Contribution of frequency modulation to speech recognition in noise[a)]

Ginger S. Stickney,[b)] Kaibao Nie, and Fan-Gang Zeng[c)]
Department of Otolaryngology - Head and Neck Surgery, University of California, Irvine, 364 Medical Surgery II, Irvine, California 92697-1275

(Received 28 February 2005; revised 5 July 2005; accepted 13 July 2005)

Cochlear implants allow most patients with profound deafness to successfully communicate under optimal listening conditions. However, the amplitude modulation (AM) information provided by most implants is not sufficient for speech recognition in realistic settings where noise is typically present. This study added slowly varying frequency modulation (FM) to the existing algorithm of an implant simulation and used competing sentences to evaluate FM contributions to speech recognition in noise. Potential FM advantage was evaluated as a function of the number of spectral bands, FM depth, FM rate, and FM band distribution. Barring floor and ceiling effects, significant improvement was observed for all bands from 1 to 32 with the additional FM cue both in quiet and noise. Performance also improved with greater FM depth and rate, which might reflect resolved sidebands under the FM condition. Having FM present in low-frequency bands was more beneficial than in high-frequency bands, and only half of the bands required the presence of FM, regardless of position, to achieve performance similar to when all bands had the FM cue. These results provide insight into the relative contributions of AM and FM to speech communication and the potential advantage of incorporating FM for cochlear implant signal processing. [DOI: 10.1121/1.2031967]

PACS number(s): 43.64.Me, 43.66.Ts, 43.71.Ky [BLM] Pages: 2412–2420

I. INTRODUCTION

Current signal processing for cochlear implants allows adequate speech perception in quiet environments for most users. However, their speech recognition performance in more realistic settings, where interfering noise is common, is severely limited. The current multichannel cochlear implant utilizes 12–22 electrodes distributed along the scala tympani to transmit frequency information based on place coding. Typically, each electrode receives an amplitude-modulated pulse train representing the narrow-band temporal envelope of a sound from a particular frequency band. Amplitude modulations from low frequencies are delivered to apical electrodes and amplitude modulations from high frequencies are delivered to basal electrodes. Shannon *et al.* (1995) demonstrated that amplitude modulations from as few as three frequency bands are sufficient to support sentence recognition in quiet. This observation highlighted the importance of amplitude modulation in speech perception. However, more recent studies have shown that amplitude modulation is not sufficient to support speech recognition in noise (Dorman *et al.*, 1998; Friesen *et al.*, 2001; Nelson *et al.*, 2003), and the greatest perceptual difficulties arise when the noise is also speech (Qin and Oxenham, 2003; Stickney *et al.*, 2004).

One means of improving performance for speech in noisy backgrounds is for the listener to perceptually identify, group, and track acoustic segments belonging to the target speech. Although it is not known for certain what role formant transitions play in segregating speech sounds, it has been suggested that gradual changes in the pattern of formant peaks of either the target or masker (or both) might provide cues for grouping and subsequently tracking sounds belonging to one of the two talkers (Assmann, 1995; Bregman, 1990). An alternative suggestion is that one or more of the masker's formants could move into a different frequency range unoccupied by the frequency components of the target, offering a temporal window in which portions of the target speech might be glimpsed (Assmann, 1995). It has also been suggested that the pitch of the voice, specifically the f0 contour, might allow listeners to improve their performance by attending to the f0 of one voice while ignoring a competing voice (Darwin and Hukin, 2000).

The pitch of the voice can be conveyed by the temporal envelope, however this cue provides a relatively weak representation of pitch (Burns and Viemeister, 1976). Green *et al.* (2004) addressed this issue using a traditional cochlear implant simulation, which transmitted amplitude modulations from several frequency bands which modulate a white noise carrier (Shannon *et al.*, 1995). They examined the separate contributions of spectral and temporal cues to pitch by varying the number of bands (single band or four-band) and envelope cutoff frequencies (rates of 32 or 400 Hz), respectively. Subjects were asked to label a single glide (sawtooth or dipthong) as "rising" or "falling" in pitch for f0's of 146, 208, and 292 Hz. They noted that the simulation's spectral cues contributed very little to pitch perception and that the weaker temporal envelope cues were useful only at lower

[a)]Portions of this work were presented at the Pan-American/Iberian Meeting on Acoustics (2002) and the Conference on Auditory Implants and Prostheses (2003).

[b)]Electronic mail: stickney@uci.edu
[c)]Electronic mail: fzeng@uci.edu

pitches. This indicates that pitch information is not effectively coded either by the envelope modulation or by a spectrally based distribution of temporal envelopes, and the greatest detriment occurs at higher pitches approximating the f0 of a female voice.

More recently, Green *et al.* (2004) modified the carrier of the implant simulation to include the periodicity of the input vowel. Compared to the traditional simulation, which used a noise carrier, the carrier containing periodicity information significantly improved pitch labeling. Lan *et al.* (2004) conducted a similar study in which they modified a traditional implant simulation to extract and include f0 for voiced segments in addition to amplitude modulations to represent the temporal envelope. They found that normal-hearing listeners presented with the novel algorithm could more accurately identify the pitch patterns of four Chinese tones than with the traditional simulation; performance also improved for phrases and sentences. These results are encouraging and indicate that modulation of the carrier frequency in addition to the temporal envelope could improve speech recognition in cochlear implant users, and perhaps also in noise.

In a study by Nie *et al.* (2005), a traditional cochlear implant simulation, containing amplitude modulations (AM) of a sinusoidal carrier, was combined with an additional frequency modulation cue (FM) to represent a slowed down version of the original sound's temporal fine structure. The instantaneous frequency was slowed so that it may be more applicable to cochlear implants. With electric stimulation, increases in temporal pitch can be perceived with increases in stimulation rate only up to 500–1000 Hz. Beyond this upper limit, there is no perceived change in pitch (Chen and Zeng, 2004). Therefore, the instantaneous frequency information, coded by the FM rate, was restricted by passing it through a low-pass filter with a cutoff frequency of 500 Hz. Nie *et al.* demonstrated that the combined AM and FM cues provided better representations of not only pitch information, but also formant transitions. They also state that because of this, higher levels of performance could be attained in tasks that involve melody recognition, speaker identification, and speech recognition with other competing talkers. In their study, sentences were processed into a 2-, 4-, or 8-band AM or AM+FM implant simulation and presented to normal-hearing listeners at five target-to-masker levels (TMR), with the masker being a competing sentence. They showed that, overall, the additional FM cue improved performance relative to AM only, and by as much as 71% for the 8-band condition at a 5 dB TMR. They state that the additional FM cue helps the listener better segregate the envelope of the target separate from the masker (Nie *et al.*, 2005; Zeng *et al.*, 2005).

The following study extends the work of Nie *et al.* by (1) further examining the benefits provided by the additional FM cue and (2) investigating FM processing parameters most critical for sentence recognition with a competing talker. The first experiment directly compared the FM advantage for sentences presented in quiet or with a competing talker as a function of the number of bands from 1 to 32. Aside from demonstrating a significant benefit provided by the additional FM cue, it was hypothesized that the greatest differences between AM and AM+FM processing would occur when the number of spectral bands was small and that maximum performance would be observed for AM+FM processing with far fewer bands than with AM-only processing. Additionally, AM+FM processing, because it provides more information for speech tracking (e.g., cues to pitch and formant transitions), was hypothesized to show a greater FM advantage in noise than in quiet. The influence of FM processing parameters on speech recognition with a competing talker was examined in experiments 2–4. Experiment 2 examined sentence recognition as a function of FM depth for bandwidths of 50 or 500 Hz at a fixed FM rate of 400 Hz. Since formants can sweep over a wide range of frequencies, much more than a range of 50 Hz, it was hypothesized that performance would improve as the FM depth was increased from 50 to 500 Hz since wider bandwidths would best capture the full formant transition. Experiment 3 examined the effect of FM rate on sentence recognition performance by comparing two rates (50 and 400 Hz) at a fixed FM depth of 500 Hz. Both normal-hearing and cochlear implant listeners can perceive changes in frequency for rates of 300–500 Hz and this frequency range is sufficient for coding the pitch of both male and female voices (adults and children). It was hypothesized that FM rate could influence performance so long as it captured the pitch of the voices used in the experiment. In other words, performance would improve by increasing the FM rate from 50 to 400 Hz. Last, in experiment 4, speech recognition performance was measured using hybrid AM and FM conditions in which the FM cue was systematically added to a subset of bands from low to high frequency and vice versa. The parameters of interest were (1) the number of AM+FM bands needed to reach a performance plateau and (2) the frequency bands (high vs. low) where AM+FM showed the greatest benefit. It was hypothesized that FM information would provide the greatest benefit when added to low-frequency bands since the range of frequencies for these bands is more likely to be associated with the low FM rate, which was limited to only 400 Hz in this study. Furthermore, the most salient formant transitions can be conveyed by only a subset of FM bands, therefore performance should improve gradually as the number of FM bands is increased, eventually reaching a plateau when the formant pattern is adequately represented.

II. SIGNAL PROCESSING

Frequency modulation (FM) was used to code the instantaneous frequency, or temporal fine structure of the speech waveform, independently from its instantaneous amplitude. A diagram of this algorithm is shown in Fig. 1. The sound is filtered into n narrow bands. Each of the narrow bands is then subjected to an AM-extraction pathway and an FM-extraction pathway. The AM pathway obtains the slowly varying envelope, using full-wave rectification followed by a low-pass cutoff filter which controls the amplitude modulation rate. The FM pathway extracts the slowly varying frequency modulation. This is obtained by first removing each narrow band's center frequency through phase-orthogonal

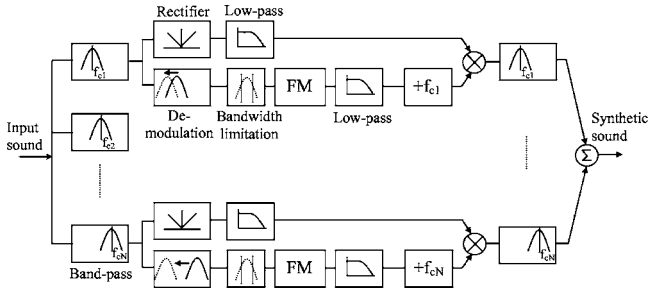

FIG. 1. Diagram of the speech processing strategy which combines AM and FM information. This speech processing strategy was used for the simulations.

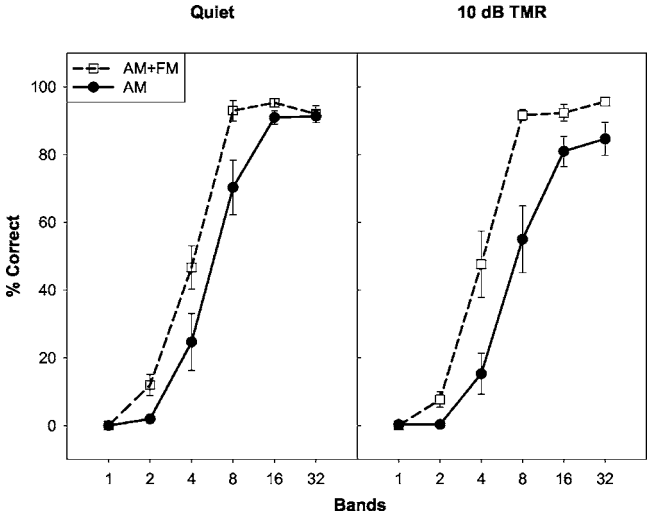

FIG. 2. Comparison of AM (filled circle, solid line) and AM+FM (unfilled squares, dashed lines) speech recognition performance in quiet (left panel) or with a competing talker at +10 dB TMR (right panel). Speech recognition performance in percent correct (y axis) is shown as a function of the number of bands (x axis). The error bars represent the standard error of the mean.

demodulators as used in implementing phase vocoders (Flanagan and Golden, 1966). This is followed by low-pass filtering to limit the FM depth and rate to relatively slowly varying FM components that can be potentially perceived by cochlear-implant users (Chen and Zeng, 2004). The delay between the two pathways is adjusted before combining the AM and FM components into a subband signal. The subband signal is then further bandpass filtered to remove frequency components that are introduced by AM and FM but fall outside the original analysis filter's bandwidth. Finally, the band-passed signals are summed to form the synthesized signal that contains only slowly varying AM and FM components around each analysis filter's center frequency.

In this study, the sentence stimuli were first preemphasized with a high-pass, first-order Butterworth filter with a cutoff frequency of 1.2 kHz. The sentences were then filtered into narrow bands 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. The AM cutoff filter was set to 500 Hz, while the FM rate and depth were manipulated in accordance with the aims of the specific experiment.

III. EXPERIMENT 1: SPEECH RECOGNITION WITH A SINGLE COMPETING TALKER

A. Methods

1. Listeners

A total of 24 normal-hearing listeners participated in this experiment. There were six subjects in each of four conditions. All subjects were native English speakers with no reported hearing loss. Subjects were recruited from the University of California, Irvine Social Science subject pool and received course extra credit for their participation.

2. Test materials

Sixty IEEE sentences (Rothauser *et al.*, 1969) were presented to the listeners, producing a total of ten sentences for each of the six conditions. The sentences consisted of five keywords, for a total of 50 stimuli per condition. Every subject received a different ordering of sentences for each condition according to a digram-balanced design. The sentences were spoken by a male talker (mean f0=108 Hz) either in quiet or in the presence of a competing sentence, which was spoken by a different talker of the same gender (mean f0 =136 Hz). The f0 values were estimated with the TEMPO program (Kawahara *et al.*, 1999). The competing sentence was "Port is a strong wine with a smoky taste." The target and masker sentences had the same onset, but the masking sentence was always longer in duration. No sentences were repeated.

3. Signal processing

The sentences were filtered into 1 to 32 narrow bands. In this experiment, the AM cutoff filter was set to 500 Hz. The FM rate was set to 400 Hz and the FM depth was set to 500 Hz or the critical bandwidth, whichever was narrower.

4. Procedure

The stimuli were presented in a sound-attenuated chamber monaurally through the right headphone. The target sentence was presented at an average rms level of 65 dB SPL. Prior to testing, subjects were asked to complete two practice sessions. The first presented natural sentences in quiet. A score of 85% or higher was required to continue with testing. The second practice session was used to familiarize the listener with the testing condition to which they were assigned. One group of 12 listeners heard the sentences in quiet and the second group heard the sentences masked by the competing talker at a 10 dB TMR. For each of these two groups of 12 listeners, six heard AM-processed stimuli and the other six heard the combined AM+FM stimuli. All subjects heard each of the six band conditions: 1, 2, 4, 8, 16, or 32 bands. Listeners were asked to type in the words of the target sentence into the computer and to guess if unsure. Each keyword was scored automatically with a MATLAB program.

B. Results

Figure 2 shows the results for AM and AM+FM speech recognition in quiet (left panel) and at a 10 dB TMR (right panel) as a function of the number of bands.[1] A mixed design

ANOVA was performed, with the type of processing and presence or absence of masking as between subject factors and the number of bands as the repeated factor. The results showed a main effect for the number of bands [$F(2,50)=441.20, p<0.001$], and Bonferroni-adjusted planned comparisons showed significant differences between all but the 16- and 32-band conditions. There was a strong effect for the type of processing [$F(1,20)=22.73, p<0.001$], with AM+FM processing producing sentence recognition scores that were on average 13% higher than AM-only scores. A significant interaction was also found between the type of processing and the number of bands [$F(5,16)=5.03, p<0.01$]. An analysis with each band condition showed that higher performance for AM+FM processing occurred in the 2-, 4-, 8-, and 16-band conditions but not in the 1- or 32-band conditions. To investigate the improvement plateau with more bands for each type of processing, separate ANOVAs were conducted for AM and for AM+FM conditions followed by Bonferroni-adjusted planned comparisons. The results showed that performance with AM processing improved from 2 to 16 bands, whereas AM+FM processing showed improved performance from 1 to 8 bands. Because of ceiling and floor effects in many of the band conditions, there were no significant main effects or interactions for the type of masking (i.e., target sentence presented in quiet or masked). However, an inspection of Fig. 2 for the mid-band conditions (e.g., 8 and 16 bands) shows that there was a greater drop in performance with the addition of noise with AM compared to AM+FM processing. With 8 bands, performance dropped by 18% with the addition of noise for AM processing, but there was no difference in performance for AM+FM processing. Similarly, with 16 bands, performance with AM processing dropped by 10% with the addition of noise, whereas AM+FM processing showed relatively little change in performance.

IV. EXPERIMENT 2: EFFECT OF FM DEPTH

A. Methods

1. Listeners

A second group of 24 listeners were recruited from the Social Science subject pool for experiment 2. All subjects reported normal hearing and were native English speakers.

2. Test materials

The same target and masking sentences from experiment 1 were used. The target sentence was presented in quiet or combined with the masking sentence at several target-to-masker ratios: 20, 15, 10, 5, and 0 dB.

3. Signal processing

In this experiment, the stimuli were processed into 4 or 8 bands. The FM depth (i.e., bandwidth) was set to 50 or 500 Hz, or the critical bandwidth, whichever was narrower. The FM rate (i.e., cutoff frequency) was fixed at 400 Hz.

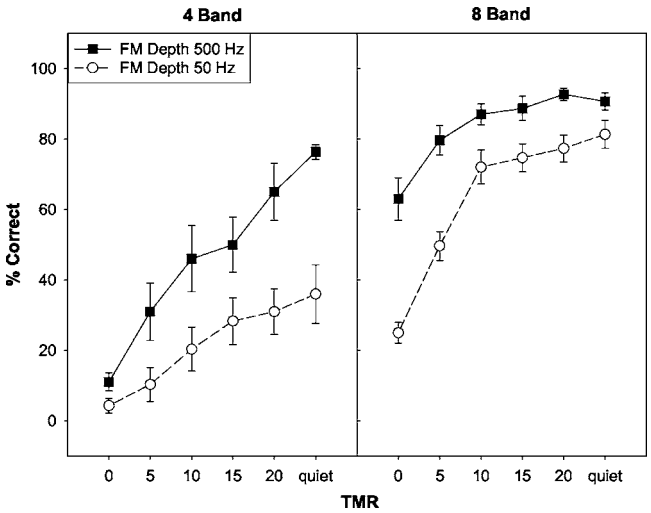

FIG. 3. The effects of FM depth on speech recognition performance (y axis) as a function of the TMR condition (x axis). Comparisons are made between a 500-Hz (filled squares, solid line) and a 50-Hz depth (unfilled circles, dashed line). In experiment 2, the FM rate was fixed at 400 Hz. Separate plots show the results for the 4- (left panel) and 8-band conditions (right panel).

4. Procedure

Four groups (2 band conditions $\times$ 2 FM depths) of six listeners each participated in the practice and test sessions. Each group heard speech processed into 4 or 8 bands with an FM depth of 50 or 500 Hz. All subjects heard the AM+FM sentences in quiet and at five TMRs. All other procedures were the same as experiment 1.

B. Results

Figure 3 shows speech recognition performance with a competing talker as a function of FM depth and TMR for sentences processed into 4 (left panel) or 8 bands (right panel). A mixed design ANOVA was performed with the number of bands and FM depth as between-subjects factors and the TMR condition as a within-subjects factor. There was a main effect of the number of bands [$F(1,20)=96.07, p<0.001$], TMR [$F(5,16)=75.86, p<0.001$], and FM depth [$F(1,20)=31.59, p<0.001$], and a significant interaction between these three factors [$F(5,16)=8.00, p<0.01$]. As expected, the 8-band condition produced higher scores than the 4-band condition, and scores improved with increasing TMRs. Of greater interest was the higher levels of performance attained with the 500-Hz depth (65.1%) than with the 50-Hz depth (42.5%). The three-factor interaction can be explained by greater differences between the two depths at high TMRs with 4 bands and at low TMRs with 8 bands, an outcome due to floor and ceiling effects, respectively.

V. EXPERIMENT 3: EFFECT OF FM RATE

A. Methods

1. Listeners

Twenty-four additional subjects participated in experiment 3. All criteria and recruiting procedures were the same as the previous experiments.

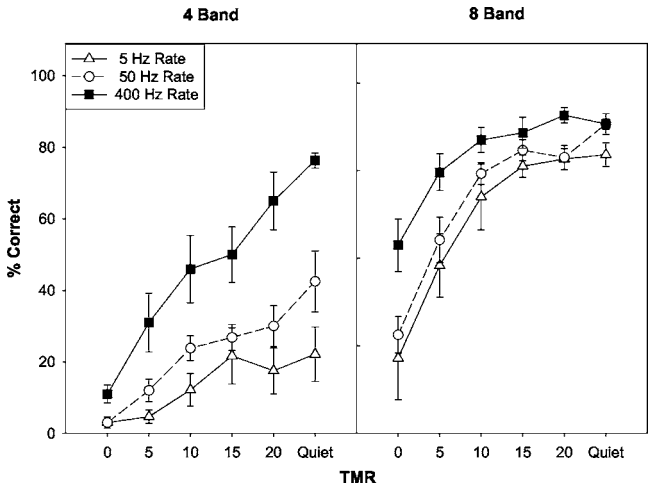

FIG. 4. The effects of FM rate on speech recognition performance (y axis) as a function of the TMR condition (x axis). Results are shown for the three rate conditions: 400 Hz (filled squares, solid line), 50 Hz (unfilled circles, dashed line), and 5 Hz (unfilled triangles, solid line). The FM depth was fixed at 500 Hz. Note that the data for the 400 Hz condition were replotted from experiment 2 (Fig. 2). Separate plots show the results for the 4- (left panel) and 8-band conditions (right panel).

2. Test materials

Experiment 3 included the same sentence stimuli and TMR conditions as experiment 2.

3. Signal processing

The only modification in experiment 3 from the previous experiment was that the FM rate was set to 5 or 50 Hz, and the FM depth was fixed at 500 Hz.

4. Procedure

Four groups (2 band conditions $\times$ 2 FM rates) of six listeners each participated in the practice and test sessions. All other procedures and conditions were the same as experiment 2.

B. Results

For comparison, the data from the 500-Hz depth, 400-Hz rate (from experiment 2) was included in the analysis. The results are shown in Fig. 4. A mixed design ANOVA was performed with FM rate and number of bands as between-subjects factors and TMR as the within-subjects factor. There was a main effect of TMR $[F(5,26)=45.91, p<0.001]$, number of bands $[F(1,30)=275.52, p<0.001]$, and FM rate $[F(2,30)=22.94, p<0.001]$. There was also a significant interaction among these three factors $[F(10,52)=2.08, p<0.05]$ and a significant interaction between the number of bands and FM rate $[F(2,30)=3.76, p<0.05]$. A simple effects analysis of each band condition showed that, with 4 bands, the 400-Hz rate produced higher performance than either the 50- or 5-Hz rates (Scheffé *posthoc*: $p<0.01$). In contrast, the 50-Hz and 400-Hz rate conditions produced equivalent performance in the 8-band condition, and both produced significantly higher performance than the 5-Hz rate (Scheffé *posthoc*: $p<0.05$).

VI. EXPERIMENT 4: EFFECTS OF THE LOCATION AND NUMBER OF FM BANDS

A. Methods

1. Listeners

Twenty-two additional subjects were recruited from the UCI Social Science subject pool.

2. Test materials

The target sentences were taken from the same corpus of IEEE sentences and the same masking sentence and talker were used from the previous experiments. Because of the large number of conditions, a new group of sentences were processed in addition to those used in the previous experiments. To reduce the number of conditions and avoid ceiling and floor effects, the masker sentence was combined with the target sentence at a 10 dB TMR for the 8-band group and at a 20 dB TMR for the 4-band group. Based on the results from experiment 1 and a pilot study, these TMRs avoid ceiling and floor effects for the 8- and 4-band conditions, respectively.

3. Procedure and signal processing

Two groups of twenty-two listeners (7 for the 4-band group and 15 for the 8-band group) participated in a practice and test session. The number of subjects varied in the two band groups because of the use of a digram-balanced Latin square design which uses the same number of subjects as conditions so that all conditions are received in a different order for each subject.

For the 4-band group, there were five conditions where FM information was added to a subset of the total number of bands: (1) AM+FM was on band 4 only; (2) bands 4, 3, and 2, hereafter referred to as 4–2; (3) band 1 only; (4) bands 1 and 2, hereafter referred to as 1–2; or (5) all 4 bands, hereafter referred to as 1–4. Remaining bands, if any, contained only AM. The higher the band number, the higher the frequencies coded within that band. The FM rate and depth were 400 and 500 Hz, respectively. In two additional conditions, performance was compared for all-AM-bands using either a noise carrier or sinusoidal carrier. This resulted in seven conditions total for the 4-band group.

For the 8-band group, AM+FM was on band 8 only, 8–7, 8–6, 8–5, 8–4, 8–2, 1 only, 1–2, 1–3, 1–4, 1–5, 1–6, or 1–8, with remaining bands consisting of AM information only. An all-AM-band comparison was also included using a noise or sinusoidal carrier, producing a total of 15 conditions for the 8-band group. All other procedures were the same as the previous experiments.

B. Results

Results for the 4-band data are shown in Fig. 5, with filled bars representing the all-AM condition (which used a sinusoidal or noise carrier), unfilled bars representing conditions where the FM information ranged from low- to high-frequency bands, and hatched bars representing conditions

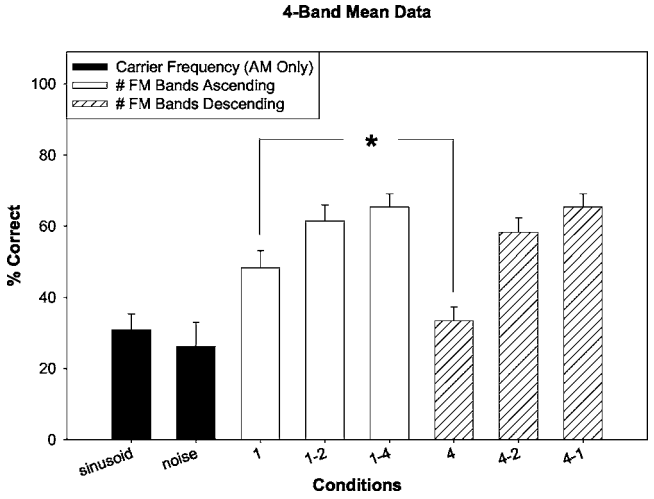

FIG. 5. Results from a 4-band hybrid AM and FM simulation. The *y* axis shows the conditions, with numbers representing the frequency band(s) containing FM information. Unfilled bars represent conditions with FM information added from low to high frequency bands. Hatched bars represent conditions with FM information added from high- to low-frequency bands. The left-most, dark vertical bars show results for the all-AM band condition comparing noise and sinusoidal carriers. Asterisks identify significant differences between conditions.

where the FM information ranged from high- to low-frequency bands. Similar labeling was used for the 8-band data shown in Fig. 6.

The first analysis examined whether a subset of AM+FM bands could reach levels of performance attained with AM+FM information on all bands. This question was addressed for AM+FM information on the lower frequency bands ("FM low") and also for AM+FM information on the higher frequency bands ("FM high"). Separate repeated measures ANOVAs were performed for the 4- and 8-band conditions and these were divided into separate ANOVAs for the

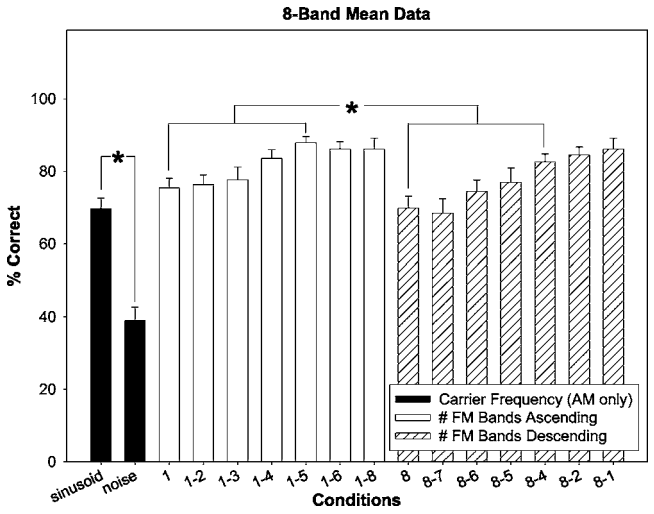

FIG. 6. Results from an 8-band hybrid AM and FM simulation. The *y* axis shows the conditions, with numbers representing the frequency band(s) containing FM information. Unfilled bars represent conditions with FM information added from low- to high-frequency bands. Hatched bars represent conditions with FM information added from high- to low-frequency bands. The left-most, dark vertical bars show results for the all-AM band condition comparing noise and sinusoidal carriers. Asterisks identify significant differences between conditions.

"FM low" and "FM high" conditions, resulting in four ANOVAs total. The results did indeed demonstrate that only a subset of AM+FM bands was needed to obtain similar performance to the all-band AM+FM condition. For the 4-band "FM high" analysis, Bonferroni adjusted planned comparisons showed that although there was a significant improvement from AM+FM on band 4 only to AM+FM on all bands $[F(1,6)=68.92, p<0.001]$, AM+FM on bands 4–2 was not significantly different from having AM+FM on all 4 bands ($p=0.29$). In other words, AM+FM information on the 3 upper bands produced similar levels of performance as the AM+FM all-band condition. Likewise, for the 4-band "FM low" analysis, planned comparisons resulted in significant differences between the all-band compared to the band-1-only AM+FM condition $[F(1,6)=11.88, p<0.02]$, but not with the band 1-2 condition ($p=0.36$). In this case, AM+FM information on only the lowest 2 bands was needed for performance to reach levels found for the all-band condition.

A similar analysis was performed for the 8-band group. For the 8-band "FM high" analysis, Bonferroni-adjusted planned comparisons showed that AM+FM information on at least 5 of the upper frequency bands produced similar performance as having all 8 AM+FM bands ($p>0.008$). For the 8-band "FM low" analysis, AM+FM information on only the lowest 3 AM+FM bands was sufficient to produce performance levels that were similar to the all-band condition ($p>0.008$).

The results discussed above indicate that fewer AM+FM bands were required to reach a plateau in performance when FM information was added to low- than to high-frequency bands. To examine this in more detail, direct comparisons were made between FM on high vs. low bands for conditions sharing the same number of bands. In the 4-band condition, performance with FM information only in the lowest band produced significantly higher scores (48.3% compared to 33.4%) compared to FM information only in the highest band (paired *t* test: $p<0.01$). In the 8-band condition, similar single condition comparisons (i.e., bands 1–2 vs. 8–7) failed to reach significance. However, a comparison of the five "FM high" vs. "FM low" conditions with the same number of bands (repeated measures ANOVA with FM region and band combination as factors) demonstrated higher scores (80.2%) with low-frequency FM bands compared to high-frequency FM bands (74.5%) $[F(1,14)=12.95, p<0.01]$.

In the final analysis, performance levels were compared for AM stimuli using sinusoidal or noise carriers. Significant differences were found between sinusoidal and noise carriers with 8 AM-only bands (paired *t* test: $p<0.001$), but not with 4 bands. In the 8-band condition, performance with sinusoidal carriers was significantly higher (by 30%) than with a noise carrier. The lower performance with the noise carrier can be attributed to the introduction of frequency modulation artifacts brought about by filtering the signal into narrow bands. The 4-band condition was more resistant to potential artifacts because of its broader bandwidth. This outcome is described in more detail in Sec. VII.

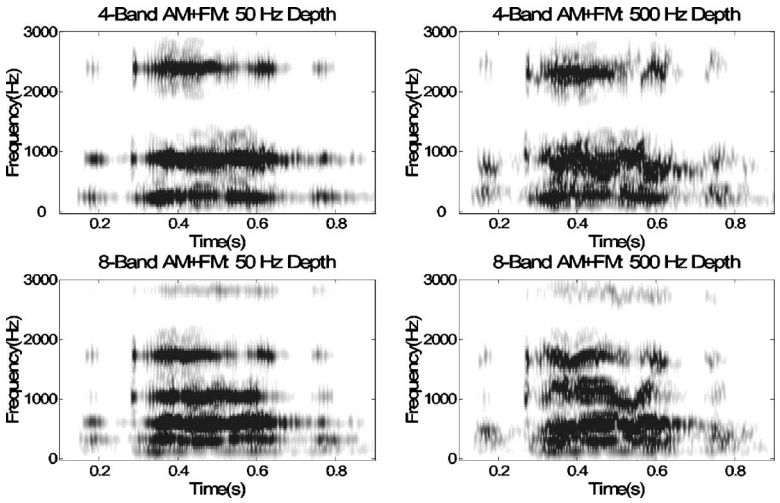

FIG. 7. Spectrogram of the phrase ("The girl at") for the 4-band (top panels) and 8-band conditions (bottom panels) with an FM depth (i.e., bandwidth) of 50 (left) and 500 Hz (right). The FM rate was 400 Hz. The spectrum is shown from 0 to 3 kHz, however the original bandwidth was 8.8 kHz.

VII. DISCUSSION

A. Roles of AM and FM in speech recognition

To compensate for the reduced spectral resolution in cochlear implants, an alternative and complementary code for transmitting frequency information was proposed, namely, frequency modulation. The AM+FM processing significantly improved performance relative to AM processing in all but the two most extreme band conditions: with 1 band, speech recognition performance with either type of processing was not intelligible, and with 32 bands, speech recognition performance was producing scores that were close or equal to 100%. Because cochlear implant users receive at most eight effective information bands (Friesen *et al.*, 2001), the additional information provided by AM+FM processing would be of great benefit when listening to speech in realistic listening environments. A consistent finding in each of the experiments was the improvement in speech recognition scores with more spectral bands. This was true regardless of the type of processing (e.g., AM vs. AM+FM, rate, depth), acknowledging that although alternative processing techniques may improve the performance of cochlear implant users when listening to speech in the presence of a competing talker, further benefit can be achieved by increasing the number of effective spectral bands. From experiment 1, the results showed that performance improved up to 16 bands with AM processing. Performance also improved by adding more bands with AM+FM processing, however performance with AM+FM processing reached maximum performance levels with fewer bands (8 bands) compared to AM processing.

B. Effects of FM depth and rate

A further examination of the effects of AM+FM processing on speech recognition performance revealed sensitivity to the FM depth (experiment 2) and FM rate (experiment 3). Specifically, when the depth was increased from 50 to 500 Hz, higher scores were observed in both the 4- and 8-band conditions. The spectral cues available with FM depths of 50 and 500 Hz are demonstrated in Fig. 7. In this figure, 4- and 8-band spectra of the phrase "The girl at" are shown. A comparison of the 50-Hz depth (left panels) and 500-Hz depth conditions (right panels) demonstrates that formant movement is most accurately represented with the greater FM depth (i.e., larger bandwidth). For the phrase shown in the figure, this is particularly noticeable between approximately 0.4 and 0.6 s. With increasing depth, there is a greater range of frequencies per band to capture the formant transition, potentially allowing the listener to better track the target sentence.

In experiment 3, a comparison of FM rates showed higher performance with higher rates. However, unlike FM depth, the rate that provided the most benefit varied depending on the number of bands. The 8-band condition resulted in similarly high levels of performance with an FM rate of 400 and 50 Hz. In contrast, with 4 bands, the 50-Hz rate produced significantly poorer performance than with 400 Hz. These results indicate that there is a tradeoff between FM rate and the number of bands. To clarify these results, Fig. 8 shows the f0 contour of natural and processed speech for the vowel /u/ at each of the two FM rates. As can be seen in the figure, the f0 contour is more adequately represented with the higher rate. The figure also demonstrates that if the number of bands is large enough to provide the spectral detail, then increasing the FM rate above 50 Hz will contribute little, if at all. On the other hand, more spectral smearing occurs as the number of bands is decreased and, consequently, higher FM rates can provide the listener with f0 information not readily available from the 4-band envelopes. In such cases, the listener could take advantage of the FM rate to follow the f0 contour of one or both talkers, and since the f0 of most talkers is at least 100 Hz, the higher rate would allow for better performance.

The FM rates and depths used in the present study can be perceived by users of cochlear implants. In a study by Chen and Zeng (2004), three adults with the Nucleus-22 cochlear implant and three normal-hearing listeners were presented with three types of frequency modulation: an upward sweep, a downward sweep, and a sinusoidal frequency modulation. They demonstrated that although the frequency difference limen increased with increases in the standard frequency in cochlear implant subjects, their difference limens were comparable to the normal-hearing listeners at low standard frequencies ($<$1000 Hz) and low sinusoid modulation

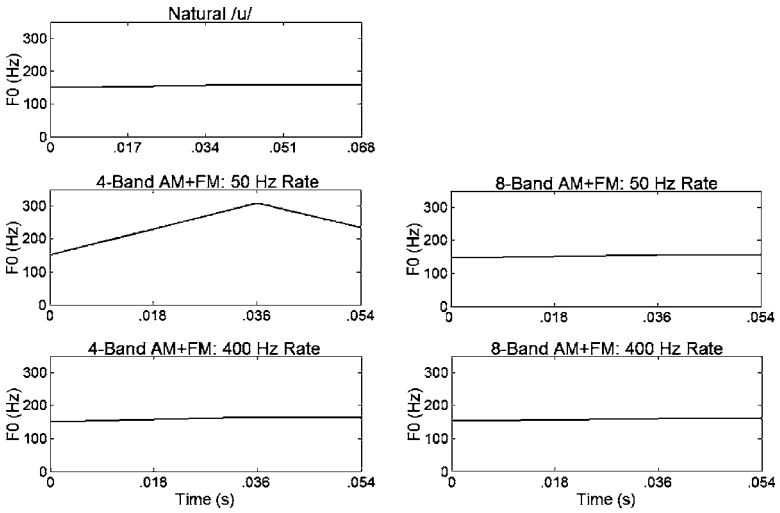

FIG. 8. The F0 contours of the vowel /u/ for the natural, unprocessed condition (top, left panel). Separate F0 contours are shown for the 4-band (left panels) and 8-band conditions (right panels) with FM rates of 50 (middle panels) and 400 Hz (bottom panels).

rates (<320 Hz). Beyond this, discrimination performance decreased monotonically. Their results suggest that cochlear implant users could have access to the FM information offered in the present study, i.e., 400 Hz FM rate and 500 Hz FM depth.

C. Effects of the number and distribution of FM bands

Experiment 4 demonstrated that only half of the bands required AM+FM information to reach similar levels of performance as the all-band AM+FM condition. This would correspond to a cutoff frequency of 1318 Hz for both the 4- and 8-band conditions. Providing FM information in this frequency range would allow for a better representation of key formants F1 and F2 known to be crucial for the identification of most speech sounds.

Experiment 4 also demonstrated that FM information provides greater benefit in low- than high-frequency bands. The low-frequency FM likely provided pitch information that could be used to segregate the two competing voices. This finding is not surprising since temporal fine structure, which is coded by FM, is critical for pitch perception (Smith *et al.*, 2002; Zeng *et al.*, 2004). In support of this, there are several recent studies showing that cochlear implant users benefit greatly from low-frequency, residual acoustic hearing when listening to speech in the presence of other speech sounds. This has been demonstrated both in cochlear implant listeners who combine an implant with a hearing aid on the non-implanted ear (Kong *et al.*, 2005) as well as cochlear implant users who have received a short electrode cochlear implant (Turner *et al.*, 2004). In sum, the results from experiment 4 are consistent with those from the bandwidth and rate experiments, and highlight the acoustic features coded by the additional FM cue and their potential role in improving speech perception with a competing talker.

D. Effects of the AM carrier

The comparison of carrier frequencies for the all-AM conditions revealed higher performance for sinusoidal than noise carriers, but only when the number of bands was increased from 4 to 8. The better performance with sinusoidal than noise carriers was likely due to additional envelope fluctuations present in the narrow-band noise carriers. To demonstrate this point, Fig. 9 compares the highest- and lowest-band waveforms of the sentence "The girl at the booth sold fifty bonds" possessing either a sinusoidal or noise carrier, or left unprocessed. As can be seen in the "Lowest Band" panels of the figure, the unprocessed waveform (top panel) is more accurately replicated with the sinusoidal carrier (bottom panels of each band condition) than with the noise carrier (middle panels of each band condi-

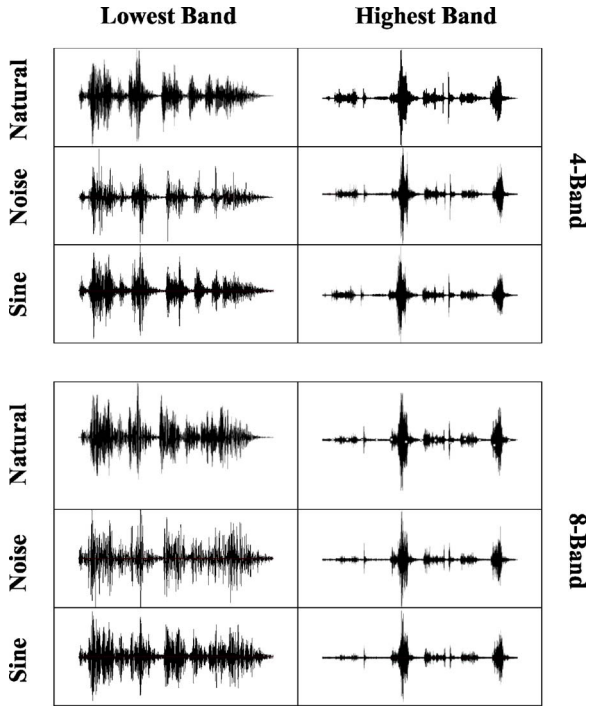

FIG. 9. Waveforms of a single-frequency band from the sentence "The girl at the booth sold fifty bonds" for the natural, unprocessed condition or the AM-only processed speech having either a sinusoidal or noise carrier. The upper three rows show waveforms from a single band in the 4-band condition (natural condition shown in the top panels and AM-processed in the middle and lower panels). Likewise, the lower three rows show waveforms for a single band in the 8-band condition (natural condition shown in the top panels and AM-processed in the middle and lower panels). The lowest band waveforms (band 1) are shown in the left column and the highest band waveforms (band 4 or 8) are shown in the right column.

tion). The noise carrier introduces additional spikes to the waveform (amplitude modulations). Thus, the reason that sinusoidal carriers outperformed noise carriers with 8 bands, but not with 4, can be explained by the greater amplitude modulation associated with narrower bandwidths when the number of bands was increased. Sinusoidal and noise carriers will therefore produce different levels of performance for stimuli processed into the midrange (e.g., approximately 8–16 bands), but not at the extremes. For this reason, previous studies using a noise carrier for an 8–16-band simulation might have underestimated performance due to modulations introduced during stimulus processing.

VIII. CONCLUSIONS

(i) These results underscore the importance of FM in speech recognition under realistic listening situations, particularly when the competing sound is speech. However, FM may have its greatest role when speech is severely impoverished, as it is with cochlear implants.

(ii) Formant transitions and voice pitch can be useful for segregating competing speech sounds. However, these cues are not adequately coded in current cochlear implant speech processing algorithms. The addition of FM could potentially provide these cues.

(iii) Low-frequency FM information contributes more to speech perception with a competing talker than high-frequency FM. This finding suggests that listeners may rely more on low-frequency temporal fine structure cues to segregate the target from the masking voice.

(iv) The slowly varying FM cue can be readily extracted from the temporal fine structure and may enhance cochlear implant performance.

ACKNOWLEDGMENTS

The authors thank Jivesh Sabnani and Neil Biswas for their help in data collection. The IEEE sentences were created by Dr. Lou Braida and recorded by Dr. Monica Hawley and Dr. Ruth Litovsky. This work was supported by grants from the National Institutes of Health (F32 DC05900 to GSS and 2R01 DC02267 to FGZ).

[1]Results with a 5 dB TMR have been presented in Zeng *et al.* (2005).

Assmann, P. F. (**1995**). "The role of formant transitions in the perception of concurrent vowels," J. Acoust. Soc. Am. **97**, 575–584.

Bregman, A.. (1990). "Auditory scene analysis." Cambridge, MA, MIT Press.

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., and Hukin, R. W. (**2000**). "Effectiveness of spatial cues, prosody, and talker characteristics in selective attention," J. Acoust. Soc. Am. **107**, 970–977.

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

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

Friesen, L., Shannon, R., 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.

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.

Kawahara, K., 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.

Kong, Y-Y., Stickney, G., and Zeng, F-G. (**2005**). "Contribution of acoustic low-frequency information in speech and melody recognition in cochlear implants," J. Acoust. Soc. Am. **117**, 1351–1361.

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**(5), 752–760.

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

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**(1), 64–73.

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.

Rothauser, E. H., Chapman, W. D., Guttman, N., Nordby, K. S., Silbiger, H. R., Urbanek, G. E., and Weinstock, M. (**1969**). "I.E.E.E. recommended practice for speech quality measurements," IEEE Trans. Audio Electroacoust. **17**, 227–246.

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

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

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

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," J. Acoust. Soc. Am. **115**, 1729–1735.

Zeng, F-G., Nie, K-B., Liu, S., Stickney, G., Del Rio, E., Kong, Y-Y., and Chen, H. (**2004**). "On the dichotomy in auditory perception between temporal envelope and fine structure cues," J. Acoust. Soc. Am. **116**, 1351–1354.

Zeng, F-G., Nie, K-B., Stickney, G., Kong, Y-Y., Vongphoe, M., Wei., C, and Cao, K. (**2005**). "Speech recognition with slowly-varying amplitude and frequency modulation cues," Proc. Natl. Acad. Sci. U.S.A. **102**(7), 2293–2298.