Speech recognition with amplitude and frequency modulations

Fan-Gang Zeng*†‡, Kaibao Nie*, Ginger S. Stickney*, Ying-Yee Kong*, Michael Vongphoe*, Ashish Bhargave*, Chaogang Wei†, and Keli Cao†

*Departments of Anatomy and Neurobiology, Biomedical Engineering, Cognitive Sciences, and Otolaryngology–Head and Neck Surgery, University of California, Irvine, CA 92697; and †Department of Otolaryngology–Head and Neck Surgery, Peking Union Medical College Hospital, Beijing 100730, China

Edited by Michael M. Merzenich, University of California, San Francisco, CA, and approved December 21, 2004 (received for review September 1, 2004)

Amplitude modulation (AM) and frequency modulation (FM) are commonly used in communication, but their relative contributions to speech recognition have not been fully explored. To bridge this gap, we derived slowly varying AM and FM from speech sounds and conducted listening tests using stimuli with different modulations in normal-hearing and cochlear-implant subjects. We found that although AM from a limited number of spectral bands may be sufficient for speech recognition in quiet, FM significantly enhances speech recognition in noise, as well as speaker and tone recognition. Additional speech reception threshold measures revealed that FM is particularly critical for speech recognition with a competing voice and is independent of spectral resolution and similarity. These results suggest that AM and FM provide independent yet complementary contributions to support robust speech recognition under realistic listening situations. Encoding FM may improve auditory scene analysis, cochlear-implant, and audio-coding performance.

auditory analysis | cochlear implant | neural code | phase | scene analysis

Acoustic cues in speech sounds allow a listener to derive not only the meaning of an utterance but also the speaker's identity and emotion. Most traditional research has taken a reductionist's approach in investigation of the minimal cues for speech recognition (1). Previous studies using either naturally produced whispered speech (2) or artificially synthesized speech (3, 4) have isolated and identified several important acoustic cues for speech recognition. For example, computers relying on primarily spectral cues and human cochlear-implant listeners relying on primarily temporal cues can achieve a high level of speech recognition in quiet (5–7). As a result, spectral and temporal acoustic cues have been interpreted as built-in redundancy mechanisms in speech recognition (8). However, this redundancy interpretation is challenged by the extremely poor performance of both computers and human cochlear implant users in realistic listening situations where noise is typically present (7, 9).

The goal of this study was to delineate the relative contributions of spectral and temporal cues to speech recognition in realistic listening situations. We chose three speech perception tasks that are known to be notoriously difficult for computers and human cochlear-implant users, including speech recognition with a competing voice, speaker recognition, and Mandarin tone recognition. We approached the issue by extracting slowly varying amplitude modulation (AM) and frequency modulation (FM) from a number of frequency bands in speech sounds and testing their relative contributions to speech recognition in acoustic and electric hearing. The AM-only speech has been used in previous studies (3, 10) and is considered to be an acoustic simulation of the cochlear implant (5). Different from previous studies using relatively "fast" FM to track formant changes in speech production (4, 11) or fine structure in speech acoustics (12, 13), the "slow" FM used here tracks gradual changes around a fixed frequency in the subband. We evaluated AM-only, AM+FM, and the original unprocessed stimuli by using speech recognition tasks in quiet and in noise, as well as speaker and Mandarin tone recognition. We hypothesized that if AM is sufficient for speech recognition, then the additional FM cue would not provide any advantage.

Methods

We conducted three experiments to test this hypothesis and additionally to resolve the difference in speech recognition between the previous and present studies. Exp. 1 processed stimuli to contain either the AM cue alone or both the AM and FM cues. The main parameter was the number of frequency bands varying from 1 to 34. Different from previous studies, Exp. 1 found that four AM bands were not enough to support good speech performance even in quiet. Exp. 2 was conducted to replicate previous studies with systematically controlled speech materials and processing parameters. Exp. 3 extended Exp. 1 by using a more sensitive and reliable speech reception threshold (SRT) measure, defined as the signal-to-noise ratio (SNR) necessary for a listener to achieve performance of 50% correct (14, 15).

Stimuli. Exp. 1 used three sets of stimuli to assess sentence recognition, speaker recognition, and Mandarin tone recognition. First, 60 Institute of Electrical and Electronic Engineers (IEEE) low-context sentences were used for speech recognition in quiet and in noise (16). A male talker produced all of the target sentences while a different male talker produced a single sentence serving as the competing voice ("Port is a strong wine with a smoky taste," duration = 3.6 sec). The SNR was fixed at 5 dB in the noise experiment. Second, 10 speakers (3 men, 3 women, 2 boys, and 2 girls) who spoke h/V/d words (where /V/ stands for a vowel), e.g., had, head, and hood, were used in the speaker recognition experiment (17). Third, 100 Chinese words, representing 25 consonant and vowel combinations and four tones, were used in the Mandarin tone recognition experiment (18). These 100 words were randomly selected from a corpus of 200 words produced by a female and a male talker. The Chinese words were presented in noise (5 dB SNR) with the noise spectral shape being identical to the long-term average spectrum of the 200 original words. Andio samples of these stimuli can be found in supporting information on the PNAS web site.

Exp. 2 compared speech recognition with four AM bands by using three types of speech materials, including the City University of New York (CUNY) sentences (19) used in the original study of Shannon *et al.* (3), the Hearing in Noise Test (HINT) sentences (15) used in the study of Dorman *et al.* (10), and the IEEE sentences used in Exp. 1 of the present study. The CUNY

This paper was submitted directly (Track II) to the PNAS office.

Freely available online through the PNAS open access option.

Abbreviations: AM, amplitude modulation; FM, frequency modulation; SRT, speech reception threshold; SNR, signal-to-noise ratio; IEEE, Institute of Electrical and Electronic Engineers; CUNY, City University of New York; HINT, Hearing in Noise Test.

‡To whom correspondence should be addressed. E-mail: fzeng@uci.edu.

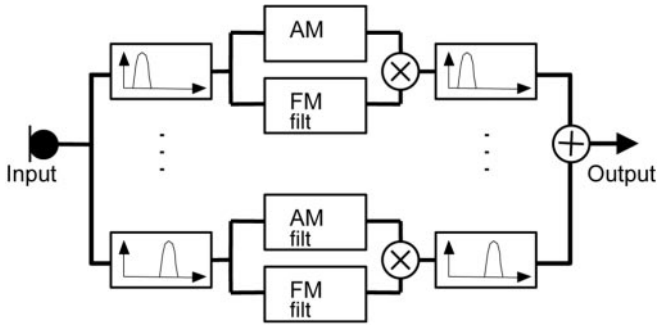

Fig. 1. Signal processing block diagram. The input signal is first filtered into a number of bands, and the band-limited AM and FM cues are then extracted. In the AM-only condition, the AM is modulated by either a noise or a sinusoid whose frequency is the bandpass filter's center frequency (not shown). In the AM+FM condition, the FM is smoothed in terms of both rate and depth and then modulated by the AM. In either condition, the same bandpass filter as in the analysis filter is applied before summation to control spectral overlap and resolution.

sentences are topic-related (e.g., "Make my steak well done." and "Do you want a toasted English muffin?"), the HINT sentences are all short declarative sentences (e.g., "A boy fell from the window."), whereas the IEEE sentences contain more complicated sentence structure and information (e.g., "The kite dipped and swayed, but stayed aloft."). Two other major differences between the present and previous studies were the overall processing bandwidth (4,000 vs. 8,800 Hz) and the type of carrier (noise vs. sinusoid). Both of these processing parameters were used in Exp. 2 to remove any potential confounding factors in data interpretation.

Exp. 3 used the HINT sentences to measure the speech reception threshold under three masker conditions, including a speech-spectrum-shaped steady-state noise from the original HINT program (15), a relatively long-duration sentence produced by the same male talker in the HINT program ("They broke all of the brown eggs," duration = 2.6 sec, fundamental frequency range = 100–140 Hz), and a similarly long sentence produced by a female talker in the IEEE sentence ("A pot of tea helps to pass the evening," duration = 2.7 sec, fundamental frequency range = 200–240 Hz). The original stimuli were presented to both normal-hearing and cochlear-implant subjects, whereas the 4-band and 34-band processed stimuli with AM and AM+FM cues were presented to only normal-hearing listeners. Thirty-four bands were used to match the number of auditory filters estimated psychophysically over the 80- to 8,800-Hz bandwidth (20).

Fig. 1 shows the block diagram for stimulus processing. To produce the AM-only and AM+FM stimuli, a stimulus was first filtered into a number of frequency analysis bands ranging from 1 to 34. The distribution of the cutoff frequencies of the bandpass filters was approximately logarithmic according to the Greenwood map (21). The band-limited signal was then decomposed by the Hilbert transform into a slowly varying temporal envelope and a relatively fast-varying fine structure (12, 22, 23). The slowly varying FM component was derived by removing the center frequency from the instantaneous frequency of the Hilbert fine structure and additionally by limiting the FM rate to 400 Hz and the FM depth to 500 Hz, or the filter's bandwidth, whichever was less (24). The AM-only stimuli were obtained by modulating the temporal envelope to the subband's center frequency and then summing the modulated subband signals (3, 10). The AM+FM stimuli were obtained by additionally frequency modulating each band's center frequency before amplitude modulation and subband summation. Before the subband summation, both the AM and the AM+FM processed subbands were subjected to the same bandpass filter as the corresponding analysis bandpass filter to prevent crosstalk between bands and the introduction of additional spectral cues produced by frequency modulation. All stimuli were presented at an average root-mean-square level of 65 dB (A weighted) with the exception of the SRT measure in Exp. 3, in which the noise was presented at 55 dBA and the signal level was varied adaptively.

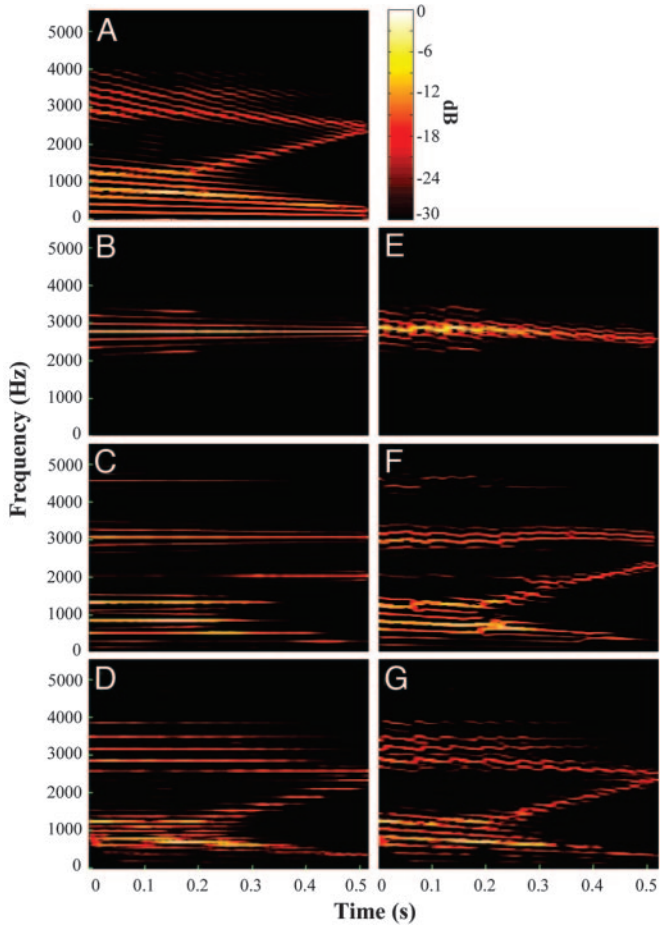

Fig. 2. Spectrograms of the original and processed speech sound /ai/. (*A*) The original speech: the thinner, slightly slanted lines represent a decrease in the fundamental frequency and its harmonics, whereas the thicker, slanted lines represent formant transitions. (*B–D*) The 1-, 8-, and 32-band amplitude modulation (AM) speech. (*E–G*) The 1-, 8-, and 32-band combined amplitude modulation and frequency modulation (AM+FM) speech.

Fig. 2*A* shows the original spectrogram for a synthetic speech syllable, /ai/, that contains both rich formant transitions (movement of energy concentration as a function of time) and fundamental frequency and its harmonic structure (downward slanted lines reflecting the decreasing fundamental frequency). Fig. 2 *B–D* shows spectrograms of 1-, 8-, and 32-band AM-only speech, respectively, whereas Fig. 2 *E–G* shows 1-, 8-, and 32-band AM+FM speech, respectively. First, we note that the original formant transition is not represented in the AM-only speech with few spectral bands (Fig. 2 *B* and *C*), and only crudely represented with 32 bands (Fig. 2*D*). In contrast, with as few as 8 bands, the AM+FM speech (Fig. 2*F*) preserves the original formant transition. Second, we note that the decreasing fundamental frequency in the original speech is represented with even the 1-band AM+FM speech (Fig. 2*E*) but not in any AM-processed speech. The acoustic analysis result indicates that the present slowly varying FM signal preserves dynamic information regarding formant and fundamental frequency movements.

Subjects. Thirty-four normal-hearing and 18 cochlear-implant subjects participated in Exp. 1. Twenty-four normal-hearing subjects (6 subjects × 4 conditions) heard IEEE sentences monaurally through headphones. One group of 12 subjects was presented with AM-only stimuli and the other group of 12 subjects received the AM+FM stimuli. Nine postlingually deafened cochlear-implant subjects (5 males and 4 females; 3 Clarion, 1 Med El, and 5 Nucleus users) also participated in the sentence recognition experiments in quiet and in noise. Six additional normal-hearing subjects and 8 of the 9 cochlear-implant subjects who participated in the sentence recognition experiment participated in the speaker recognition experiment. All subjects were native English speakers in the sentence and speaker recognition experiments. Finally, four normal-hearing and 9 cochlear-implant (Nucleus) subjects who were all native Mandarin speakers participated in the tone recognition experiment.

Eight additional normal-hearing subjects participated in Exp. 2. The same 8 normal-hearing subjects from Exp. 2 and the same 9 cochlear-implant subjects from Exp. 1 also participated in Exp. 3. Local Institutional Review Board approval and informed consent were obtained for all subjects.

Procedures. For sentence recognition in Exp. 1, all subjects heard six band conditions (1, 2, 4, 8, 16, and 32), each having 10 sentences per condition for a total of 60 sentences per subject. Sentences within each of the blocked band conditions were randomized, as well as the presentation order of the blocks themselves. No sentence was repeated. Before formal data collection, all subjects received a practice session of 12 sentences including 2 sentences per band condition. Percent correct scores, in terms of the number of keywords correctly identified, were obtained as the quantitative measure for each test condition.

A closed-set, forced-choice task was used in the speaker and tone recognition experiments. In the speaker experiment, the choice was one of the ten speakers labeled as Boy 1, Boy 2, Girl 1, Girl 2, Female 1, Female 2, Female 3, Male 1, Male 2, and Male 3, respectively. To familiarize subjects with the speaker's voice, all subjects received extensive training with the original speech samples until asymptotic performance was achieved. The training included an informal listening session with the subject either going through the entire speaker set or selecting a particular speaker for repeated listening, and an additional formal training session with the subject going through five or more practice sessions. In the tone experiment, the choice was one of the four tones labeled as tone 1 (-), tone 2 (/), tone 3 (V), and tone 4 (\), respectively. All subjects received a practice run for each condition in the tone experiment.

A within-subject design was used for sentence recognition in Exps. 2 and 3. Percent correct scores in terms of key words identified were used as the measure in Exp. 2. SRT was estimated by a one-down and one-up procedure in Exp. 3, in which a sentence from the HINT list was presented at a predetermined SNR based on the pilot study. The first sentence was repeated until it was 100% correctly identified. The SNR would be decreased by 2 dB should the subject correctly identify the sentence or increased by 2 dB should the subject incorrectly identify the sentence. The SRT was calculated as the mean SNR at which the subject went from correct to incorrect identification or vice versa, producing a performance level of 50% correct (15).

Results

Exp. 1: Effect of the Number of Analysis Bands. Fig. 3*A* shows sentence recognition in quiet as a function of the number of bands for normal-hearing subjects listening to both the AM-only (open circles) and AM+FM speech (solid triangles) as well as the cochlear-implant subjects listening to the original speech (hatched bars). Note first that the normal-hearing subjects

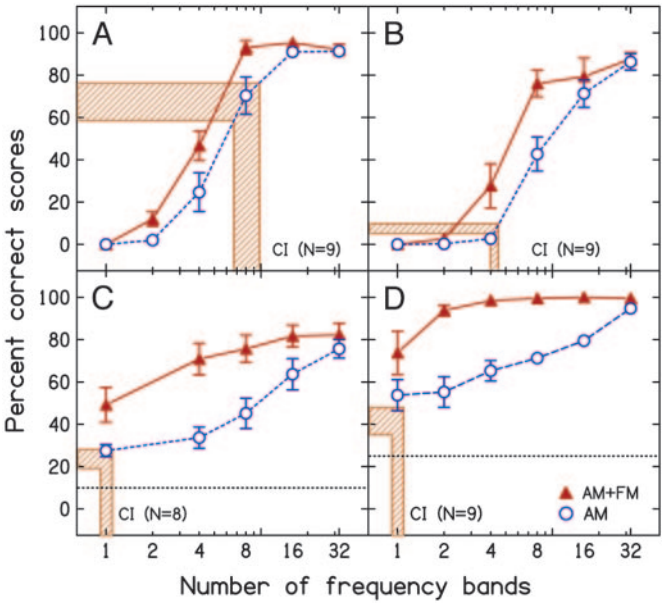

Fig. 3. Performance with AM and FM sounds. The normal-hearing subjects' data are plotted as a function of the number of frequency bands (symbols), whereas the cochlear-implant subjects' data are plotted as the hatched areas corresponding to the mean ± SE score on the *y* axis and its equivalent number of AM bands on the *x* axis. ○, AM condition; ▲, AM+FM condition. The error bars represent standard errors. CI, cochlear implant. (*A*) Sentence recognition in quiet. (*B*) Sentence recognition in noise (5 dB SNR). (*C*) Speaker recognition. Chance performance was 10%, represented by the dotted line. (*D*) Mandarin tone recognition. Chance performance was 25%, represented by the dotted line.

improved performance from zero percent correct with one band to nearly perfect performance as the number of bands increased [$F(5,50) = 291.8$, $P < 0.001$]. Second, The AM+FM speech produced significantly better performance than the AM-only speech [$F(1,10) = 6.9$, $P < 0.05$], with the largest improvement being 23 percentage points in the 8-band condition. Third, cochlear-implant subjects produced an average score of 68% correct with the original speech, which was equivalent to an estimated performance level of eight AM bands achieved by normal-hearing subjects (the hatched bar connected to the *x* axis).

Fig. 3*B* shows sentence recognition in the presence of a competing voice at a 5-dB SNR. The normal-hearing subjects produced effects similar to the quiet condition for both the number of bands [$F(5,50) = 177.6$, $P < 0.001$] and the AM+FM speech advantage [$F(1,10) = 11.4$, $P < 0.01$]. The largest FM advantage was 33 percentage points in the eight-band condition. Noise significantly degraded performance in both normal-hearing and cochlear-implant subjects ($P < 0.001$). Degraded performance was the least evident for the AM+FM speech (17 percentage points in the eight-band condition), more for the AM-only speech (27 percentage points in the eight-band condition), and the most for cochlear-implant subjects (60 percentage points). The score of 8% correct in cochlear-implant subjects was equivalent to normal performance with only four AM bands. Because significant ceiling and floor results were observed in the present experiment, we would use a more sensitive and reliable SRT measure to further quantify the FM advantage in Exp. 3.

Fig. 3*C* shows speaker recognition as a function of the number of bands. As a control, normal-hearing subjects achieved 86% correct asymptotic performance with the original unprocessed stimuli. Normal-hearing subjects produced significantly better performance with more bands [$F(4,20) = 72.8$, $P < 0.001$] and

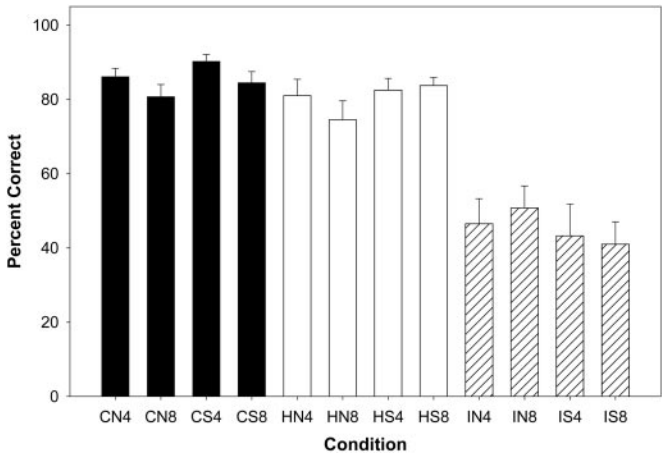

Fig. 4. Sentence recognition in quiet for the four-band condition. Speech materials include CUNY (filled bars), HINT (open bars), and IEEE sentences (hatched bars). Processing conditions include noise (N) and sinusoid (S) carriers as well as the 4,000-Hz (4) and 8,800-Hz (8) bandwidth. For example, CN4 stands for CUNY sentences with the noise carrier and a 4,000-Hz bandwidth.

with AM+FM speech than AM-only speech [$F(1,5) = 123.1, P < 0.001$]. The largest FM advantage was 37 percentage points in the four-band condition. Even with extensive training, cochlear-implant subjects scored on average only 23% correct in the speaker recognition task, equivalent to normal performance with one AM band. In contrast, the same implant subjects were able to achieve 65% correct performance on a vowel recognition task using the same stimulus set. This result indicates that current cochlear implant users can largely recognize what is said, but they cannot identify who says it.

Fig. 3*D* shows Mandarin tone recognition as a function of the number of bands. Normal-hearing subjects again produced significantly better performance with more bands [$F(5,15) = 13.2, P < 0.001$] and with the AM+FM speech than with the AM-only speech [$F(1,3) = 218.8, P < 0.001$]. The largest FM advantage was 39 percentage points in the two-band condition. Most strikingly, cochlear-implant subjects scored on average only 42% correct, barely equivalent to normal performance with one AM band.

Exp. 2: Effect of Speech Materials. One important difference between the previous and present studies was that previous studies showed nearly perfect performance in sentence recognition with only four AM bands (3, 10), whereas the present study found significantly lower performance. Fig. 4 shows speech recognition results from the same eight subjects who produced a high level of performance (80% correct or above) for the CUNY (filled bars) and HINT (open bars) but significantly lower performance for the IEEE (hatched bars) sentences [$F(2,14) = 107.0, P < 0.001$)]. Neither the bandwidth nor the carrier type produced a significant effect ($P > 0.1$), suggesting that the difference between the previous and present studies was mainly due to the type of speech material used. This result further casts doubt on the general utility of the AM cue in speech recognition even in quiet.

Exp. 3: Effect of FM on Different Maskers. Fig. 5*A* shows long-term amplitude spectra of the male talker target (solid line), the same male masker (dotted line), and the female masker (dashed line). Fig. 5*B* shows that the normal-hearing subjects (filled bars) produced a SRT that was 24 dB lower than the cochlear-implant subjects (open bars) across all three masker conditions [$F(1,7) = 107.4, P < 0.01$]. The normal-hearing subjects were able to take advantage of the talker differences, most likely in temporal fluctuation between the steady-state noise and the natural speech (25, 26) and in fundamental frequency between two different talkers (27, 28), producing systematically better SRT at −6 dB for the steady-state noise masker, −14 dB for the male masker, and −20 dB for the female masker [$F(2,14) = 47.2, P < 0.01$]. In contrast, the cochlear-implant subjects could not exploit these differences, and they performed similarly (SRT = 10–12 dB) for all three maskers [$F(2,16) = 3.1, P > 0.05$].

Fig. 5*C* shows that the amplitude spectrum for the original target sentences (solid line) was drastically different from the four-band counterparts with AM (dotted line) and AM+FM (dashed line) conditions. Because of the same postmodulation filtering, the AM and AM+FM processed spectra were similar, with peaks at the bandpass filters' center frequencies. Fig. 5*D* shows that the AM+FM four-band condition produced significantly better speech recognition in noise than the AM four-band condition [$F(1,7) = 52.1, P < 0.01$]. Interestingly, the additional FM cue produced a significant advantage in the male and female masker conditions (5- to 8-dB effect, $P < 0.01$) but not in the traditional steady-state noise condition ($P > 0.5$), reinforcing the hypothesis from the speaker identification result in Exp. 1 that the FM cue might allow the subjects to identify and separate talkers to improve performance in realistic listening situations. Finally, we noted that the four-band AM condition yielded virtually the same performance in the normal hearing subjects as in the cochlear-implant subjects [$F(1,7) = 0.4, P > 0.5$], independently verifying the result obtained in Exp. 1 that implant performance was equivalent to the four-band AM performance in noise conditions.

Fig. 5*E* shows that, when the number of bands was increased to 34, both the AM and AM+FM processed stimuli had amplitude spectra virtually identical to those of the original target stimuli. Still, the additional FM cue resulted in significantly better performance than the AM-only condition [$F(1,7) = 31.0, P < 0.01$]. Consistent with the four-band condition above, the FM advantage (3–5 dB) occurred only when a competing voice was used as a masker ($P < 0.05$) with the steady-state noise producing no statistically significant difference between the AM and AM+FM conditions ($P > 0.1$). Together, the present data suggest that the FM advantage is independent of both spectral resolution (the number of bands) and stimulus similarity (amplitude spectra).

Discussion

Traditional studies on speech recognition have focused on spectral cues, such as formants (1, 4), but recently attention has been turned to temporal cues, particularly the waveform envelope or AM cue (3, 10, 29, 30). Results from these recent studies have been over-interpreted to imply that only the AM cue is needed for speech recognition (31–33). The present result shows that the utility of the AM cue is seriously limited to ideal conditions (high-context speech materials and quiet listening environments). In addition, the present result demonstrates a striking contrast between the current cochlear implant users' ability to recognize speech and their inability to recognize speakers and tones. This finding further highlights the limitation of current cochlear implant speech processing strategies, as well as the need to encode the FM cue to improve speech recognition in noise, speaker identification, and tonal language perception.

To quantitatively address the acoustic mechanisms underlying the observed FM advantage, we calculated both modulation spectrum and amplitude spectrum for the original speech, the AM, and AM+FM processed speech as a function of the number of bands (see supporting information on the PNAS web site). We found that, independent of the number of bands, the modulation spectra were essentially identical for all three stimuli as the same AM cue was present in all stimuli. On the other hand, the

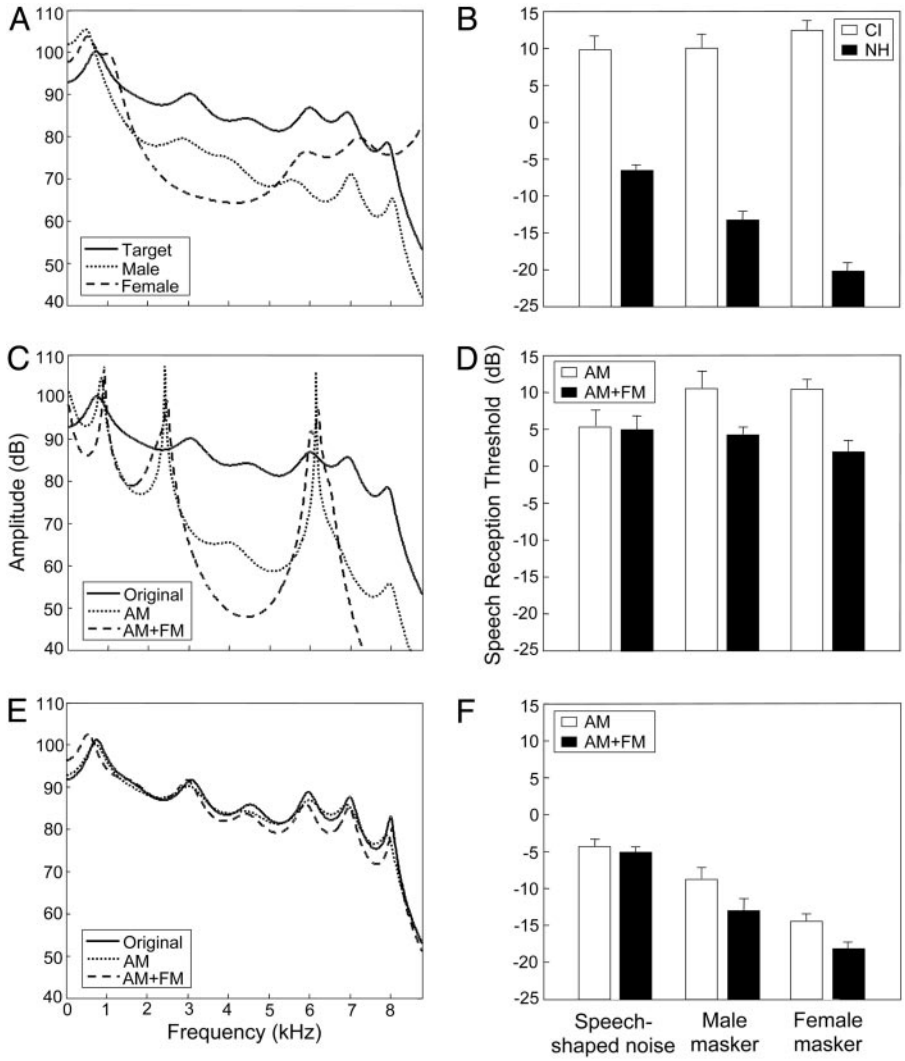

Fig. 5. Amplitude spectra (14th-order linear-predictive-coding smoothed; *Left*) and speech reception thresholds (*Right*). (*A*) Original speech spectra for the target sentences (solid line), a different sentence by the same male talker as the competing voice (dotted line), and a female competing voice (dashed line). (*B*) SRT measures for the cochlear-implant (open bars) and normal-hearing (filled bars) subjects. (*C*) Amplitude spectra for the original sentences (solid line, the same as in *A*) and their four-band counterparts with the AM (dotted line) and the AM+FM (dashed line) conditions. (*D*) SRT measures for the 4-band AM (open bars) and AM+FM (filled bars) conditions in the normal-hearing subjects. (*E*) Amplitude spectra for the original sentences (solid line, the same as in *A* and *B*) and their 34-band counterparts with the AM (dotted line) and the AM+FM (dashed line) conditions. (*F*) SRT measures for the 34-band AM (open bars) and AM+FM (filled bars) conditions in the normal-hearing subjects.

amplitude spectra were always similar between the AM and AM+FM processed speech but were significantly different from the original speech's amplitude spectrum when the number of bands was small (e.g., Fig. 5*C*). Careful examination further suggests that the FM advantage cannot be explained by the traditional measure in spectral similarity. Measured by the Euclidean distance between two spectra, a 16-band AM stimulus was 1.5 times more similar to the original stimulus than an 8-band AM+FM stimulus. However, the 8-band AM+FM stimulus outperformed the 16-band AM stimulus by 2, 11, 12, and 20 percentage points for sentence recognition in quiet, sentence recognition in noise, speaker recognition, and tone recognition, respectively (Fig. 3).

Because the FM cue is derived from phase, the present study argues strongly for the importance of phase information in realistic listening situations. We note that for at least two decades phase has been suggested to play a critical role in human perception (34), yet it has received little attention in the auditory field. If anything, recent studies seemed to have implicated a diminished role of the phase in speech recognition (3, 35, 36). The present result shows that phase information may not be needed in simple listening tasks but is critically needed in challenging tasks, such as speech recognition with a competing voice.

Implications for Cochlear Implants. The most direct and immediate implication is to improve signal processing in auditory prostheses. Currently, cochlear implants typically have 12–22 physical electrodes, but a much smaller number of functional channels as measured by speech performance in quiet (37). The present result strongly suggests that frequency modulation in addition to amplitude modulation should be extracted and encoded to improve cochlear implant performance. Recent perceptual tests have shown that cochlear implant subjects are capable of detecting these slowly varying frequency modulations by electric stimulation (38).

Implications for Audio Coding. Current audio coding schemes mostly have taken advantage of perceptual masking in the

frequency domain (39). The present result suggests that, within each subband, encoding of the slowly varying changes in intensity and frequency may be sufficient to support high-quality auditory perception. The coding efficiency can be improved, particularly for high-frequency bands, because the required sampling rate would be much lower to encode intensity and frequency changes with several hundred Hertz bandwidths than to encode, for example, a high-frequency subband at 5,000 Hz.

Implications for Speech Coding. The present study also suggests that frequency modulation serves as a salient cue that allows a listener to separate, and then assign appropriately, the amplitude modulation cues to form foreground and background auditory objects (40–42). Frequency modulation extraction and analysis can be used to serve as front-end processing to help solve, automatically, the segregating and binding problem in complex listening environments, such as at a cocktail party or in a noisy cockpit.

Implications for Neural Coding. FM has been shown to be a potent cue in vocal communication and auditory perception, such as in echo localization (43, 44). However, it is debatable whether amplitude and frequency modulations are processed in the same or different neural pathways (45–47). It is possible to generate novel stimuli that contain different proportions of amplitude and frequency modulations to probe the structure–function pathways, such as the hypothesized "speech" vs. "music" and "what" vs. "where" processing streams in the brain (48, 49).

We thank Lin Chen, Abby Copeland, Sheng Liu, Michelle McGuire, Mitch Sutter, Mario Svirsky, and Xiaoqin Wang for comments on the manuscript and the PNAS member editor, Michael M. Merzenich, and reviewers for motivating Exps. 2 and 3. This study was supported in part by the U.S. National Institutes of Health (Grant 2R01 DC02267 to F.-G.Z.) and the Chinese National Natural Science Foundation (Grants 39570756 and 30228021 to K.C. and Lin Chen).

1. Cooper, F. S., Liberman, A. M. & Borst, J. M. (1951) *Proc. Natl. Acad. Sci. USA* **37,** 318–325.
2. Tartter, V. C. (1991) *Percept. Psychophys.* **49,** 365–372.
3. Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J. & Ekelid, M. (1995) *Science* **270,** 303–304.
4. Remez, R. E., Rubin, P. E., Pisoni, D. B. & Carrell, T. D. (1981) *Science* **212,** 947–949.
5. Wilson, B. S., Finley, C. C., Lawson, D. T., Wolford, R. D., Eddington, D. K. & Rabinowitz, W. M. (1991) *Nature* **352,** 236–238.
6. Skinner, M. W., Holden, L. K., Whitford, L. A., Plant, K. L., Psarros, C. & Holden, T. A. (2002) *Ear Hear.* **23,** 207–223.
7. Rabiner, L. (2003) *Science* **301,** 1494–1495.
8. Schroeder, M. R. (1966) *Proc. IEEE* **54,** 720–734.
9. Zeng, F. G. & Galvin, J. J. (1999) *Ear Hear.* **20,** 60–74.
10. Dorman, M. F., Loizou, P. C. & Rainey, D. (1997) *J. Acoust Soc. Am.* **102,** 2403–2411.
11. Alexandros, P. & Maragos, P. (1999) *Speech Commun.* **28,** 195–209.
12. Smith, Z. M., Delgutte, B. & Oxenham, A. J. (2002) *Nature* **416,** 87–90.
13. Sheft, S. & Yost, W. A. (2001) *Air Force Research Laboratory Progress Report No. 1, Contract SPO700-98-D-4002* (Loyola University, Chicago).
14. Plomp, R. & Mimpen, A. M. (1979) *Audiology* **18,** 43–52.
15. Nilsson, M., Soli, S. D. & Sullivan, J. A. (1994) *J. Acoust. Soc. Am.* **95,** 1085–1099.
16. Hawley, M. L., Litovsky, R. Y. & Colburn, H. S. (1999) *J. Acoust. Soc. Am.* **105,** 3436–3448.
17. Hillenbrand, J., Getty, L. A., Clark, M. J. & Wheeler, K. (1995) *J. Acoust. Soc. Am.* **97,** 3099–3111.
18. Wei, C. G., Cao, K. & Zeng, F. G. (2004) *Hear. Res.* **197,** 87–95.
19. Boothroyd, A. (1985) *J. Speech Hear. Res.* **28,** 185–196.
20. Moore, B. C. & Glasberg, B. R. (1987) *Hear. Res.* **28,** 209–225.
21. Greenwood, D. D. (1990) *J. Acoust. Soc. Am.* **87,** 2592–2605.
22. Flanagan, J. L. & Golden, R. M. (1966) *Bell Syst. Tech. J.* **45,** 1493–1509.
23. Flanagan, J. L. (1980) *J. Acoust. Soc. Am.* **68,** 412–419.
24. Nie, K., Stickney, G. & Zeng, F. G. (2005) *IEEE Trans. Biomed. Eng.* **52,** 64–73.
25. Bacon, S. P., Opie, J. M. & Montoya, D. Y. (1998) *J. Speech Lang. Hear. Res.* **41,** 549–563.
26. Nelson, P. B. & Jin, S. H. (2004) *J. Acoust. Soc. Am.* **115,** 2286–2294.
27. Chalikia, M. H. & Bregman, A. S. (1993) *Percept. Psychophys.* **53,** 125–133.
28. Culling, J. F. & Darwin, C. J. (1993) *J. Acoust. Soc. Am.* **93,** 3454–3467.
29. Van Tasell, D. J., Soli, S. D., Kirby, V. M. & Widin, G. P. (1987) *J. Acoust. Soc. Am.* **82,** 1152–1161.
30. Rosen, S. (1992) *Philos. Trans. R. Soc. London B* **336,** 367–373.
31. Hopfield, J. J. (2004) *Proc. Natl. Acad. Sci. USA* **101,** 6255–6260.
32. Faure, P. A., Fremouw, T., Casseday, J. H. & Covey, E. (2003) *J. Neurosci.* **23,** 3052–3065.
33. Liegeois-Chauvel, C., Lorenzi, C., Trebuchon, A., Regis, J. & Chauvel, P. (2004) *Cereb. Cortex* **14,** 731–740.
34. Oppenheim, A. V. & Lim, J. S. (1981) *Proc. IEEE* **69,** 529–541.
35. Greenberg, S. & Arai, T. (2001) in *Proceedings of the Seventh Eurospeech Conference on Speech Communication and Technology (Eurospeech-2001)*, ed. Dalsgaard, P. (Int. Speech Commun. Assoc., Grenoble, France), Vol. 3, pp. 473–476.
36. Saberi, K. & Perrott, D. R. (1999) *Nature* **398,** 760.
37. Fishman, K. E., Shannon, R. V. & Slattery, W. H. (1997) *J. Speech Lang. Hear. Res.* **40,** 1201–1215.
38. Chen, H. & Zeng, F. G. (2004) *J. Acoust. Soc. Am.* **116,** 2269–2277.
39. Schroeder, M. R., Atal, B. S. & Hall, J. L. (1979) *J. Acoust. Soc. Am.* **66,** 1647–1652.
40. Bregman, A. S. (1994) *Auditory Scene Analysis: The Perceptual Organization of Sound* (MIT Press, Cambridge, MA).
41. Culling, J. F. & Summerfield, Q. (1995) *J. Acoust. Soc. Am.* **98,** 837–846.
42. Marin, C. M. & McAdams, S. (1991) *J. Acoust. Soc. Am.* **89,** 341–351.
43. Suga, N. (1964) *J. Physiol.* **175,** 50–80.
44. Doupe, A. J. & Kuhl, P. K. (1999) *Annu. Rev. Neurosci.* **22,** 567–631.
45. Liang, L., Lu, T. & Wang, X. (2002) *J. Neurophysiol.* **87,** 2237–2261.
46. Moore, B. C. & Sek, A. (1996) *J. Acoust. Soc. Am.* **100,** 2320–2331.
47. Saberi, K. & Hafter, E. R. (1995) *Nature* **374,** 537–539.
48. Rauschecker, J. P. & Tian, B. (2000) *Proc. Natl. Acad. Sci. USA* **97,** 11800–11806.
49. Zatorre, R. J., Belin, P. & Penhune, V. B. (2002) *Trends Cogn. Sci.* **6,** 37–46.