Encoding Frequency Modulation to Improve Cochlear Implant Performance in Noise

Kaibao Nie, *Member, IEEE*, Ginger Stickney, and Fan-Gang Zeng*, *Member, IEEE*

***Abstract*—Different from traditional Fourier analysis, a signal can be decomposed into amplitude and frequency modulation components. The speech processing strategy in most modern cochlear implants only extracts and encodes amplitude modulation in a limited number of frequency bands. While amplitude modulation encoding has allowed cochlear implant users to achieve good speech recognition in quiet, their performance in noise is severely compromised. Here, we propose a novel speech processing strategy that encodes both amplitude and frequency modulations in order to improve cochlear implant performance in noise. By removing the center frequency from the subband signals and additionally limiting the frequency modulation's range and rate, the present strategy transforms the fast-varying temporal fine structure into a slowly varying frequency modulation signal. As a first step, we evaluated the potential contribution of additional frequency modulation to speech recognition in noise via acoustic simulations of the cochlear implant. We found that while amplitude modulation from a limited number of spectral bands is sufficient to support speech recognition in quiet, frequency modulation is needed to support speech recognition in noise. In particular, improvement by as much as 71 percentage points was observed for sentence recognition in the presence of a competing voice. The present result strongly suggests that frequency modulation be extracted and encoded to improve cochlear implant performance in realistic listening situations. We have proposed several implementation methods to stimulate further investigation.**

***Index Terms*—Amplitude modulation, cochlear implant, fine structure, frequency modulation, signal processing, speech recognition, temporal envelope.**

I. Introduction

COCHLEAR implants electrically stimulate the auditory nerve to restore hearing to profoundly deaf persons. The modern multichannel devices produce word recognition scores around 80% for sentences in quiet, allowing the majority of their users to talk on the phone. However, these cochlear-implant users have much greater difficulty than normal-hearing listeners in recognizing speech under realistic listening situations such as at a cocktail party or a restaurant, where background noise is always present [1]–[4]. Here, we first review acoustic cues and their perceptual roles in speech recognition, then we examine limitations of speech encoding strategies in current cochlear implants and, finally, we propose an innovative coding strategy that may be used to improve cochlear implant performance in noise.

Traditionally, acoustic cues in speech sounds have been analyzed from the speech production point of view [5], [6]. For example, waveform periodicity indicates the vocal cord vibration status and determines whether a sound is voiced or unvoiced (e.g., vowel /a/ versus consonant /s/); temporal cues such as sound duration and silent gaps typically reflect the manner of articulation (e.g., stop /b/ versus fricative /f/); spectral cues such as formants and their transitions reflect the place of articulation (e.g., labial /b/ versus glottal /g/).

Alternatively, acoustic cues in speech sounds can be examined from the speech perception point of view. For example, the auditory system is sensitive to amplitude and frequency modulations and may have developed specific mechanisms to extract them to form different neural representations [7]–[9]. As early as the 1930s, Dudley [10] invented vocoders and demonstrated that intelligible speech could be produced by amplitude modulations or temporal envelopes from only ten frequency bands. Shannon *et al.* [11] later found that amplitude modulations from as few as 3–4 bands are sufficient to support speech recognition in quiet. However, recent studies have indicated that the amplitude modulation cue cannot support robust speech recognition in noise, particularly when the noise is another competing voice [1]–[4]. Instead, frequency modulation derived from the temporal fine structure is needed to support speech recognition in noise and other critical functions such as speaker identification, music perception, tonal language perception and sound localization [12]–[14].

All current cochlear implants, except for those delivering the analog waveforms [15], have adopted speech processing strategies that focus on extracting and representing the amplitude modulation cue. For example, Continuous Interleaved Sampling (CIS) is such a strategy that has been implemented in all three major cochlear implants (Clarion, Nucleus, and Med-El) [16]. In the CIS strategy, a sound is filtered into a number of subbands with the number ranging from 4 to 22. The subband signal is then typically full-wave rectified and low-pass filtered to extract the temporal envelope. The temporal envelope is used to amplitude modulate a biphasic pulse train delivered to a stimulating electrode. Another example is the n-of-m, or peak picking strategy, which only uses the temporal envelopes from several bands that have the highest energy to stimulate a subset of electrodes [17]. A common feature in these strategies is that the pulse carrier has a constant rate, containing no information regarding the speech sound. The only exception was an obsolete strategy implemented in the early Nucleus device, which

Manuscript received September 25, 2003; revised March 25, 2004. This work was supported in part by the National Institutes of Health (NIH) under Grant 2RO1-DC02267 and Grant F32-DC05900. The work of K. Nie was supported in part by the Natural Science Foundation of China (NSFC) under Grant 30000041. *Asterisk indicates corresponding author.*

K. Nie and G. Stickney are with the Departments of Otolaryngology-Head and Neck Surgery and Biomedical Engineering, University of California, Irvine, CA 92697 USA (e-mail: knie@uci.edu).

*F. G. Zeng is with the Departments of Otolaryngology-Head and Neck Surgery and Biomedical Engineering, University of California, Irvine, CA 92697 USA (e-mail: fzeng@uci.edu).

Digital Object Identifier 10.1109/TBME.2004.839799

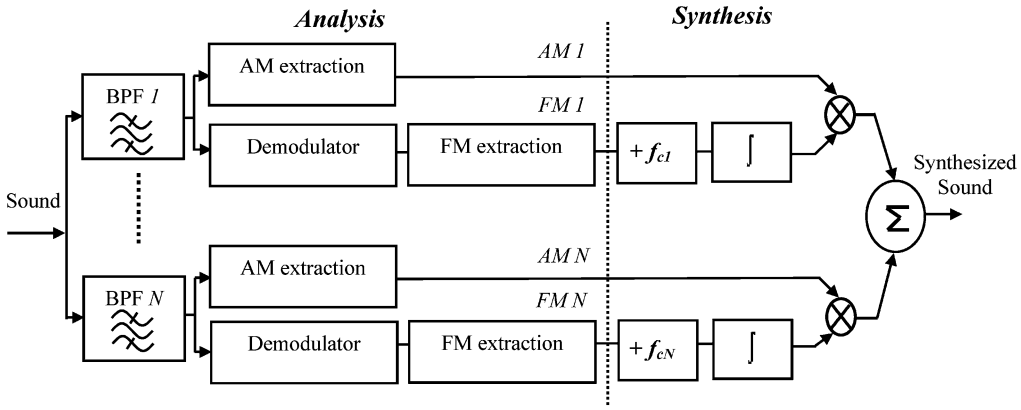

Fig. 1. Functional block diagram of the proposed speech processing strategy. The diagram contains an analysis part and a synthesis part. A sound is divided into N subbands by a bank of bandpass filters ($\mathrm{BPF}1 \sim \mathrm{BPF}N$). Within each subband, amplitude and frequency modulations are extracted in separate pathways (details are shown in Fig. 2). In acoustic simulations, the frequency modulation signal is added to the original center frequency, then integrated to recover the original phase information and, finally, multiplied by the amplitude modulation signal to recover the original subband signal. The processed sound is synthesized by summing these processed subband signals.

extracted the fundamental frequency in speech sounds and accordingly adjusted the stimulation rate delivered to the electrode [18].

A large body of experimental evidence has indicated that the normal auditory nerve does not produce a spike train at a fixed rate, but instead it produces spike trains that are capable of synchronizing with stimulus waveform periodicity up to 5000 Hz [19]. The neural spike train can follow phase changes of complex stimuli [20] and extract frequency modulation-related information [21]–[23]. Previous studies have also shown that temporal pitch in electric stimulation can follow rate changes up to 500–1000 Hz [35], [36]. Motivated by this body of physiological and psychophysical evidence, we propose to encode frequency modulation in cochlear implants to improve their performance in noise. In the following sections, we will first present an algorithm (Section II) that decomposes a signal into slowly varying amplitude and frequency modulations. We term this algorithm the frequency amplitude modulation encoding (FAME) strategy. We will verify in Section III the algorithm's accuracy and efficiency from a signal processing point of view. We will then describe the methodology for the acoustic simulation experiments in Section IV and evaluate their results in Section V.

II. Frequency-Amplitude-Modulation-Encoding Algorithm

A. General Structure

A common method to derive amplitude and frequency modulations is to compute the temporal envelope and the instantaneous frequency via the Hilbert transform. The problem with this method is that the estimated instantaneous frequency usually varies rapidly and over a broad range, producing values that often have no clear physical meaning [24]. Following Flanagan's methodology in his classic study on speech coding [25], particularly with the phase vocoders [26], we sought to extract slowly varying, band-limited amplitude and frequency modulations in speech sounds. We hypothesize that these modulations can be encoded in cochlear implants to improve their performance.

A signal, $s(t)$, can be approximated by a sum of N band-limited components, $x(t)$, containing both amplitude and frequency modulations

$$s(t) \approx \sum_{k=1}^{N} x_k(t) = \sum_{k=1}^{N} A_k(t) \cos\left[2\pi f_{ck} t + 2\pi \int_0^t g_k(\tau)\, d\tau + \theta_k\right] \quad (1)$$

where $A_k(t)$ and $g_k(t)$ are the kth band's amplitude and frequency modulations, whereas f_{ck} and θ_k are the kth band's center frequency and initial phase, respectively. The goal here is to remove the center frequency f_{ck} and to apply low-pass filters to limit and smooth the amplitude and frequency modulations. Fig. 1 displays the functional block diagram of this analysis-by-synthesis algorithm. In the analysis part, the original sound signal is divided into N subbands using a filter bank with center frequencies equally distributed on a logarithmic scale to mimic cochlear filters. Two independent parallel pathways are then used to extract amplitude modulation (AM) and frequency modulation (FM) in each band. The AM pathway extracts the slowly varying envelope, while the FM pathway extracts the slowly varying frequency modulation using a demodulator to remove the subband's center frequency (details are given below). In actual electric stimulation, the slowly varying envelope would amplitude modulate a slowly varying pulse rate to be delivered to the electrode. In acoustic simulations of the cochlear implant, the center frequency (f_c) has to be re-introduced and the instantaneous frequency has to be integrated for the recovery of the original subband phase value. In the last stage of the synthesis part, the recovered bandpassed signals are summed to form the synthesized speech containing the processed slowly varying amplitude and frequency modulations.

B. AM Extraction

Fig. 2(a) illustrates extraction of AM in the kth subband. The AM is extracted by full-wave rectification of the output of the bandpass filter, followed by a low-pass filter, LPF 1. The cutoff frequency of LPF 1 controls the maximal AM rate preserved in

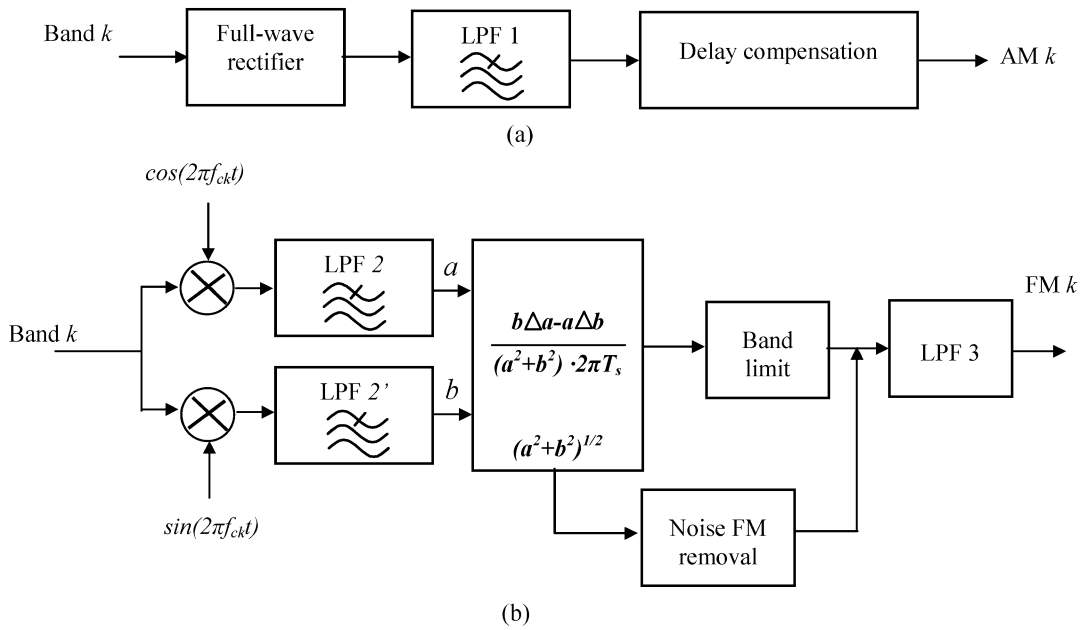

Fig. 2. Algorithms for extracting AM and FM. (a) Shows that the output of subband k is full-wave rectified and then low-passed (LPF 1) to obtain a slowly varying amplitude modulation signal AM k. Delay compensation is also introduced to synchronize the amplitude and frequency modulation pathways. (b) Shows that a pair of orthogonal sinusoidal signals at the center frequency of the k th subband is used to remove the center frequency from the original signal and to extract frequency modulation around the center frequency. Low-pass filters LPF 2 and LPF 2′ are used to limit the frequency modulation range (e.g., 500 Hz). Instantaneous frequency is then calculated from the in-phase signal a and the out-of-phase signal b using the formula shown in the middle block. The instantaneous frequency is further band-limited and low-passed (LPF 3) to limit the frequency modulation rate (e.g., 400 Hz). Additionally, an amplitude threshold device is used to remove artificial frequency modulations exaggerated by a differential process in the algorithm.

the AM signal. Additionally, the delay compensation box synchronizes signals between the AM and FM pathways.

C. FM Extraction

Fig. 2(b) details the FM extraction pathway. First, the output of the k th subband, $x_k(t)$, is subjected to a quadrature oscillator with the center frequency f_{ck}. This manipulation is equivalent to shifting the spectrum of $x(k)$ from f_{ck} to *zero* and $2f_{ck}$ in the frequency domain. The following low-pass filters (*LPF 2* and *LPF 2'*) then extract the slowly varying frequency components (a and b) by removing the high frequency component ($2f_{ck}$). In signal processing nomenclature, the slowing-varying components a and b are termed in-phase and out-of-phase signals of the original subband signal $x_k(t)$, respectively.

Mathematically, if $x_k(t)$ can be described as $x_k(t) = m(t)\cos[2\pi f_{ck}t + \varphi(t)]$, where $m(t)$ is the amplitude, f_{ck} is the center frequency and $\psi(t)$ is the phase, then the in-phase signal can be derived

$$\begin{aligned} &x_k(t) \times \cos(2\pi f_{ck}t) \\ &\quad = m(t)\cos[2\pi f_{ck}t + \varphi(t)]\cos(2\pi f_{ck}t) \\ &\quad = \frac{1}{2}m(t)\cos[2\pi f_{ck}t + 2\pi f_{ck}t + \varphi(t)] \\ &\qquad + \frac{1}{2}m(t)\cos[2\pi f_{ck}t + \varphi(t) - 2\pi f_{ck}t] \\ &\quad = \frac{1}{2}m(t)\cos[2(2\pi f_{ck})t + \varphi(t)] + \frac{1}{2}m(t)\cos\varphi(t). \end{aligned} \quad (2)$$

The first term in the above equation can be filtered out by the low-pass filter LPF2 to produce

$$a = \frac{1}{2}m(t)\cos\varphi(t). \quad (3)$$

Similarly, the out-of-phase signal can be derived as

$$\begin{aligned} &x_k(t) \times \sin(2\pi f_{ck}t) \\ &\quad = m(t)\cos[2\pi f_{ck}t + \varphi(t)]\sin(2\pi f_{ck}t) \\ &\quad = \frac{1}{2}m(t)\sin[2\pi f_{ck}t + \varphi(t) + 2\pi f_{ck}t] \\ &\qquad - \frac{1}{2}m(t)\sin[2\pi f_{ck}t + \varphi(t) - 2\pi f_{ck}t] \\ &\quad = \frac{1}{2}m(t)\sin[2\pi(2f_{ck})t + \varphi(t)] - \frac{1}{2}m(t)\sin\varphi(t). \end{aligned} \quad (4)$$

Again, the first term in the above equation can be filtered out

$$b = -\frac{1}{2}m(t)\sin\varphi(t) = \frac{1}{2}m(t)\cos[\varphi(t) + \pi/2]. \quad (5)$$

Dividing b (5) by a (3) will produce,

$$\begin{aligned} \frac{b}{a} &= -\tan\varphi(t) \\ \varphi(t) &= \tan^{-1}\left(-\frac{b}{a}\right). \end{aligned} \quad (6)$$

Finally, the instantaneous frequency can be obtained

$$\mathrm{FM} = \frac{1}{2\pi}\frac{d\varphi(t)}{dt} = \frac{d\tan^{-1}\left(-\frac{b}{a}\right)}{2\pi dt} = \frac{b(da/dt) - a(db/dt)}{2\pi(a^2+b^2)}. \quad (7)$$

In discrete implementation, differentiation in (7) can be substituted by calculating the difference in time (Δ) to obtain the slowly varying frequency modulation

$$\mathrm{FM} = \frac{b\Delta a - a\Delta b}{2\pi(a^2+b^2) \times T_s} \quad (8)$$

where Ts represents sampling period.

In Fig. 2(b), two filters with the same bandwidth as the initial low-pass filters (*LPF 2* and *LPF 2'*) are then used to remove

high-frequency distortions generated by the instantaneous frequency calculation. An additional amplitude calculation $[(a^2 + b^2)^{1/2}]$ is used as a threshold device to remove erroneous frequency modulation produced by low-level noise due to the differential process in FM extraction. In this case, the FM bandwidth was set to 0 Hz. Finally, the band-limited frequency modulation is subject to a low-pass filter *(LPF 3)* to produce the desired slowly varying frequency modulation.

In signal processing terms, the cutoff frequency of *LPF 2* and *LPF 2'* determines FM depth or FM bandwidth, whereas the cutoff frequency of *LPF 3* determines FM rate. In this study, the FM depth is set to 500 Hz or the filter bandwidth, depending on the number of frequency bands and the center frequency of these bands. For example, the bandwidth in the first subband of an 8-band processor is only 139 Hz, a 500-Hz FM bandwidth would produce undesirable cross-talk between adjacent bands. The FM rate is set to 400 Hz or below so that it may be perceived by cochlear implant listeners [27], [28].

III. Algorithm Verification

Two synthetic sounds, a frequency sweep and a speech syllable */bai/*, were used to verify the processing algorithm. Fig. 3(a) shows a frequency sweep with its instantaneous frequency being linearly increased from 80 to 8800 Hz (the bandwidth used in the present implementation) over a 4-s duration. Fig. 3(b) shows AM components and Fig. 3(c) shows FM components extracted from an eight-band FAME implementation. As expected, the AM pathway output contained a short-duration amplitude envelope at low frequencies [bottom traces in Fig. 3(b)] and a long-duration envelope at high frequencies (top traces) because the frequency sweep was linear whereas the analysis filters' bandwidth was logarithmic. Similarly, the FM pathway output contained a smooth linear frequency sweep with a bandwidth corresponding to the subband bandwidth at low frequencies [bottom traces in Fig. 3(c)] and 500 Hz at high frequencies (top traces). These results verified that the present FAME strategy was able to extract and synchronize the slowly varying AM and FM cues with the same FM depth and rate as specified in the design and implementation.

Additionally, the synthetic speech syllable */bai/* was used to show the functional significance of the proposed FAME strategy. The selected syllable contains rich frequency modulations in the formant transition from consonant /b/ to vowel /a/, as well as from /a/ to /i/ [see arrows in the syllable's spectrogram shown in Fig. 4(a)]. Fig. 4(b) shows the output of the 8-band AM pathway and Fig. 4(c) shows the output of the FM pathway. At the FM output, the thick lines represent extracted FM with a 50-Hz FM rate while the thin lines represent extracted FM with a 400-Hz FM rate. As expected, the AM components reflect the dynamic energy distribution across different bands, i.e., the temporal envelope cues extracted and encoded in the majority of current cochlear implants. On the other hand, the FM components reflect formant transitions, particularly the slowly varying frequency changes at the output of bands 4 and 5 (from bottom) between time intervals of 0.3 and 0.5 s.

Fig. 5 shows the spectrograms of the original syllable (a), the AM-only processed signal with 8 bands (b), and the FM-added

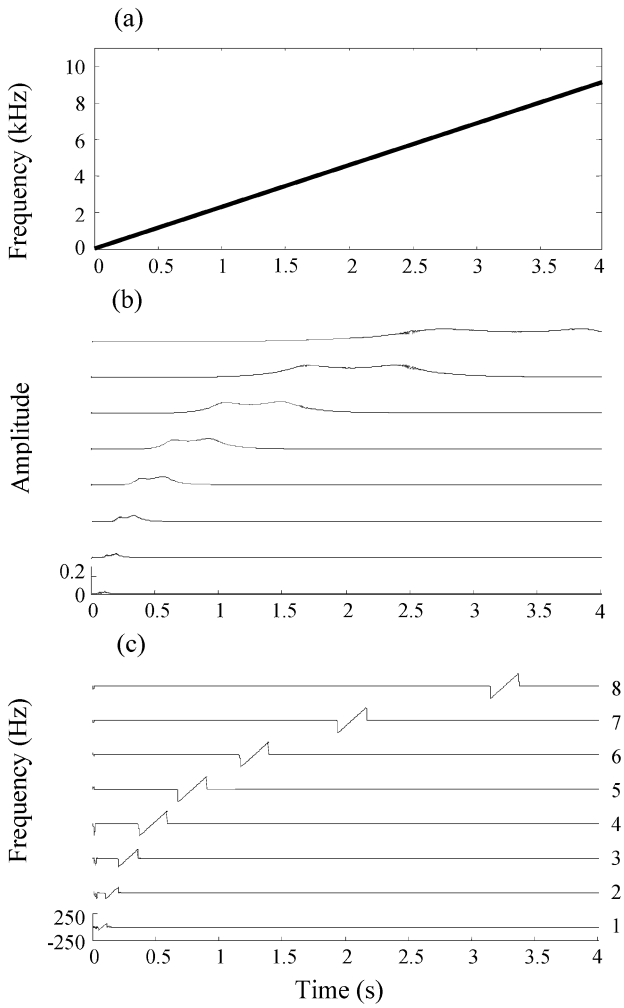

Fig. 3. (a) Shows the spectrogram of a frequency sweep from 80 to 8 800 Hz over a 4-s time interval. (b) and (c) Show extracted amplitude and frequency modulation signals from an 8-band processor, respectively. The center frequency of band 1 is the lowest whereas that of band 8 is the highest.

signal also with 8 bands (c). There are several noticeable differences between the AM only strategy and the FAME strategy. First, note the apparently better representation of the formant transition in the FAME strategy than in the AM-only strategy. This better representation is evident for the upward transition for formant 2 between /b/ and /a/ for the duration from 0 to 0.03 s, and additionally the downward transition for formant 1 and the upward transition for formant 2 for the diphthong /ai/ from 0.3 to 0.5 s. Second, note an artifact of downward frequency modulation in formant 3 for durations from 0.5 to 0.6 s [also evident in trace 6 from bottom in Fig. 4(c)], which was not present in the original spectrogram. This artifact was due to the specific setting of band six's center frequency, which was between formants 2 and 3 and contained no physical energy. Third, a less apparent but potentially important difference was noted in the representation of the fundamental frequency in the FAME spectrogram (subtle downward slanted lines) but not in the AM-only spectrogram (horizontally straight lines).

This detailed comparison clearly shows that the proposed FAME strategy extracts and encodes both formant transition and fundamental frequency cues. In contrast, the AM-only strategy encodes these cues either indirectly in the time domain

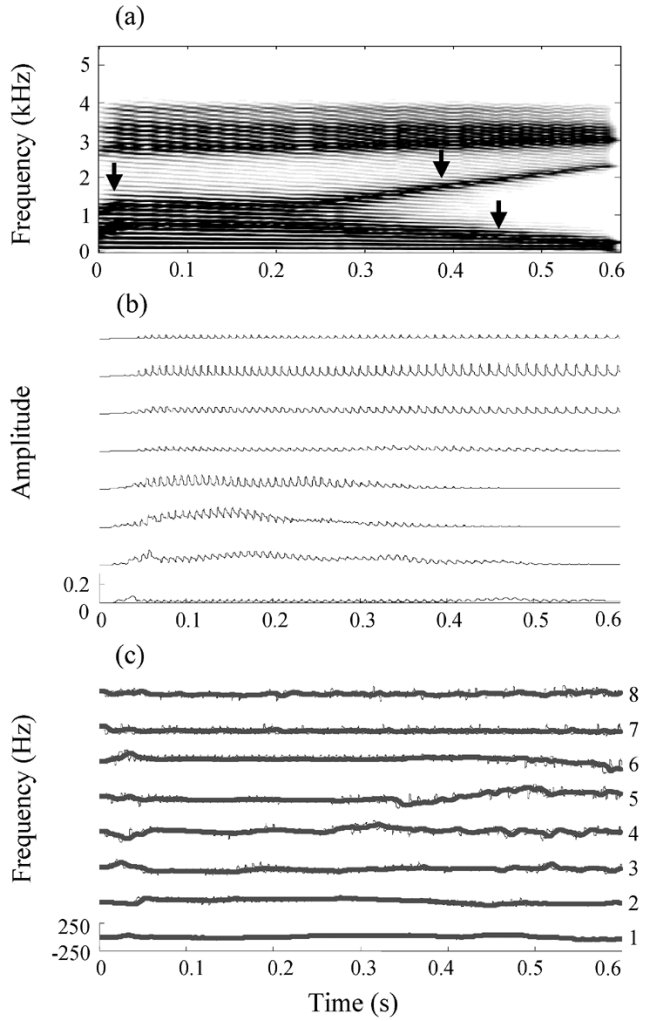

Fig. 4. (a) Shows the spectrogram of a synthetic syllable */bai/*, while (b) and (c) show amplitude and frequency modulation signals from an 8-band processor. Arrows indicate formant transitions.

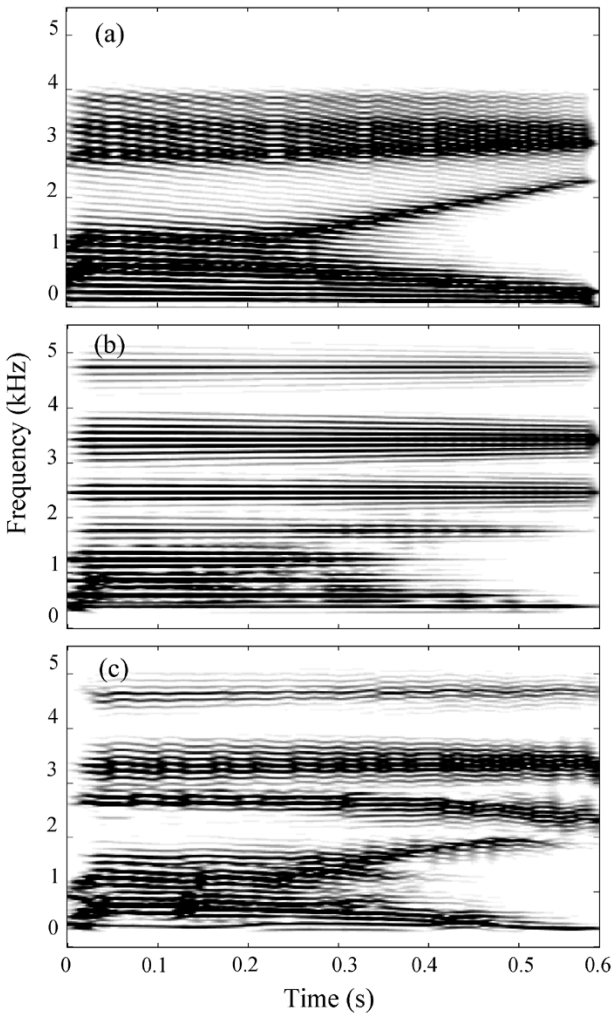

Fig. 5. (a) The spectrum for the original, (b) AM-only, and (c) the AM+FM processed synthetic syllable */bai/*.

(e.g., fundamental frequency) or crudely in the frequency domain (e.g., formant transition), particularly when the number of bands is low. It is also clear from this detailed comparison that the present FAME strategy contains more acoustic information than the previously proposed strategies that explicitly encoded the fundamental frequency cue [18], [29].

IV. Experimental Design

Two psychoacoustic experiments were conducted with normal-hearing listeners to evaluate the contribution of the additional FM information to speech recognition in noise. Experiment 1 examined phoneme recognition in quiet or in the presence of a steady-state speech-shaped noise. Experiment 2 mimicked a more realistic listening situation by measuring sentence recognition in the presence of a competing voice.

A. Subjects

A total of 40 normal-hearing subjects were recruited to participate in the experiments. Five subjects participated in the phoneme recognition experiment and 35 participated in the sentence recognition experiment. The 35 subjects in the sentence experiment were divided into seven groups of five subjects each. Each group participated in one of the seven experimental conditions, including three processing conditions (AM versus AM+FM, or Natural speech) and three band conditions (2, 4, and 8 bands). Informed consent and local IRB approval were obtained for this study.

B. Test Materials

The phoneme materials included 12 /hvd/ vowels spoken by three male adults, three female adults, and three girls [30], as well as 20 /aCa/ consonants spoken by two male and two female adults [31]. The phonemes were presented either in quiet or in a speech-spectrum-shaped noise at 0 and -5-dB signal-to-noise ratios (SNRs). The target sentence materials consisted of 50 IEEE sentences spoken by a male talker [32]. Each sentence consisted of five keywords. The background noise, the masker, for the sentence test was a competing IEEE sentence (i.e., "Port is a strong wine with a smoky taste") spoken by a different male talker. Both the target and the masker had the same onset, with the masker being always longer in duration. The SNRs were set at 0, 5, 10, 15, and 20 dB in the sentence experiment with ten randomized sentences for each block.

C. Signal Processing

All stimuli were pre-emphasized with a first-order, high-pass Butterworth filter above 1.2 kHz. The analysis filters were fourth-order elliptic bandpass filters with a 50-dB attenuation in the stop band and a 2-dB ripple in the passband. The overall processing bandwidth was 300–5500 Hz in the phoneme experiment and 80–8 800 Hz in the sentence experiment. The low-pass filters (*LPF 1, LPF 2, LPF 2' and LPF 3* in Fig. 2) that were used to extract AM and FM cues were fourth-order Bessel filters, producing constant group delays in the passband. In the phoneme test, the cutoff frequency was set at 5, 50, or 500 Hz for the AM-rate filter (i.e., *LPF 1*). The cutoff frequency was set at 400 Hz for the FM-rate filter (i.e., *LPF 3*). The FM bandwidth (*LPF 2 and LPF 2'*) was set at 50, 200, or 500 Hz or the subband bandwidth if it was narrower than the LPF 2's cutoff frequency. In the sentence test, the cutoff frequency was 500 Hz for the AM-rate filter and 400 Hz for the FM-rate filter. The FM bandwidth was set at 500 Hz or the subband bandwidth.

D. Procedure

All subjects performed the experiments in a double-wall, sound-attenuated booth. In the phoneme recognition experiment, a graphic user interface was created with 12 vowels or 20 consonants displayed as buttons on a computer screen. After a phoneme was presented, the subject was instructed to choose the correct answer by clicking the button corresponding to the presented phoneme. Feedback regarding the correct answer was provided after each trial. The stimulus was played in a random order. When one test condition was finished, the percent correct score was calculated and stored to a file for further data analysis. The stimulus was presented at 70 dBA and monaurally via a Sennheiser headphone. A pretest with the unprocessed phonemes was given to each subject to screen out those who scored below 90% correct. The order of all experimental conditions was randomized for each subject.

In the sentence recognition experiment, the subject was presented with the target sentence in the presence of the competing sentence. The subject was asked to type in as many words as possible from the target sentence via a computer keyboard. The number of correctly identified keywords was calculated to produce the final percent correct score for each of the five SNR blocks. Guessing was encouraged but no feedback was given during or after the experiment. All blocks were randomized as well as all ten sentences within each block. A practice session with unprocessed speech was offered prior to testing to screen out subjects who scored below 85%. Following this practice session, a second practice session was given to familiarize the subjects with the processed speech. No score was calculated for this second practice session. The sentences were presented at 65 dBA and monaurally via a Sennheiser headphone.

V. Experimental Results

A. Phoneme Recognition

Fig. 6 shows vowel recognition scores as a function of SNR for the 2-band [Fig. 6(a)], 4-band [Fig. 6(b)], 8-band [Fig. 6(c)], and 16-band [Fig. 6(d)] conditions. In general, the FAME strategy produced better performance than the AM-only strategy. ANOVA revealed that this better performance was significant for all $(\mathrm{p} < 0.01)$ except the 2-band condition $[\mathrm{F}(1,4) = 5.6, \mathrm{p} = 0.8]$. In addition, more frequency bands $[\mathrm{F}(3,12) = 122.1, \mathrm{p} < 0.05]$ and higher SNRs $[\mathrm{F}(2,8) = 80.3, \mathrm{p} < 0.05]$ produced better performance than fewer bands and lower SNRs.

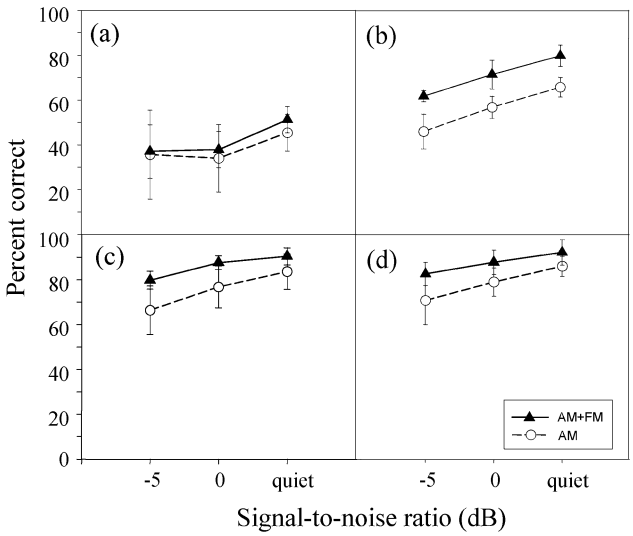

Fig. 6. Vowel recognition as a function of the number of bands (panels) and SNR (x-axis). (a)-(d) Show data from 2, 4, 8, and 16 bands, respectively. The AM-only condition is represented by open circles and the AM+FM condition by filled triangles.

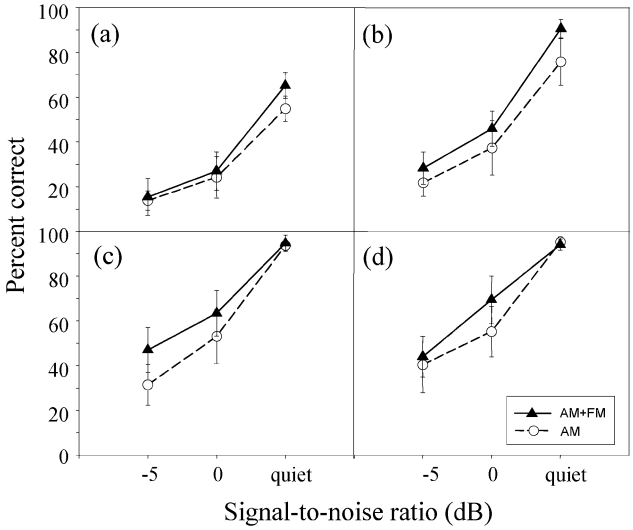

Fig. 7. Consonant recognition as a function of the number of bands (panels) and SNR (x-axis).(a)-(d) Show data from 2, 4, 8, and 16 bands, respectively. The AM-only condition is represented by open circles and the AM+FM condition by filled triangles.

Fig. 7 shows consonant recognition scores as a function of SNR for the 2-band [Fig. 7(a)], 4-band [Fig. 7(b)], 8-band [Fig. 7(c)], and 16-band [Fig. 7(d)] conditions. Similar to the vowel recognition result, the FAME strategy produced better performance than the AM strategy $[\mathrm{F}(1,4) = 56.4, \mathrm{p} < 0.01]$. Both the number of bands $[\mathrm{F}(3,12) = 181.2, \mathrm{p} < 0.01]$ and the SNR $[\mathrm{F}(2,8) = 143.5, \mathrm{p} < 0.01]$ were also significant factors; the noise condition appeared to produce greater improvement than the quiet condition for the 8- and 16-band processors.

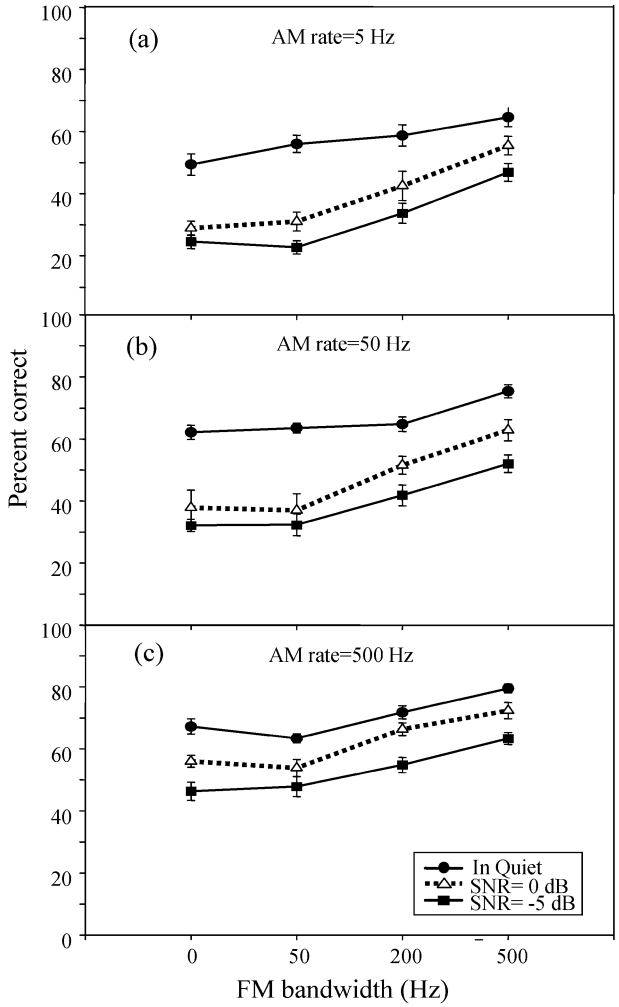

Fig. 8. Vowel recognition as a function of AM rate (panels) and FM range (x-axis) in a 4-band processor. (a)-(c) Show data with the AM rate at 5, 50, and 500 Hz, respectively.

Fig. 8 shows vowel recognition scores from a four-band processor as a function of the FM bandwidth (0, 50, 200, and 500 Hz) and the AM rate [5, 50, and 500 Hz, corresponding to Fig. 8 (a), (b), and (c), respectively]. The 0-Hz FM bandwidth was equivalent to the AM-only condition. The FM rate was fixed at 400 Hz. First, note the significant improvement in performance with the FM bandwidth $[\mathrm{F}(3,12) = 103.2, \mathrm{p} < 0.01]$. It appears that a 200-Hz FM bandwidth was needed to improve performance significantly. Second, note the monotonic improvement with the AM rate $[\mathrm{F}(2,8) = 115.6, \mathrm{p} < 0.01]$ and the SNR $[\mathrm{F}(2,8) = 50.2, \mathrm{p} < 0.01]$. Finally, note the relatively shallower slope for the quiet condition (filled circles) than for the two noise conditions, suggesting that the additional FM cue improved performance more in noise than in quiet.

B. Sentence Recognition

Fig. 9, shows sentence recognition scores as a function of SNR from 0 to 20 dB for the 2-band [Fig. 9(a)], 4-band [Fig. 9(b)], and 8-band [Fig. 9(c)] conditions. The inverted triangles represent the natural, unprocessed condition, whereas the triangles and circles represent the AM+FM and the AM conditions, respectively. The same results from the natural condition are presented in all panels for comparison.

First, note that sentence recognition with the natural stimuli was relatively resistant to the noise over the SNR range tested, with almost perfect performance at 20-dB SNR and gradually dropping to 75% at 0-dB SNR. Second, note that the performance with the AM strategy was 0% with the 2-band condition and 50% or less with the 4- and 8-band conditions. The drop in performance for the AM+FM condition at the 15-dB SNR was due to the greater difficulty of the sentences used for this condition than at other SNRs. The present low performance with the AM strategy is different from Shannon *et al.*'s original cochlear implant simulation study in which nearly perfect sentence recognition was observed with as few as three bands [11]. This difference reflected most likely the low-context IEEE sentences and the sinusoidal carrier used in the present study as opposed to the high-context HINT sentences and the narrowband noise carrier used in the Shannon *et al.* study. Third, the FAME strategy produced significantly better performance than the AM-only strategy $[\mathrm{F}(1,24) = 249.4, \mathrm{p} < .001]$, with the 8-band FAME speech achieving similar performance to the natural speech. Finally, note that the greatest improvement was 35 percentage points at a 20-dB SNR for the 2-band processor, 54 percentage points at a 20-dB SNR for the 4-band processor, and 71 percentage points at a 5-dB SNR for the 8-band processor. This large improvement in sentence recognition suggests that extraction and encoding of frequency modulation is crucial for improving cochlear implant performance under realistic listening situations.

VI. Discussion

A. Contributions of AM and FM to Speech Perception

The present study has provided strong evidence for the complementary contribution of AM and FM cues to speech perception. While the AM cue from several frequency bands is sufficient to support speech recognition in quiet, the FM cue is critical for speech recognition in noise, particularly when the noise is a competing voice reflecting more realistic listening situations. How does FM complement AM to improve speech recognition in noise? Let us examine the 8-band, 0-dB SNR condition for sentence recognition with a competing voice, in which the AM-only cue produced a 5% correct score whereas the additional FM cue improved the performance to 55% correct (Fig. 9). Fig. 10 shows temporal envelopes from all eight frequency bands for the target sentence ("The meal was cooked before the bell rang") in Fig. 10(a), the masker sentence ("Port is a strong wine with a smoky taste") in Fig. 10(b), and the mixed target and masker sentences in Fig. 10(c). We can see clearly from Fig. 10 (a) and (b) that distinctive envelope peaks are present in the target and masker sentences. Some of these distinctive envelope peaks from the target and the masker are separately preserved even in the mixed signal, particularly in the high-frequency bands, while others combine to form modified envelope peaks due to their temporal overlap in the mixed signal, particularly in the low-frequency bands.

In the AM-only processing, the mixed envelopes were used to amplitude modulate a *common* carrier, impairing the listener's

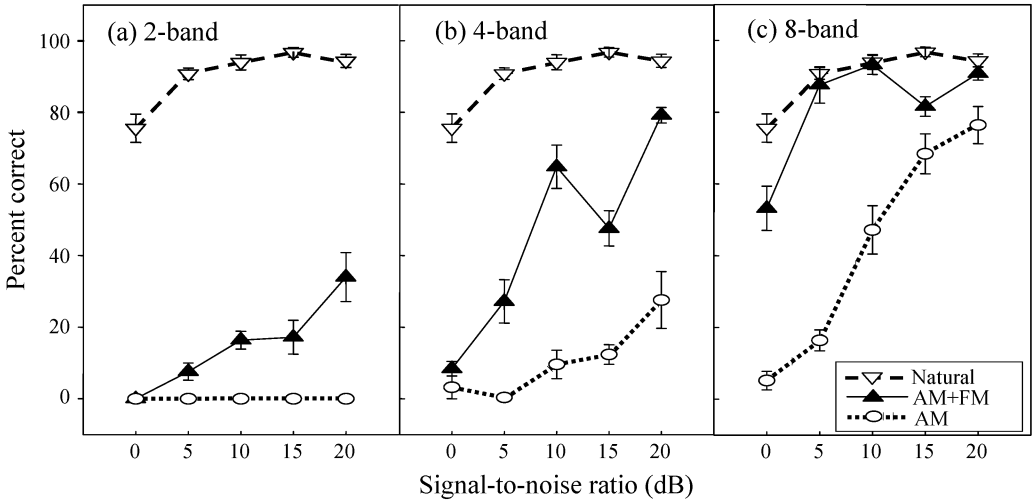

Fig. 9. Sentence recognition as a function of the number of bands (panels) and SNR (x-axis). (a)-(c) Show data from the 2-, 4-, and 8-band processor, respectively. The natural (unprocessed) data are represented by inverted triangles, while the AM-only and the AM+FM data are presented by open circles and filled triangles, respectively.

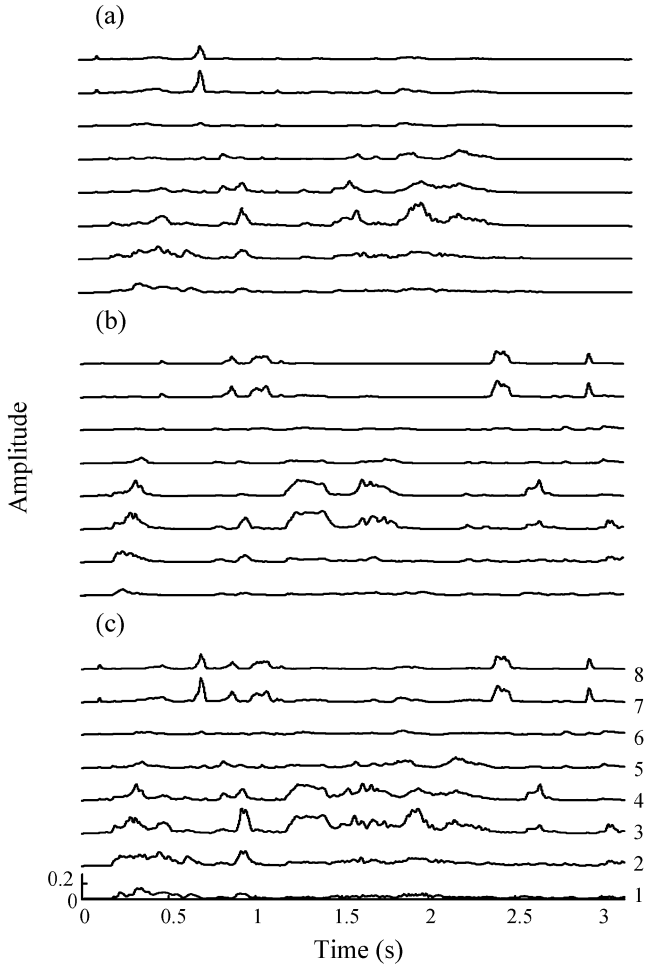

Fig. 10. (a) Shows the 8-band amplitude modulation signals of a target sentence ("The meal was cooked before the bell rang"). (b) Plots the same 8-band amplitude modulation signals of a competing sentence ("Port is a strong wine with a smoky taste"). (c) Shows the combined amplitude modulation signals from the mixed sentences at a 0-dB SNR.

ability to separate the signal from the noise. In cognitive science terms, the listener would not be able to tell which is the foreground and which is the background, and would most likely combine the envelope cues from both the target and the masker to form a single, warped perceptual stream [33]. In the FAME strategy, the carrier likely contains different FM cues for the target and masker, allowing the target envelopes to form one perceptual stream and the masker envelopes to form another stream for better sound segregation. Further study is needed to support this hypothesis.

B. Signal Processing for AM and FM Extraction

The present algorithm is based mostly on pioneering work on phase vocoders by Flanagan and his colleagues [25], [26]. We also note several recent studies on AM and FM representations of speech sounds. Mathematically, there are an infinite number of combinations that can decompose a signal into AM and FM components [24]. Both physical and functional constraints are needed to identify those combinations that are likely useful in real implementation. For example, combinations that produce negative frequencies would be difficult to interpret functionally [34], [35]. The Hilbert transform [12] and the energy operator model [36] are two popular methods to derive AM and FM components. These methods have been used to probe the independent contribution of the envelope and fine structure to auditory perception [12] and to derive efficient coding algorithms by dynamically tracking the AM and FM cues in speech sounds [37], [38]. However, these methods are difficult to apply directly to cochlear implants because the extracted FM generally varies too widely in range and too rapidly in rate.

C. Applications to Cochlear Implants

Both FM range and rate need to be limited for cochlear implant users, because perceptual data have demonstrated that they cannot detect FM range and rate above several hundred Hz [27], [28]. The present FAME strategy has reduced FM for cochlear implant users by removing the center frequency from the subband signal. The center frequency removal might be acceptable because, presumably, it would already be encoded by the position of the stimulating electrode. The present strategy has additionally limited the FM range and rate to several hundred Hertz,

which is within the perceptual abilities of most cochlear implant users. Below we propose several possible ways to implement the FAME strategy in current cochlear implants.

One suggestion would be to frequency-modulate the fixed pulse-rate carrier with the slowly varying FM signal in cochlear implants employing the CIS strategy. This frequency modulation would be additional to the amplitude modulation already implemented in the CIS strategy. Another way would be to replace the fixed pulse-rate carrier entirely with just the slowly varying FM signal. The latter should yield additional power saving because it employs a much slower rate of stimulation than the high-rate stimulation in typical CIS processors. In cochlear implants employing the N-of-M strategy, the FM signal can be implemented at least for the voiced segment of speech, which tends to be more stable and longer than the unvoiced segment. In cochlear implants employing analog-waveform strategies, both AM and FM components are technically present but might not be readily available to the cochlear implant user. This is because not only are the AM and FM cues in the subband signals still not delineated, but also the FM rate in the subband signals, particularly in the high-frequency bands, is too fast to be perceived. However, the basic idea behind the present strategy could be incorporated into these strategies by removing the center frequency of the analog subband electric signals. Finally, we note that the complicated interaction between place and time cues in actual cochlear implants requires future research to implement and evaluate frequency modulation in improving cochlear implant performance in noise.

VII. Conclusion

Based on previous work on phase vocoders, we have developed a novel algorithm to derive slowly varying amplitude and frequency modulations in speech sounds. We presented psychoacoustic data showing the complementary contribution of amplitude and frequency modulations to speech perception: amplitude modulation from a limited number of spectral bands is sufficient to support speech recognition in quiet but frequency modulation is needed to support speech recognition in noise. We have hypothesized an underlying mechanism by which frequency modulation is used to segregate different voices, allowing enhanced performance in noise, particularly when the noise is a competing voice. The present results strongly suggest that frequency modulation be encoded in cochlear implants. We have proposed several implementation methods to stimulate further investigation.

Acknowledgment

The authors would like to thank E. del Rio and S. Desai for data collection and manuscript proofreading. They also thank Associate Editor, Dr. I. Rybak, and three anonymous reviewers for their helpful comments on the manuscript.

References

[1] Q. J. Fu and R. V. Shannon, "Phoneme recognition by cochlear implant users as a function of signal-to-noise ratio and nonlinear amplitude mapping," *J. Acoust. Soc. Am.*, vol. 106, pp. L18–L23, 1999.

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

[3] F. G. Zeng and J. J. Galvin, 3rd, "Amplitude mapping and phoneme recognition in cochlear implant listeners," *Ear. Hear.*, vol. 20, pp. 60–74, 1999.

[4] G. Stickney, F.G. Zeng, R. Litovsky, and P. Assmann, "Cochlear implant speech recognition with speech masker," *J. Acoust. Soc. Am.*, vol. 116, pp. 1081–1091, 2004.

[5] G. Fant, A. Kruckenberg, and J. Liljencrants, "The source-filter frame of prominence," *Phonetica*, vol. 57, pp. 113–127, 2000.

[6] K. N. Stevens, "Acoustic correlates of some phonetic categories," *J. Acoust. Soc. Am.*, vol. 68, pp. 836–842, 1980.

[7] T. Dau, B. Kollmeier, and A. Kohlrausch, "Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers," *J. Acoust. Soc. Am.*, vol. 102, pp. 2892–2905, 1997.

[8] N. Suga, "Recovery cycles and responses to frequency modulated tone pulses in auditory neurons of echo-locating bats," *J. Physiol.*, vol. 175, pp. 50–80, 1964.

[9] X. Wang and M. B. Sachs, "Coding of envelope modulation in the auditory nerve and anteroventral cochlear nucleus," *Philos. Trans. Roy. Soc. Lond B. Biol. Sci.*, vol. 336, pp. 399–402, 1992.

[10] H. Dudly, "The vocoder," *Bell Labs Rec.*, vol. 17, pp. 122–126, 1939.

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

[12] Z. M. Smith, B. Delgutte, and A. J. Oxenham, "Chimaeric sounds reveal dichotomies in auditory perception," *Nature*, vol. 416, pp. 87–90, 2002.

[13] Y. Y. Kong, M. Vongphoe, and F. G. Zeng, "Independent contributions of amplitude modulation and frequency modulation to auditory perception: II. Melody, tone and speaker identification," in *Abstract 26th Annu. Midwinter Res. Meeting*, vol. 26, 2003, pp. 213–214.

[14] S. Sheft and W. A. Yost, "Auditory Abilities of Experienced Signal Analysts," Loyola University, Chicago, IL, AFRL Prog. Rep. 1, Contract SPO700-98-D-4002, 2001.

[15] D. K. Eddington, W. H. Dobelle, D. E. Brackmann, M. G. Mladevosky, and J. L. Parkin, "Auditory prosthesis research with multiple channel intracochlear stimulation in man," *Ann. Otol. Rhinol. Laryngol.*, vol. 87, pp. 1–39, 1978.

[16] B. S. Wilson, C. C. Finley, D. T. Lawson, R. D. Wolford, D. K. Eddington, and W. M. Rabinowitz, "Better speech recognition with cochlear implants," *Nature*, vol. 352, pp. 236–238, 1991.

[17] H. J. McDermott, C. M. McKay, and A. E. Vandali, "A new portable sound processor for the University of Melbourne/Nucleus Limited multielectrode cochlear implant," *J. Acoust. Soc. Am.*, vol. 91, pp. 3367–3371, 1992.

[18] M. W. Skinner, L. K. Holden, T. A. Holden, R. C. Dowell, P. M. Seligman, J. A. Brimacombe, and A. L. Beiter, "Performance of postlinguistically deaf adults with the wearable speech processor (WSP III) and mini speech processor (MSP) of the nucleus multi-electrode cochlear implant," *Ear. Hear.*, vol. 12, pp. 3–22, 1991.

[19] D. H. Johnson, "The relationship between spike rate and synchrony in responses of auditory-nerve fibers to single tones," *J. Acoust. Soc. Am.*, vol. 68, pp. 1115–1122, 1980.

[20] J. E. Rose, J. F. Brugge, D. J. Anderson, and J. E. Hind, "Phase-locked response to low-frequency tones in single auditory nerve fibers of the squirrel monkey," *J. Neurophysiol.*, vol. 30, pp. 769–793, 1967.

[21] D. Regan and B. W. Tansley, "Selective adaptation to frequency-modulated tones: Evidence for an information-processing channel selectively sensitive to frequency changes," *J. Acoust. Soc. Am.*, vol. 65, pp. 1249–1257, 1979.

[22] B. C. Moore and A. Sek, "Detection of frequency modulation at low modulation rates: Evidence for a mechanism based on phase locking," *J. Acoust. Soc. Am.*, vol. 100, pp. 2320–2331, 1996.

[23] B. W. Tansley and D. Regan, "Separate auditory channels for unidirectional frequency modulation and unidirectional amplitude modulation," *Sens. Processes*, vol. 3, pp. 132–140, 1979.

[24] P. Loughin and B. Tacer, "On the amplitude-and-frequency modulation decomposition of signals," *J. Acoust. Soc. Am.*, vol. 100, pp. 1594–1601, 1996.

[25] J. L. Flanagan, "Parametric coding of speech spectra," *J. Acoust. Soc. Am.*, vol. 68, pp. 412–419, 1980.

[26] J. L. Flanagan and R. M. Golden, "Phase vocoder," *Bell Syst. Tech. J.*, vol. 45, pp. 1493–1509, 1966.

[27] H. Chen and F. G. Zeng, "Frequency modulation detection in cochlear implant subjects," *J. Acoust. Soc. Am.*, vol. 116, pp. 2269–2277, 2004.

[28] F. G. Zeng, "Temporal pitch in electric hearing," *Hear. Res.*, vol. 174, pp. 101–106, 2002.
[29] N. Lan, K. B. Nie, S. K. Gao, and F. G. Zeng, "A novel speech-processing strategy incorporating tonal Information for cochlear implants," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 5, pp. 752–760, May 2004.
[30] J. Hillenbrand, L. A. Getty, M. J. Clark, and K. Wheeler, "Acoustic characteristics of American English vowels," *J. Acoust. Soc. Am.*, vol. 97, pp. 3099–3111, 1995.
[31] R. V. Shannon, A. Jensvold, M. Padilla, M. E. Robert, and X. Wang, "Consonant recordings for speech testing," *J. Acoust. Soc. Am.*, vol. 106, pp. L71–L74, 1999.
[32] M. L. Hawley, R. Y. Litovsky, and H. S. Colburn, "Speech intelligibility and localization in a multi-source environment," *J. Acoust. Soc. Am.*, vol. 105, pp. 3436–3448, 1999.
[33] A. S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound*. Cambridge, MA: The MIT Press, 1994.
[34] R. Kumaresan and Y. Wang, "Model-based approach to envelope and positive instantaneous frequency estimation of signals with speech application," *J. Acoust. Soc. Am.*, vol. 105, pp. 1912–1924, 1999.
[35] R. Kumaresan and Y. Wang, "On representing signals using only timing information," *J. Acoust. Soc. Am.*, vol. 110, pp. 2421–2439, 2001.
[36] P. Maragos, J. Kaiser, and T. Quatieri, "On amplitude and frequency demodulation using energy operators," *IEEE Trans. Signal Process.*, vol. 41, no. 4, pp. 1532–1550, Apr. 1993.
[37] P. Alexandros and P. Maragos, "A comparison of the energy operator and the Hilbert transform approach to signal and speech demodulation," *Signal Processing*, vol. 37, pp. 95–120, 1994.
[38] ——, "Speech analysis and synthesis using an AM-FM modulation," *Speech Commun.*, vol. 28, pp. 195–209, 1999.

Kaibao Nie (S'98–M'03) received the B.S. and M.S. degrees in electrical engineering from Shandong University, Jinan, China, in 1988 and 1991, respectively. In 1999, he received the Ph.D. degree in biomedical engineering from Tsinghua University, Beijing, China.

He is currently a Postdoctoral Researcher in the Departments of Biomedical Engineering and Otolaryngology at the University of California, Irvine. From 1999 to 2001, he was an Associate Professor at the Department of Electronic Engineering, Shandong University, China. He has published more than 50 scientific papers in peer-reviewed journals and conference proceedings. He is the author of three books chapters and he holds two patents. His research interests include speech signal processing, cochlear implants, speech perception, and biomedical instrumentation.

Dr. Nie is a Specially invited Expert of the city of Jinan, China. He is a reviewer for the Natural Science Foundation of China and the journal of IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING.

Ginger Stickney received the B.S. degree in cognitive science from the University of California, San Diego, in 1992, the M.S. degree in cognition and neuroscience in 1997, and audiology in 1998, and the Ph.D. degrees in communication sciences in 2001 from the University of Texas at Dallas.

From 1998 to 1999, she worked as an audiology clinical fellow with the cochlear implant team at the House Ear Institute. She is currently a Postdoctoral Scholar in the Department of Otolaryngology at the University of California, Irvine. Her research interests include speech perception, neural plasticity, and cochlear implants.

Dr. Stickney is a member of the Association for Research in Otolaryngology (ARO), the Acoustical Society of America (ASA), and the American Speech-Language-Hearing Association (ASHA).

Fan-Gang Zeng (S'88–M'91) received the B.S. degree in electrical engineering from the University of Science and Technology of China, in 1982, the M.S. degree in biomedical engineering from Shanghai Institute of Physiology, Academia Sinica in 1985, and the Ph.D. degree in hearing science from Syracuse University, New York, in 1990.

He joined the House Ear Institute in Los Angeles as a Research Associate from 1990 to 1992, Assistant Scientist from 1992 to 1994, and Associate Scientist and Director of the Auditory Perception Laboratory from 1994 to 1998. He held an Adjunct Associate Professorship in Electrical Engineering and taught graduate-level courses at the University of Southern California from 1996 to 1998. He was a tenured faculty member in Hearing and Speech Science, Neuroscience and Cognitive Science at the University of Maryland, College Park, from 1998 to 2000. Since 2000, he has been Research Director in the Department of Otolaryngology—Head and Neck Surgery, and Professor in the Departments of Anatomy and Neurobiology, Biomedical Engineering, Cognitive Sciences, and Otolaryngology, University of California, Irvine. He has published 50 peer-reviewed journal articles, 20 book chapters, and is the senior editor for a volume on cochlear implants in Springer Handbook of Auditory Research (Springer-Verlag, New York). He holds 5U.S. Patents and has given 80 invited presentations worldwide.

Dr. Zeng is an Associate Editor for IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING and *Journal of Speech, Language, and Hearing Research*. He has reviewed grants as a regular or *ad hoc* member for the National Institutes of Health, National Science Foundation, US-Israel Binational Science Foundation, The Australian Garnett Passe and Rodney Williams Memorial Foundation, British Wellcome Trust, and The Royal National Institute for Deaf People. He is on the Scientific Advisory Board at Audia Technology, Inc., Santa Clara, CA.