J. (**1998**). "Speech recognition with altered spectral distribution of envelope cues," J. Acoust. Soc. Am. **104**, 2467-2476.

Smith, Z. M., Delgutte, B., and Oxenham, A. J. (**2002**). "Chimaeric sounds reveal dichotomies in auditory perception," Nature **416**, 87-90.

Steeneken, H. J. M., and Houtgast, T. (**1980**). "A physical method for measuring speech-transmission quality," J. Acoust. Soc. Am. **69**, 318-326.

Stevens, K. N. (**1998**). Acoustic Phonetics (MIT Press, Cambridge, MA).

Stickney, G., Zeng, F.-G., Litovsky, R., and Assmann, P. (**2004**). "Cochlear implant speech recognition with speech maskers," J Acoust Soc Am **116**, 1081 - 1091.

Stone, M. A., Glasberg, B. R., and Moore, B. C. J. (**1992**). "Simplified measurement of impaired auditory filter shapes using the notched-noise method," Brit. J. Audiol. **26**, 329-334.

Studebaker, G. (**1985**). "A "Rationalized" Arcsine Transform," Journal of Speech and Hearing Research **28**, 455-462.

Summerfield, A. Q., and Assmann, P. F. (**1991**). "Perception of concurrent vowels: effects of pitch-pulse asynchrony and harmonic misalignment," J. Acoust. Soc. Am. **89**, 1364-1377.

Summerfield, Q., and Culling, J. F. (**1992**). "Auditory segregation of competing voices: Absence of effects of FM or AM coherence," Philos. Trans. Roy. Soc. Lond. B **336**, 357-365.

Summers, V., and Leek, M. R. (**1998**). "F0 processing and the separation of competing speech signals by listeners with normal hearing and with hearing loss," Journal of Speech Language and Hearing Research **41**, 1294-1306.

ter Keurs, M., Festen, J. M., and Plomp, R. (**1992**). "Effect of spectral envelope smearing on speech reception," J. Acoust. Soc. Am. **91**, 2872-2880.

Terhardt, E. (**1974**). "Pitch, consonance, and harmony," J. Acoust. Soc. Am. **55**, 1061-1069.

Throckmorton, C. S., and Collins, L. M. (**2002**). "The effect of channel interactions on speech recognition in cochlear implant subjects: predictions from an acoustic model," J. Acoust. Soc. Am. **112**, 285-296.

Turner, C. W., Gantz, B. J., Vidal, C., Behrens, A., and Henry, B. A. (**2004**). "Speech recognition in noise for cochlear implant listeners: Benefits of residual acoustic hearing," Journal of the Acoustical Society of America **115**, 1729-1735.

Tyler, R. S., Parkinson, A. J., Wilson, B. S., Witt, S., Preece, J. P., and Noble, W. (**2002**). "Patients utilizing a hearing aid and a cochlear implant: Speech perception and localization," Ear and Hearing **23**, 98-105.

Vickers, D. A., Moore, B. C. J., and 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.

Vliegen, J., and Oxenham, A. J. (**1999**). "Sequential stream segregation in the absence of spectral cues," J. Acoust. Soc. Am. **105**, 339-346.

von Ilberg, C., Kiefer, J., Tillein, J., Pfenningdorff, T., Hartmann, R., Sturzebecher, E., and Klinke, R. (**1999**). "Electric-acoustic stimulation of the auditory system - New technology for severe hearing loss," Orl-Journal for Oto-Rhino-Laryngology and Its Related Specialties **61**, 334-340.

Wilson, B. S. (**1997**). "The future of cochlear implants," British Journal of Audiology **31**, 205-225.

Wilson, B. S., Finley, C. C., Lawson, D. T., Wolford, R. D., Eddington, D. K., and Rabinowitz, W. M. (**1991**). "Better speech recognition with cochlear implants," Nature **352**, 236-238.

Zeng, F.-G. (**2004**). "Trends in cochlear implants," Trends in Amplification **8**, 1-34.