Effects of cochlear implant processing and fundamental frequency on the intelligibility of competing sentences[a)]

Ginger S. Stickney[b)]
Hearing Instrument Consultants, 3090 Bristol Street, Suite 150, Costa Mesa, California 92626

Peter F. Assmann
School of Behavioral and Brain Sciences, University of Texas at Dallas, GR41, Box 830688, Richardson, Texas 75083

Janice Chang
Hearing and Speech Research Laboratory, University of California, Irvine, 364 Medical Surgery II, Irvine, California 92697-1275

Fan-Gang Zeng[c)]
Department of Otolaryngology—Head and Neck Surgery, University of California, Irvine, 364 Medical Surgery II, Irvine, California 92697-1275

(Received 15 October 2005; revised 17 May 2007; accepted 26 May 2007)

Speech perception in the presence of another competing voice is one of the most challenging tasks for cochlear implant users. Several studies have shown that (1) the fundamental frequency (F0) is a useful cue for segregating competing speech sounds and (2) the F0 is better represented by the temporal fine structure than by the temporal envelope. However, current cochlear implant speech processing algorithms emphasize temporal envelope information and discard the temporal fine structure. In this study, speech recognition was measured as a function of the F0 separation of the target and competing sentence in normal-hearing and cochlear implant listeners. For the normal-hearing listeners, the combined sentences were processed through either a standard implant simulation or a new algorithm which additionally extracts a slowed-down version of the temporal fine structure (called Frequency-Amplitude-Modulation-Encoding). The results showed no benefit of increasing F0 separation for the cochlear implant or simulation groups. In contrast, the new algorithm resulted in gradual improvements with increasing F0 separation, similar to that found with unprocessed sentences. These results emphasize the importance of temporal fine structure for speech perception and demonstrate a potential remedy for difficulty in the perceptual segregation of competing speech sounds. [DOI: 10.1121/1.2750159]

PACS number(s): 43.66.Ts, 43.71.Ky, 43.71.Bp, 43.66.Hg [KWG] Pages: 1069–1078

I. INTRODUCTION

The fundamental frequency (F0) of voiced speech, which determines the pitch of the voice, can be a useful cue for segregating competing speech sounds (Bregman, 1990). When two voices compete, it is easier to hear what one voice is saying if the competing voice has a different pitch, or occupies a different F0 range (Bird and Darwin, 1998; Brokx and Nooteboom, 1982; Darwin and Hukin, 2000). While normal-hearing listeners are capable of using differences in the pitch of the voice to improve their performance with competing speech sounds, there are many cochlear implant users who show no benefit when two competing sentences are spoken by talkers of different genders (Stickney *et al.*, 2004).

At present, speech coding strategies used by most cochlear implants encode only the slowly varying amplitude modulations of the speech wave form (the temporal envelope), while the temporal fine structure is discarded. Therefore, the F0 can only be conveyed by the temporal modulations. Although pitch can be conveyed by the temporal envelope (Burns and Viemeister, 1976), sounds that include the temporal fine structure evoke a stronger pitch percept than sounds that preserve only the temporal envelope (Oxenham *et al.*, 2004; Smith *et al.*, 2002). The lack of explicit encoding of the temporal fine structure is one reason that speech perception in the presence of other competing voices is such a challenging task for cochlear implant users (Qin and Oxenham, 2003; Stickney *et al.*, 2004; Zeng *et al.*, 2004).

Stickney *et al.* (2004) presented normal-hearing and cochlear implant listeners with competing sentences spoken by the same or different talkers. The normal-hearing listeners were presented with natural speech or a cochlear implant simulation. The simulation transmits amplitude modulations from a series of frequency bands, and within each frequency band, the amplitude modulation is applied to a white noise carrier (Shannon *et al.*, 1995). Stickney *et al.* demonstrated that normal-hearing listeners presented with an implant

[a)]Portions of this work were presented at the 148th Meeting of the Acoustical Society of America (2004) in San Diego, CA.
[b)]Electronic mail: gsstickney@yahoo.com;
[c)]Electronic mail: fzeng@uci.edu

simulation and cochlear implant users had as much difficulty when the competing talker was a female voice as when the talker was the same male voice as the target. In contrast, normal-hearing listeners presented with natural speech did not encounter these difficulties and instead showed an improvement of 50 percentage points at a 0 dB signal to noise ratio (SNR) with the female masker compared to the same male masker. These results demonstrate that with implant simulations that extract only the envelope information within each frequency band, listeners appear to be unable to take advantage of differences in voice F0 to segregate competing speech sounds.

Some of the earlier multichannel devices, such as Cochlear Corporation's Nucleus 22-electrode device, used the F0 to modulate a pulsatile carrier during voiced speech and spectral information from one or two formants to selectively stimulate a subset of electrodes with the greatest energy. The direct coding of F0 was eventually abandoned with the introduction of a new speech processing algorithm (Continuous Interleaved Sampling) that stimulated the electrodes sequentially to avoid potential current field interactions (Wilson *et al.*, 1991). In this algorithm, F0 information could be inherently conveyed by the temporal envelope, provided the carrier pulse rate and envelope cutoff frequency were sufficiently high (Geurts and Wouters, 2001). To better represent the instantaneous temporal envelope, Rubenstein *et al.* (1999) have recommended increasing the rate at which amplitude modulation information is cycled through the electrodes (i.e., a stimulation rate greater than 2000 Hz), analogous to increasing the sampling rate. They suggest that high-rate stimulation also has the potential to reintroduce more natural stochastic effects in auditory nerve responses (see Kiang *et al.*, 1965) that would allow for a greater dynamic range. The combination of higher rates and stochastic resonance may improve speech perception in implant users as a consequence of enhancing the representation of the temporal envelope. The temporal envelope however can only convey a weak representation of pitch.

How pitch information can be better represented in cochlear implant processing by additionally modifying the carrier frequency has therefore been of great interest in recent years. Green *et al.* (2004) compared the pitch labeling of processed diphthongal glides in normal-hearing and cochlear implant listeners. In one condition, the processing of stimuli for the normal-hearing listeners involved an implant simulation with a noise carrier. In a second condition, the carrier (a sawtooth-shaped wave form) included the periodicity of the vowel. They found that the carrier which additionally coded the periodicity information improved pitch labeling for both normal-hearing listeners presented with the modified implant simulation and cochlear implant users. While pitch perception was improved with the modified processing, a more recent study by the same group (Green *et al.*, 2005) found that formant frequency discrimination and vowel recognition were adversely affected, perhaps because the modified processing disrupted spectral cues.

Although the benefits of a periodic carrier were significant with the pitch labeling task, the amount of improvement with the addition of periodicity information was small. It could be that a pitch labeling task was not sufficient for demonstrating the true benefits provided by this additional cue. Speech perception tasks relying heavily on pitch information might have demonstrated a larger effect. In a study by Lan *et al.* (2004), an implant simulation was similarly modified to include F0 information for voiced segments of Mandarin Chinese tones. They found that the pitch patterns of four Mandarin tones were more accurately identified with the modified than traditional processing. Lan *et al.* also noted improved performance for phonemes, words, and sentences. The present study examines another type of modification to the cochlear implant simulation that codes F0 indirectly by extracting the temporal fine structure of sound.

A new signal processing algorithm has recently been developed to code the temporal fine structure by means of a frequency-modulated (FM) carrier. Speech recognition performance in the presence of a competing talker was examined using the new strategy (Frequency-Amplitude-Modulation-Encoding, FAME) and compared with an implant simulation using a tone-excited vocoder (Nie *et al.*, 2005; Stickney *et al.*, 2005). Details of the new algorithm are explained in the following (see Sec. II). In both studies, the target and masker sentences were spoken by different talkers, allowing for differences in voice pitch to be captured by FM. They found that the addition of the FM significantly improved performance relative to the standard simulation. With speech maskers, performance dropped by as much as 18 percentage points with the standard simulation relative to performance in quiet. In contrast, there was no significant drop in performance when a competing talker was added for the FAME processing that included FM. This result suggests that the listeners had access to additional cues with FM, most likely F0 information, which helped them segregate the two competing talkers.

The present study tests the hypothesis that FM, added to an implant simulation, can convey sufficient F0 information such that when the mean F0 of the masker is shifted relative to the target, there will be an improvement in speech recognition. Speech recognition was measured as a function of the F0 separation of the target and masking sentence (both of which were spoken by the same talker) over several semitones in normal-hearing and cochlear implant listeners. The normal-hearing listeners were presented with natural speech, the standard cochlear implant simulation, or the FAME processing. Because FAME codes both temporal envelope and temporal fine structure cues, F0 cues should be transmitted more effectively with FAME compared to the standard implant simulation to assist in perceptually segregating the competing sentences.

II. FREQUENCY-AMPLITUDE-MODULATION-ENCODING (FAME)

The new strategy (FAME) separately extracts the slowly varying amplitude (AM) and frequency (FM) modulations within each frequency band (see Fig. 1). The FM codes the temporal fine structure of the speech wave form, whereas the AM separately codes the temporal envelope. The instantaneous amplitude of the FM carrier frequency is determined from the temporal envelope in the corresponding band. The

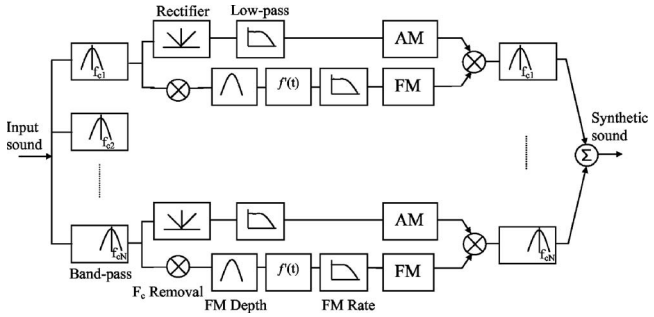

FIG. 1. Signal processing diagram for a 4-channel FAME processor.

signal is first divided into n narrowbands. The narrowband signals are then transmitted to separate AM and FM extraction pathways. The AM pathway involves full-wave rectification followed by low-pass filtering at 500 Hz to obtain the slowly varying envelope. The FM pathway involves removal of each narrowband's center frequency through phase-orthogonal demodulators (Flanagan and Golden, 1966) followed by low-pass filtering of the FM bandwidth (with a cutoff of 500 Hz) and rate (with a cutoff of 400 Hz). Limiting the FM rate is important since the eventual goal of a speech coding strategy, such as FAME, is to provide FM information that can be perceived by the majority of cochlear implant users.

As demonstrated by Chen and Zeng (2004), cochlear implant users' ability to detect a change in pitch for a frequency sweep or sinusoidal FM decreased as the standard frequency or modulation rate was increased. The delay between the AM and FM extraction pathways is adjusted prior to the combination of these two components within each subband and the signals are further bandpass filtered to remove frequency components introduced by AM and FM that fall outside of the original analysis filter's bandwidth. The waveforms from all bands are then summed to form the synthesized signal that contains the slowly varying AM and FM components.

III. EXPERIMENT

A. Methods

1. Listeners

The subjects were 49 young native English speakers, comprising undergraduate and graduate students. All subjects reported normal hearing. The subjects were recruited at the University of California, Irvine. There were seven subjects for each of the seven processing conditions. Additionally, seven cochlear implant users were recruited (see Table I for subject demographics). Subjects were compensated $10/h for their participation.

2. Test materials

Subjects were presented with IEEE sentences paired with a masker taken from the same set of sentences. All IEEE sentences in this study consisted of a subset of the 72 phonetically balanced lists of 10 sentences. The sentences were obtained from recordings by Hawley *et al.* (1999). The target sentences were spoken by a male voice (mean F0 = 108 Hz) in the presence of a different sentence spoken by the same male voice. The same masker sentence ("A large size in stockings is hard to sell") was presented on each trial to avoid confusion of the target and masker sentences.[1] The target and masker sentence had the same onset, but the masking sentence was always longer in duration. No sentences were repeated.

The stimuli were 70 sentences, with seven F0 conditions of 10 sentences each. Each sentence consisted of five keywords, for a total of 50 keywords per condition. There were six F0 conditions where the F0 contour of the masker sentence was shifted to a higher F0. The seventh condition included a natural speech masker where the F0 was neither estimated nor modified. Unlike most previous studies that have examined the effects of F0 difference on competing sentences using a steady-state (monotone) pitch, the natural F0 contour was preserved in the present study, but shifted upwards by n semitones from the average F0 measured across the entire sentence. A high-quality speech analysis-synthesis system called STRAIGHT (Kawahara, 1997, 1999) was used to estimate the F0 and resynthesize the sentence with a shifted F0. The estimated F0 contour (analyzed in 1 ms frames) was then replaced by one that was shifted up by a fixed amount compared to the average F0 of the original sentence, i.e., in each frame the measured F0 was replaced by $F0_{new} = 2^{n/12} * F0_{original}$, where $n = 0$, 3, 6, 9, 12, or 15 semitones. These conditions were labeled "semi3," "semi6,"…, and "semi15" corresponding to an F0 shift of 3, 6,…, and 15 semitones, respectively. The label semi0 represented the condition where the STRAIGHT algorithm was applied to the masker but the F0 contour was not raised, whereas the label

TABLE I. Subject demographics.

Subject	Age	Implant	Speech strategy	Duration of hearing loss (years)	Duration of deafness (years)	Duration of implant use (years)
CI1	46	Nucleus-22	SPEAK	12	12	11
CI2	69	Nucleus-24	ACE	7	7	6
CI3	70	Nucleus-24	ACE	40	14	3
CI4	61	Nucleus-22	SPEAK	52	14	12
CI5	78	Nucleus-24	CIS	35	13	1
CI6	68	Clarion CII	MPS	22	18	2
CI7	68	Clarion CII	MPS	62	58	5

"natural" was used to represent the condition where the masker was not processed with the STRAIGHT algorithm.

There were seven processing conditions. Conditions where the stimuli were not subjected to cochlear implant processing were labeled as "unprocessed." There were two conditions that used unprocessed speech: (1) unprocessed speech presented at a 0 dB signal-to-noise ratio (SNR) and (2) unprocessed speech presented at a 10 dB SNR. The remaining five conditions used standard implant simulations with the competing sentences at a 10 dB SNR. The SNR was adjusted after the application of the STRAIGHT algorithm. The SNR was calculated by first determining the rms of the unprocessed target and masker sentences (including silent periods), then scaling the masker and target to the same rms, and for the 10 dB SNR conditions, subsequently decreasing the rms of the masker sentence relative to the target sentence.

The implant simulation (AM or AM+FM) was applied after the target and masker were mixed. The AM+FM processing used the same algorithm as in the previous study by Nie *et al.* (2005). The combined target and masker stimuli were first pre-emphasized with a high-pass, first-order Butterworth filter with a cutoff frequency of 1.2 kHz. The sentences were then filtered into 4, 8, or 32 narrowbands using fourth-order elliptic bandpass. The AM and FM extraction was accomplished with fourth-order Bessel filters. The overall processing bandwidth was 80–8800 Hz. A sinusoidal carrier was used for both AM-only and AM+FM conditions. The AM-filter cutoff was set to 500 Hz, while the FM rate and depth were 400 and 500 Hz, respectively. However, for filters with bandwidths $<$500 Hz, the FM depth was set to be the same as the bandwidth of the filter. The implant simulation conditions were: (1) 4-channel AM-only processed speech; (2) 8-channel AM-only processed speech; (3) 32-channel AM-only processed speech; (4) 4-channel AM+FM-processed speech; and (5) 8-channel AM+FM-processed speech. The seven groups of normal-hearing subjects were presented with one of these conditions. The cochlear implant subjects were presented with only the unprocessed speech at a 10 dB SNR. Based on pilot data, the SNR for the implant simulation conditions was changed from 0 to 10 dB SNR. A 0 dB SNR was too difficult for several conditions with the AM-only processing and for the cochlear implant subjects.

3. Procedure

The stimuli were presented monaurally to the right ear through headphones (Sennheiser HAD 200), with subjects seated in an IAC sound booth. The level of the combined target and masker sentence was set to approximately 65 dB SPL, on average (Brüel & Kjær 2260 Investigator sound level meter; Brüel & Kjær Type 4152 artificial ear). After each stimulus was presented, subjects typed their responses using the computer keyboard and were encouraged to guess if unsure. Subjects were given as much time as needed to type their responses and were also given an opportunity to correct their spelling errors. Their responses were scored automatically based on the percentage of target sentence keywords correctly identified. Since all scoring was done with the computer program, no allowance was made for minor spelling errors.

Prior to testing, subjects were presented with two practice sessions. The first practice session presented ten unprocessed sentences in quiet. Subjects were to identify at least 85% of the keywords in order to participate in the study. No subjects were disqualified in this practice session. The second practice session consisted of seven sentences processed in the same manner as the test stimuli, with one sentence for each condition. This portion of the test was designed to simulate the test conditions so the subjects would know what to expect in the actual test session. No score was calculated for this practice session.

In the test session, each subject was presented with seventy sentences. There were ten sentences per condition for each subject. Each subject received a different set of sentences for each condition (digram-balanced Latin square design) to distribute the effects of sentence difficulty across the conditions. For example, subject 1 heard sentences 1–10 in the natural condition, while subject 2 heard the same sentences 1–10 in the condition with the masker's F0 contour shifted by 0 semitones. The order of the ten sentences in each set was randomized, as was the order of the conditions presented to the subject. Each test session lasted for approximately 20 min.

B. Results

1. Unprocessed speech

Figure 2 illustrates the wave form and corresponding F0 contours for the target and masker stimuli presented at a 0 dB SNR. The F0 contours were extracted from the sentences in isolation (i.e. prior to mixing) in 1 ms frames using the F0 estimation algorithm used by the STRAIGHT analysis system developed by Kawahara (1997). Because the stimuli shown here were not further processed with the algorithm that was used to create the cochlear implant simulations, they are referred to as "unprocessed." The panels from top to bottom show increasing differences in F0 between the target and masker sentence. Note that the F0 varies over time and that when the average F0s are the same for the target and masker, there are temporal intervals where they are well separated (i.e. the instantaneous F0 is different for the two sentences). Also, note that as the F0 shift increases, so does the size of this temporal interval where the instantaneous target and masker F0s are well separated.

Figure 3 shows the sentence recognition accuracy for the normal-hearing listeners presented with the unprocessed speech at the two SNRs. The x axis shows the F0 shift and the y axis shows the percentage of keywords correctly identified in that condition. It can be seen that at a SNR of 0 dB, performance improved as the F0 shift increased. In contrast, at a 10 dB SNR, no benefit was observed as performance was at ceiling. A mixed design analysis of variance (ANOVA) was performed with the two unprocessed speech conditions as the between-subjects factor and F0 shift as the within-subjects factor. The normal hearing listeners presented with the unprocessed speech showed higher perfor-

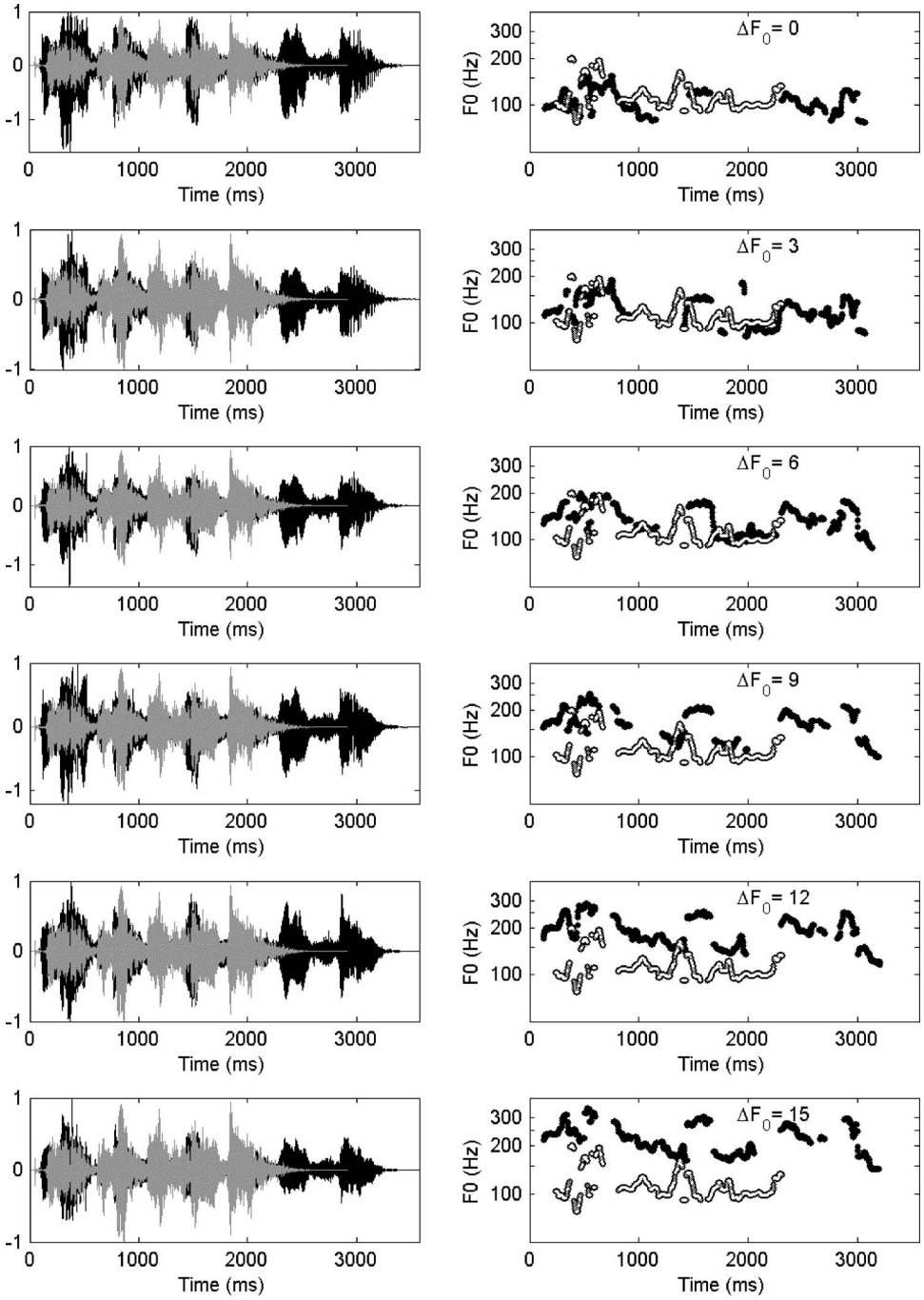

FIG. 2. Unprocessed wave forms and F0 contours for the target sentence and masker sentence. The target sentence is "The sheep were led home by a dog" and the masker sentence is "A large size in stockings is hard to sell." The wave form for the target sentence is shown in light gray and its F0 contour is represented as an unfilled line. The wave form for the masker sentence is shown in black and its F0 contour is represented as a solid black line. The F0 contours were extracted from unmixed signals that were scaled to the same rms and superimposed. The F0 contour for the masking sentence increases in frequency from the top panel to the bottom panel.

mance at the higher SNR of +10 dB [$F(1,12)=20.7, p<0.01$]. Target keyword identification generally improved as the F0 separation was increased from 0 to 6 semitones, but only when the rms level of the target and masker was matched (0 dB SNR). At a 0 dB SNR, performance improved by 12 percentage points from semi0 to semi6. However, this trend did not reach statistical significance when subjected to a Bonferroni adjustment for the six pairwise comparisons.

2. AM-only processed speech and cochlear implant performance

The left panels of Fig. 4 illustrate the estimated F0 contours for the same target and masker sentences as Fig. 2 with AM-only processing, estimated from the sentences prior to mixing but after processing. Notice that the F0 of the target and masker is relatively flat compared to the unprocessed F0 contour in Fig. 2. The sparseness of the F0 contours indicates that F0 is not as well represented in the processed versions, and that there appear to be more F0 estimation errors. What is also noticeable is that there is more overlap between the F0 components of the target and masker with AM-only processing. The reason for this greater overlap was likely due to the lack of explicit encoding of the temporal fine structure in the AM-only processed stimuli.

Figure 5 shows the results for the normal-hearing subject groups presented with 4, 8, or 32 AM-only processed channels. For purposes of comparison, the results from cochlear implant users and the normal-hearing subjects presented with the unprocessed sentences at a 10 dB SNR are

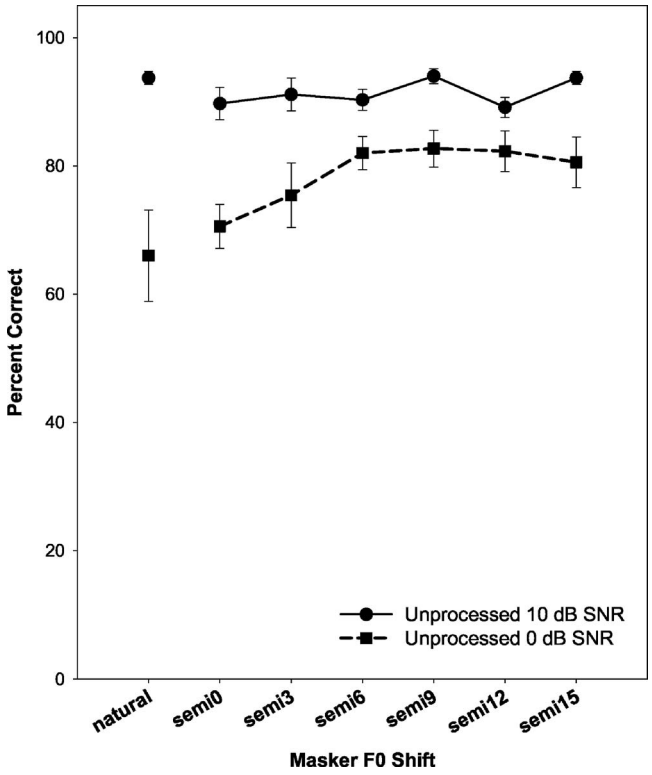

FIG. 3. Results for the normal-hearing subject groups presented with unprocessed speech at either a 0 dB SNR (square with dashed line) or 10 dB SNR (circle with solid line). The *x* axis shows each of the F0 shift conditions. The label "natural" represents the condition where the masker sentence was not processed by the STRAIGHT algorithm. The labels "semi0," "semi3," …, and "semi15" represent the conditions with an F0 shift of 0, 3, …, and 15 semitones, respectively. The error bars represent the standard error of the mean calculated from the scores of the 7 subjects within each group.

included in Fig. 5. What is most interesting about the data shown in Fig. 5 is that even though there is a dramatic difference in speech recognition scores as the number of channels is varied, there is relatively no change in performance as the F0 is shifted. A mixed design ANOVA was performed with the number of channels as the between-subjects factor and the F0 shift conditions as the within-subjects factor. Intelligibility improved dramatically as the number of channels was increased $[F(2,18)=179.3, p<0.001]$. Bonferroni pairwise comparisons showed significant improvements in performance from 4 to 8 and from 8 to 32 channels ($p<0.0125$), but not from 32 channels to the unprocessed speech. In addition, there was no significant difference in performance between the cochlear implant users and the normal-hearing listeners presented with 4-channel AM-only processed speech. Five separate ANOVAs, one for each processing group, were performed to determine whether or not there was a significant effect of F0 shift. The key finding was that for all the normal-hearing groups of subjects presented with AM-only processing and the cochlear implant users, performance as a function of F0 shift did not change. Even with 32 channels of envelope information listeners were not able to take advantage of the F0 difference between the target and masking sentence.

3. AM+FM*-processed speech*

The panels on the right side of Fig. 4 show the F0 contours of the target and masker for the AM+FM-processed speech. In contrast to the results obtained with AM-only processing (shown in the left panels), the AM+FM-processing preserves partial F0 contours. Note that as the F0 of the AM+FM-processed masker is increased, more of the F0 contour of the target sentence is revealed.

The results with AM+FM-processing are shown in Fig. 6. As observed with AM-only processing, speech recognition performance improved with more channels (i.e., the scores were higher with 8 AM+FM channels than with 4 AM +FM channels) $[F(1,12)=50.5, p<0.001]$. A comparison with Fig. 5 also shows that speech recognition performance with AM+FM processing was higher than with AM-only processing with the same number of channels. Last, Fig. 6 shows that speech recognition scores, at least for the 8-channel AM+FM group, tended to improve as the F0 shift was increased.

A mixed design ANOVA was used to compare AM and AM+FM performance. The F0 shift conditions were the within-subjects factor and the type of processing (four groups of subjects: 4-channel AM, 4-channel AM+FM, 8-channel AM, and 8-channel AM+FM) was the between-subjects factor. The results showed a significant effect of processing $[F(3,24)=50.35, p<0.001]$, F0 shift $[F(6,19)=3.12, p<0.05]$, and a significant interaction between the type of processing and the effect of the F0 shift on speech recognition performance $[F(18,54)=4.15, p<0.001]$. A *post-hoc* Scheffé analysis demonstrated significantly higher performance with the 4-channel AM+FM processing compared to the 4-channel AM-only processing, higher performance for the 8-channel AM-only processing compared to the 4-channel AM+FM processing, and the highest performance with 8-channel AM+FM processing ($p<0.05$ for all comparisons). For the 4-channel conditions, speech recognition scores, collapsed across all F0-shift conditions, were 30% for AM+FM and 13% for AM only. Similarly, for the 8-channel conditions, the scores were 57% and 45% for the AM+FM and AM conditions, respectively.

To examine the effect of the F0 shift on speech recognition with AM+FM processing, a separate mixed design ANOVA including only the AM+FM data and both channel conditions (4 and 8 channel) was performed. In contrast to the results obtained with AM-only processing, there was an improvement in performance as the F0 separation was increased $[F(6,7)=6.8, p<0.05]$. The interaction between the two channel conditions and the effect of F0 shift on speech recognition performance approached significance $[F(6,7)=3.1, p=0.08]$. For the 4-channel AM+FM group, there was much variability in the speech recognition scores as the F0 shift increased, obscuring any clear trend. In contrast, speech recognition performance for the 8-channel AM+FM group gradually improved with increases in F0 shift; the amount of improvement relative to the unshifted F0 (i.e., semi0 condition) was as much as 20 percentage points with an F0 shift of 12 semitones.

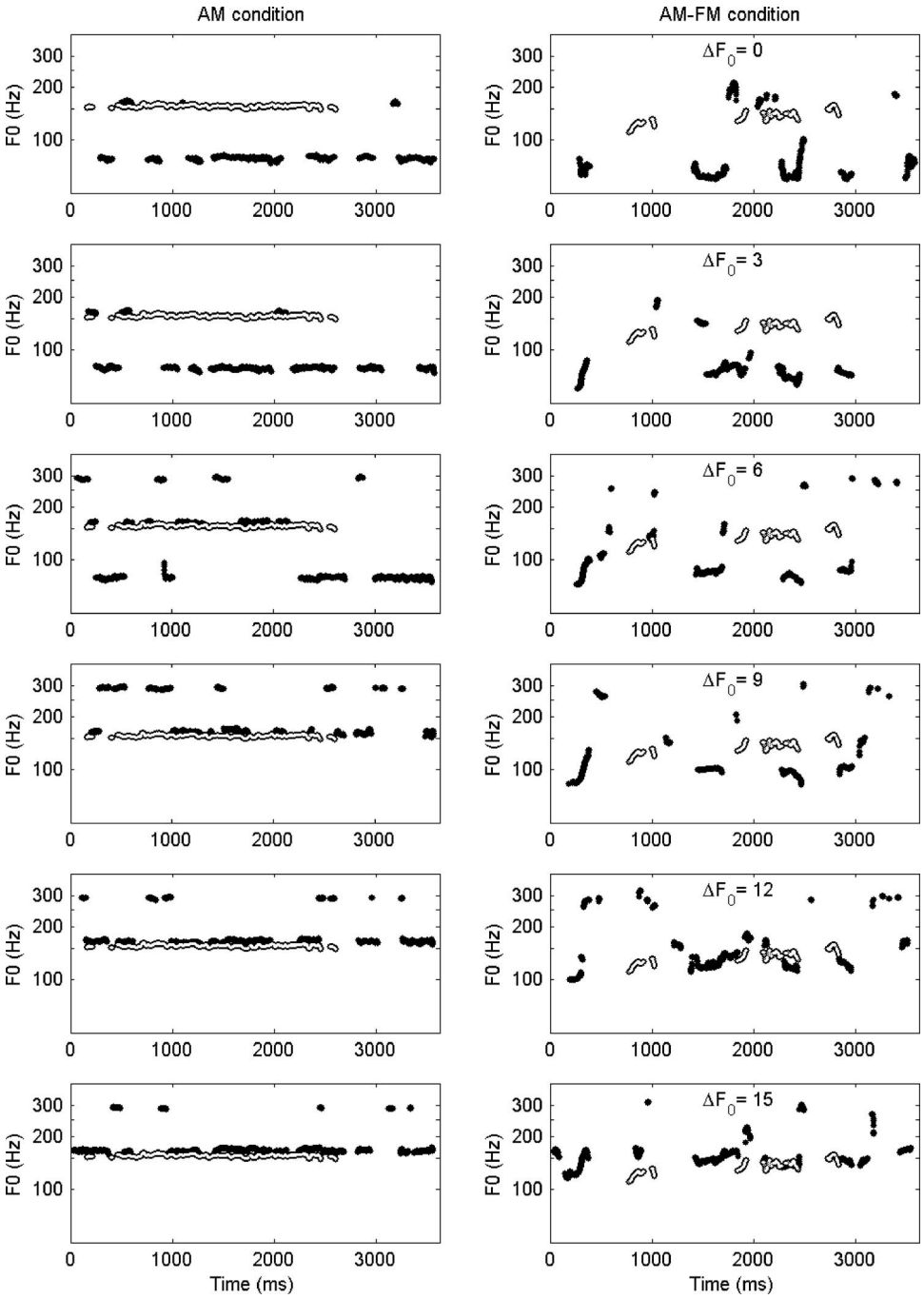

FIG. 4. The 8-channel implant simulation wave forms and F0 contours. The target sentence is "The sheep were led home by a dog" and the masker sentence is "A large size in stockings is hard to sell." The F0 contour of the target sentence is represented as an unfilled line, whereas that of the masker sentence is represented as a solid black line. The F0 contours were extracted from unmixed signals that were scaled to the same rms and superimposed. From the top to the bottom panel the F0 contour for the masker sentence increases.

IV. DISCUSSION

Users of cochlear implants have great difficulty understanding speech in the presence of one or more competing speech sounds. The results of the present study suggest that this may be due, in part, to a lack of F0 information provided by their speech processing algorithm. Several studies have demonstrated that normal-hearing listeners can take advantage of differences in voice characteristics (including the F0) between two competing talkers when presented with natural speech (Bird and Darwin, 1998; Brungart, 2001; Qin and Oxenham, 2003; Stickney *et al.*, 2004). However, when presented with strictly envelope-extracting implant simulations, even with as many as 24 channels, normal-hearing listeners did not benefit from these differences (Qin and Oxenham, 2003). Qin and Oxenham showed that although the 24-channel condition produced similar temporal envelopes as the natural speech, the differences in speech recognition thresholds for a female talker masking a male target sentence in the natural speech condition compared to the 24-channel condition were astounding: a better speech reception threshold of −11.3 dB for natural speech compared to the significantly poorer threshold of 0.6 dB for the 24-channel condition. Furthermore, Stickney *et al.* (2004) demonstrated results from actual cochlear implant users that were quite comparable to the normal-hearing listeners presented with, at most, 8 temporal envelope channels. Statistically there was no significant improvement from using the same male talker as masker and target to using a male talker as target and a female talker as a masker. Together, the results suggest that even with a relatively large number of channels (e.g., 24)

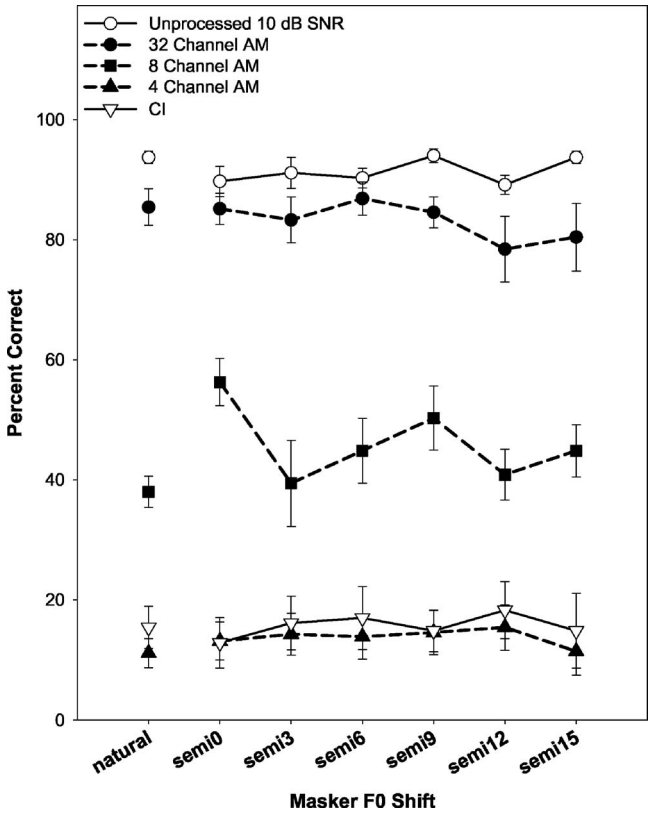

FIG. 5. Results for the normal-hearing subject groups presented with the AM-only processed speech at a 10 dB SNR (closed symbols with dashed lines). Results from the cochlear implant users and normal-hearing listeners presented with unprocessed speech at a 10 dB SNR are included for comparison (open symbols with solid lines). The *x* axis shows each of the F0 shift conditions. The label "natural" represents the condition where the masker sentence was not processed by the STRAIGHT algorithm. The labels "semi0," "semi3," …, and "semi15" represent the conditions with an F0 shift of 0, 3, …, and 15 semitones, respectively. The error bars represent the standard error of the mean calculated from the scores of the 7 subjects within each group.

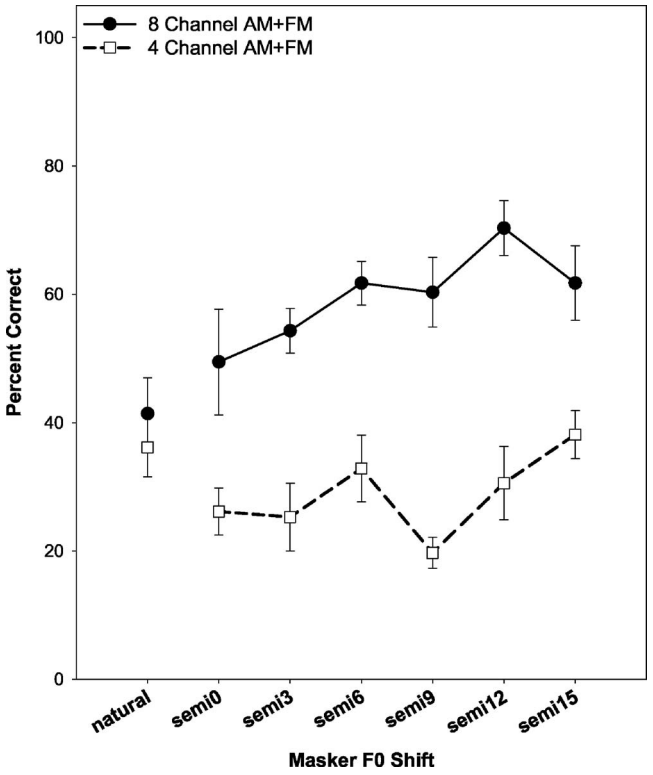

FIG. 6. Results for the normal-hearing subject groups presented with the 8-channel AM+FM-processed speech (closed circles and solid lines) or the 4-channel AM+FM speech (open squares and dashed lines) at a 10 dB SNR. The label "natural" represents the condition where the masker sentence was not processed by the STRAIGHT algorithm. The labels "semi0," "semi3," …, and "semi15" represent the conditions with an F0 shift of 0, 3, …, and 15 semitones, respectively. The error bars represent the standard error of the mean calculated from the scores of the 7 subjects within each group.

cochlear implant listeners provided only with temporal-envelope information may be unable to access the cues that normal-hearing listeners use to successfully segregate competing speech sounds.

How much do differences in F0 contribute to speech recognition under these conditions? The answer to this question can be addressed by comparing the results of the current study with that of Stickney *et al.* (2004). The mean F0 of the female talker from the previous study was 219 Hz. This F0 value is very close to that of the semi12 condition of the present experiment, which is 216 Hz. However, in the earlier study, the masker and target sentence stimuli provided various differences in talker characteristics in addition to F0 differences that could allow listeners to improve their performance. A study by Darwin *et al.* (2003) showed that differences in vocal tract length, in addition to F0, can help to segregate two competing voices, although it contributes much less than a difference in F0. Changing the vocal tract length (simulated by scaling the spectral envelope in a vocoder, which effectively multiplies the frequencies of all formants by a scale factor) produces an audible difference in voice quality that can be used to help segregate the speech of two competing voices. The male-female difference is associated with an upward shift in the frequencies of the formants by about 15%–20% (Peterson and Barney, 1952). However, in the present study, the masker and target sentences were spoken by the same individual, and the STRAIGHT algorithm shifted only the F0, leaving the formants unchanged. If F0 was the primary contributor, then there should be little difference between the two studies when comparing the amount of improvement from the same male masker to female masker condition (previous study) with that from the semi0 to semi12 condition (present study).

Normal-hearing listeners presented with the 4-channel AM-only condition at a 10 dB SNR showed comparable levels of speech recognition performance between the two studies. Stickney *et al.* (2004) found no difference in performance when the target sentences, spoken by a male voice, were masked either by a female voice or the same male voice as the target sentence. Likewise, in the present study, there was very little difference in score when the F0 was shifted from 0 semitones (analogous to the same male talker condition of the previous study) to 12 semitones (analogous to the female talker condition of the previous study); the total percent change with an F0 shift of 0–12 semitones was only 2%. This result suggests that with 4-channel AM-only processing, listeners cannot benefit from differences in F0. Adding FM though led to a 16% improvement in score in the 4-channel condition. However, there was no clear trend (no gradual

increase or decrease in performance) when the F0 was shifted from 0 to 12 semitones, suggesting that the addition of FM in this condition can improve overall intelligibility but may not provide sufficient F0 information. It is therefore possible that with a limited number of channels (such as 4), the additional F0-related information conveyed by the FM was still not sufficient to mediate an increase in performance with increasing F0 separation. This issue was addressed by increasing the number of channels to 8.

The percent change for the 8-channel AM conditions was quite different between the two studies. In Stickney *et al.* (2004), introducing a difference in voice gender between the target and masker, though not showing a statistically significant improvement, did increase the speech recognition score by an average of 28 percentage points. In contrast, in the present study, the percent change in performance when the F0 shift was increased from 0 to 12 semitones was minimal. One interpretation of these findings is that, with the stimuli in the previous study, some listeners might have been able to utilize differences in talker characteristics conveyed by the enhanced representation of temporal envelope cues in the 8-channel condition compared to the 4-channel condition. These listeners might have used cues other than F0 conveyed in the temporal and spectral envelope to improve their score, such as cues associated with differences in speaking rate between the two talkers and relatively coarse spectral differences between the male and female talker. In the present study, when FM information was added to the 8-channel condition, there was an improvement in score (20%) with an F0 shift of 12 semitones, which was not found when only AM information was provided or when listeners were presented with a smaller number of temporal envelope channels, with or without FM. This demonstrates the added benefit of F0 information conveyed by FM for segregating competing speech sounds, but the listener must also have sufficient spectral resolution to make use of the FM cue for segregating competing speech sounds on the basis of pitch (Oxenham *et al.*, 2004). With fewer channels, the analysis filters are broader, and the spectrotemporal resolution may be impaired.

Interestingly, the cochlear implant listeners in Stickney *et al.* (2004) showed a similar pattern of results as the normal-hearing listeners presented with 8-channel AM-only information. Some of the cochlear implant users could benefit from the temporal envelope differences between the two talkers, improving their score by an average of 20 percentage points with the female talker compared to the same male talker. On the other hand, in the present study, there was only a 5% improvement as the F0 was shifted from 0 to 12 semitones. The cochlear implant users evaluated here had poorer levels of performance overall compared to the implant users in the previous study, and their pattern of results were more similar to that of the normal-hearing listeners presented with 4-channel than 8-channel AM-only information. Therefore, some cochlear implant users can benefit from differences in the temporal envelopes of two competing talkers but might not benefit from differences in voice pitch. Likewise, it may also be true that only the better performing cochlear implant users will be able to utilize the additional temporal fine structure information. This is a topic that is itself interesting and will need further study.

Together these findings illustrate the combined usefulness of temporal envelope and temporal fine structure cues for auditory stream segregation (Bregman, 1990). The temporal envelope can provide cues for speaking rate and loudness. The temporal fine structure, on the other hand, can provide cues for voice pitch and formant transitions. With the natural pitch contour, F0 helps, in part, by giving momentary differences in pitch that allow listeners to segregate the two voices (Assmann, 1999). Consistent with this idea, several studies have demonstrated that the natural F0 contour leads to higher performance than a flattened F0 (Assmann, 1999; Binns and Culling, 2005; Watson and Schlauch, 2005). Furthermore, Binns and Culling discovered that the normal F0 contour provided some benefit over a flattened F0 when the target speech was presented in the presence of speech-shaped noise, but this benefit was much more pronounced when there were two competing speech sounds. It is possible that the F0 contour as well as the formant transitions provide a gradual change in frequency that can help the listener better track the target sound. Both the F0 contour and formant transitions would therefore follow the Gestalt principle of good continuation, thereby providing important cues for auditory stream segregation (Bregman, 1990).

Several investigators have identified methods for improving the encoding of the temporal envelope by reducing neural and electrical-field interactions between channels (Wilson *et al.*, 1991), enhancing the modulation depth (Geurts and Wouters, 1999), and increasing the rate of stimulation across channels (Rubinstein *et al.*, 1999). The coding of the additional temporal fine structure cue is also under investigation (Green *et al.*, 2004, 2005; Lan *et al.*, 2004; Nie *et al.*, 2005; Zeng *et al.*, 2005). The parameters used for FM coding of temporal fine structure, described in the present study, can be perceived by users of cochlear implants. This was demonstrated in a study by Chen and Zeng (2004). In their study, cochlear implant users were asked to detect the greatest change in pitch when the target was either a sinusoidal FM or a frequency sweep. Both FM depth and FM rate were varied. For the frequency sweep, the difference limen for FM depth with a 1000 Hz standard (the highest standard frequency tested) was 361 Hz. For the sinusoidal FM also with a 1000 Hz standard, the difference limens for FM depth were 400 and 549.4 Hz for FM rates of 160 and 320 Hz, respectively. At lower standard frequencies, the difference limens were significantly better, indicating an upper limit for FM coding. The FM rate and depth used in the FAME strategy have therefore been limited to 400 and 500 Hz, respectively, to be within the range that is perceivable by cochlear implant users. The FM rate could then be used to vary the interpulse interval of the pulse train carrier of a cochlear implant. Other potential implementations are described in an earlier publication by Nie *et al.*, (2005). Implementation of the FAME strategy for actual cochlear implant users is under way, as is the development of similar speech processing strategies that code temporal fine structure information. While results from actual cochlear implant users

await the implementation of these algorithms, the results from simulation studies indicate that enhancement of existing temporal envelope information and the addition of temporal fine structure have the potential to provide cues that can help cochlear implant listeners in one of the most challenging listening tasks that they are faced with: understanding speech with competing speech sounds.

ACKNOWLEDGMENTS

We are very grateful to Jennifer Lo and Rabia Farooquee who assisted in processing the stimuli and conducting the listening experiments. Dr. KaiBao Nie developed the FAME algorithm and user interface for processing the sentences. The IEEE sentences were created by Dr. Lou Braida and recorded by Dr. Monica Hawley and Dr. Ruth Litovsky. The STRAIGHT algorithm was provided by Dr. Hideki Kawahara. This work was supported by a NIH Grant No. F32 DC05900 awarded to G.S.S. and NIH Grant No. 2R01DC02267 awarded to F.G.Z.

[1]Although different sentence pairs might require greater or smaller F0 shifts for equal intelligibility, the overall pattern of the results for the standard simulation or FAME processing should be similar, specifically little to no effect of F0 shift with the implant simulation and some benefit of F0 shift for FAME processing.

Assmann, P. F. (**1994**). "The role of formant transitions in the perception of concurrent vowels," J. Acoust. Soc. Am. **96**, 1–9.

Assmann, P. F. (**1999**). "Fundamental frequency and the intelligibility of competing voices," Proceedings of the 14th International Congress of Phonetic Sciences, San Francisco, August 1–8, pp. 179–182.

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

Binns, C., and Culling, J. F. (**2005**). "The role of fundamental frequency (F0) contours in the perception of speech against interfering speech," J. Acoust. Soc. Am. **117**, 2606–2607.

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), pp. 263–269.

Bregman, A. (**1990**). *Auditory Scene Analysis* (MIT, Cambridge, MA).

Brokx, J. P. L., and Nooteboom, S. G. (**1982**). "Intonation and the perception of simultaneous voices," J. Phonetics **10**, 23–26.

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.

Chen, H., and Zeng, F.-G. (**2004**). "Frequency modulation detection in cochlear implant subjects," J. Acoust. Soc. Am. **116**, 2269–2277.

Darwin, C. J., Brungart, D. S., and Simpson, B. D. (**2003**). "Effects of fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers," J. Acoust. Soc. Am. **114**, 2913–2922.

Darwin, C. J., and Hukin, R. W. (**2000**). "Effectiveness of spatial cues, prosody, and talker characteristics in selective attention," J. Acoust. Soc. Am. **107**, 970–976.

Flanagan, J. L., and Golden, R. M. (**1966**). "Phase vocoder," Bell Syst. Tech. J. **45**, 1493–1509.

Geurts, L., and Wouters, J. (**1999**). "Enhancing the speech envelope of continuous interleaved sampling processors for cochlear implants," J. Acoust. Soc. Am. **105**, 2476–2484.

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.

Green, T., Faulkner, A., and Rosen, S. (**2004**). "Enhancing temporal cues to voice pitch in continuous interleaved sampling cochlear implants," J. Acoust. Soc. Am. **116**, 2298–2310.

Green, T., Faulkner, A., Rosen, S., and Macherey, O. (**2005**). "Enhancement of temporal periodicity cues in cochlear implants: Effects on prosodic perception and vowel identification," J. Acoust. Soc. Am. **118**, 375–385.

Hawley, M. L., Litovsky, R. Y., and Colburn, S. H. (**1999**). "Speech intelligibility and localization in a multi-source environment," J. Acoust. Soc. Am. **105**, 3436–3448.

Kawahara, H. (**1997**). "Speech representation and transformation using adaptive interpolation of weighted spectrum: VOCODER revisited," Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Vol. **2**, pp. 1303–1306.

Kawahara, H., Masuda-Katsuse, I., and de Cheveigne, A. (**1999**). "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," Speech Commun. **27**, 187–207.

Kiang, N. Y. S., Watanabe, T., Tomas, E. C., and Clark, L. F. (**1965**). *Discharge Patterns of Single Fibers in the Cat's Auditory Nerve* (MIT, Cambridge, MA).

Lan, N., Nie, K., Gao, S. K., and Zeng, F.-G., (**2004**). "A novel speech processing strategy incorporating tonal information for cochlear implants," IEEE Trans. Biomed. Eng. **51**, 752–760.

Nie, K., Stickney, G. S., and Zeng, F.-G. (**2005**). "Encoding frequency modulation to improve cochlear implant performance in noise," IEEE Trans. Biomed. Eng. **52**, 64–73.

Oxenham, A., Bernstein, J. G., and Penagos, H. (**2004**). "Correct tonotopic representation is necessary for complex pitch perception," Proc. Natl. Acad. Sci. U.S.A. **101**, 1421–1425.

Peterson, G. E., and Barney, H. L. (**1952**). "Control methods used in a study of vowels," J. Acoust. Soc. Am. **24**, 175–184.

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

Rubinstein, 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.

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

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

Stickney, G. S., Nie, K., and Zeng, F.-G., (**2005**). "Frequency modulation added to implant simulations improves speech recognition in noise," J. Acoust. Soc. Am. **118**, 2412–2420.

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

Watson, P. J., and Schlauch, R. S. (**2005**). "Spectral contributions to intelligibility of sentences with flattened fundamental frequency," J. Acoust. Soc. Am. **117**, 2606.

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 (London) **352**, 236–238.

Zeng, F.-G., Nie, K. B., Stickney, G. S., Kong, Y.-Y., Vongphoe, M., Bhargave, A., Wei, C. G., Cao, K. (**2005**). "Speech recognition with amplitude and frequency modulations," *Proceedings of the National Academy of Science*, Vol. 102, pp. 2293–2298.