Temporal properties in clear speech perception

Sheng Liu
Hearing and Speech Research Laboratory, Department of Biomedical Engineering, University of California, Irvine, Irvine, California 92697

Fan-Gang Zeng[a)]
Hearing and Speech Research Laboratory, Departments of Anatomy and Neurobiology, Biomedical Engineering, Cognitive Sciences, and Otolaryngology-Head and Neck Surgery, University of California, Irvine, Irvine, California 92697

(Received 15 January 2005; revised 28 April 2006; accepted 4 May 2006)

Three experiments were conducted to study relative contributions of speaking rate, temporal envelope, and temporal fine structure to clear speech perception. Experiment I used uniform time scaling to match the speaking rate between clear and conversational speech. Experiment II decreased the speaking rate in conversational speech without processing artifacts by increasing silent gaps between phonetic segments. Experiment III created "auditory chimeras" by mixing the temporal envelope of clear speech with the fine structure of conversational speech, and vice versa. Speech intelligibility in normal-hearing listeners was measured over a wide range of signal-to-noise ratios to derive speech reception thresholds (SRT). The results showed that processing artifacts in uniform time scaling, particularly time compression, reduced speech intelligibility. Inserting gaps in conversational speech improved the SRT by 1.3 dB, but this improvement might be a result of increased short-term signal-to-noise ratios during level normalization. Data from auditory chimeras indicated that the temporal envelope cue contributed more to the clear speech advantage at high signal-to-noise ratios, whereas the temporal fine structure cue contributed more at low signal-to-noise ratios. Taken together, these results suggest that acoustic cues for the clear speech advantage are multiple and distributed. © *2006 Acoustical Society of America.*
[DOI: 10.1121/1.2208427]

PACS number(s): 43.71.Es, 43.71.Gv, 43.71.Ky [ALF] Pages: 424–432

I. INTRODUCTION

When speech communication becomes difficult, a talker may adopt a different style of speech, "clear speech." This style differs from the everyday speech style, herein referred to as "conversational speech." Previous studies have demonstrated a significant intelligibility advantage for clear speech over conversational speech in both normal-hearing and hearing-impaired listeners across a wide range of listening conditions including quiet, noisy, and reverberant backgrounds (Chen, 1980; Picheny *et al.*, 1985; Payton *et al.*, 1994; Gagne *et al.*, 1995; Schum, 1996; Uchanski *et al.*, 1996; Helfer, 1997; Bradlow and Bent, 2002; Ferguson and Kewley-Port, 2002; Gagne *et al.*, 2002; Krause and Braida, 2002; Bradlow *et al.*, 2003; Liu *et al.*, 2004). Several acoustic differences have been identified between clear and conversational speech, including slower speaking rate, greater temporal modulation, enhanced fundamental frequency variation, expanded vowel space, and higher energy distribution at high frequencies for clear speech (Picheny *et al.*, 1986; Payton *et al.*, 1994; Uchanski *et al.*, 1996; Krause and Braida, 2002;2004; Liu *et al.*, 2004). However, the exact acoustic cues that are responsible for the clear speech advantage remain largely elusive. The present study focuses on the role of temporal information in clear speech perception. Three temporal properties, including speaking rate, temporal envelope, and temporal fine structure, are examined.

Speaking rate is determined by both word and pause durations and is one of the most extensively studied temporal characteristics in speech perception. As early as the 1950s, Fairbanks and colleagues used magnetic tape recorders with different playback speeds to perform uniform time compression and expansion of speech sounds (Fairbanks *et al.*, 1954). Depending on the original speaking rate and the talker's gender, normal-hearing listeners could generally tolerate time-compressed and expanded speech for ratios up to two (Fairbanks *et al.*, 1957; Beasley *et al.*, 1972). However, elderly listeners and persons with certain central auditory processing disorders were found to have a particular difficulty in perceiving the time-compressed speech (Kurdziel *et al.*, 1976; Gordon-Salant and Fitzgibbons, 1995;1997).

Using an explicit pitch-tracking method, coupled with manipulations of the input and output sampling rates, more recent digital time-scaling algorithms could uniformly time compress and expand speech without changing voice pitch (Malah, 1979). Picheny *et al.* (1989) used Malah's algorithm to increase the clear speech rate to match the naturally produced conversational speech rate and to decrease the conversational speech rate to match the naturally produced clear speech rate. They found that uniform time scaling degraded speech intelligibility for both sped-up clear speech and slowed-down conversational speech, suggesting that digital artifacts contaminated the results.

[a)]Author to whom correspondence should be addressed. Electronic mail: fzeng@uci.edu

 0001-4966/2006/120(1)/424/9/$22.50

Uchanski *et al.* (1996) used nonuniform time scaling to alter phonetic segment lengths to reflect previously measured segmental durational differences between clear and conversational speech. The nonuniform time-scaling method still produced lower intelligibility than unprocessed speech, but the degree of degradation was much less than uniform time scaling. Krause and Braida (2002; 2004) employed "natural" clear speech, training talkers to produce clear speech with the same speaking rate as conversational speech. The talkers were able to produce "fast" clear speech that had the same speaking rate as conversational speech. Perceptual results still showed significantly higher intelligibility for "fast" clear speech than same-rate conversational speech, but there was a global trend of decreasing intelligibility with increasing speaking rate for both clear and conversational speech.

Following this line of research, we employed two different techniques to further probe the role of speaking rate in clear speech perception. Experiment I employed newer signal-processing algorithms, which introduced fewer digital artifacts than algorithms in the 1980s, to uniformly time-scaled speech (Moulines and Laroche, 1995; Kawahara *et al.*, 1999). These newer time-scaling algorithms typically utilized sophisticated pitch extraction algorithms (e.g., pitch synchronous overlap add method or PSOLA) and avoided producing tonal noise in voiceless fricatives and degrading transitional portions in stop consonants. Experiment I is consistent with a recent trend in which classic speech studies are replicated using new digital signal-processing technology (Liu and Kewley-Port, 2004; Assmann and Katz, 2005).

Experiment II decreased speaking rate by inserting silent gaps between phonemes in the conversational speech. This experiment was partially motivated by recent studies which showed a high correlation between temporal processing and speech perception in special populations, including elderly listeners, cochlear-implant users, and persons with auditory neuropathy (Gordon-Salant and Fitzgibbons, 1997; Zeng *et al.*, 1999; Fu, 2002; Zeng *et al.*, 2005a). Inserting gaps between speech segments increased amplitude modulation and provided an extended time window, allowing the listener to process speech more efficiently. The present study differs from Uchanski's nonuniform time-scaling study in the following three ways. First, we did not attempt to match phoneme durations between clear and conversation speech. Instead, we inserted silent gaps proportionally in conversational speech so that the gap-inserted conversational speech had the same overall duration as the clear speech but was free of digital processing artifacts. Second, different from the 10% change in the overall duration in the Uchanski *et al.* (1996) study, we increased the average sentence duration (1.31 seconds) in the original conversational speech by 50% to match the average sentence duration (1.97 seconds) found in clear speech (Liu *et al.*, 2004). Finally, we used different speech materials (BKB sentences) than the nonsense sentences used in the Uchanski *et al.* study.

In addition to speaking rate, other temporal properties play a significant role in clear speech perception. Rosen (1992) divided temporal information into three categories according to the rate of wave fluctuations: envelope (2–50 Hz), periodicity (50–500 Hz), and fine structure (500–10 000 Hz). The temporal envelope cue from a limited number of spectral channels has been shown to be sufficient for speech recognition in quiet (Dudley, 1939; Houtgast and Steeneken, 1985; Van Tasell *et al.*, 1987; Drullman, 1995; Shannon *et al.*, 1995), but periodicity and fine structure are critical for speech recognition in noise, particularly when the noise is a competing voice (Nelson *et al.*, 2003; Qin and Oxenham, 2003; Stickney *et al.*, 2004; Kong *et al.*, 2005; Nie *et al.*, 2005; Zeng *et al.*, 2005b). It is possible that all three temporal cues are enhanced in clear speech (Bradlow *et al.*, 2003; Krause and Braida, 2004; Liu *et al.*, 2004); however, no study has directly assessed the relative contributions of these temporal cues to clear speech perception.

Experiment III used a novel processing scheme called "auditory chimera" to examine the relative contributions of temporal envelope and fine structure cues to clear speech perception (Smith *et al.*, 2002). The "chimera" scheme is reminiscent of previous cue-trading studies in the segmental domain, in which conflicting burst release and formant transition cues were combined in a single synthetic stimulus to examine their relative contribution to stop consonant recognition (Walley and Carrell, 1983; Dorman and Loizou, 1996). To synthesize a chimaeric sound, Smith *et al.* first divided two broadband signals into several sub-bands, then used the Hilbert transform to extract the temporal envelope and fine structure in each sub-band, and finally mixed one signal's temporal envelope with another signal's fine structure. Smith *et al.* tested the intelligibility of chimerized speech and found that the temporal envelope, rather than the temporal fine structure, made the most contributions to speech intelligibility. However, Smith *et al.* did not test speech recognition in noise, nor did they use any clear speech materials.

In summary, the present study conducted three experiments to evaluate the relative contributions of speaking rate, temporal envelope, and temporal fine structure to the clear speech advantage. Experiment I measured speech intelligibility as a function of signal-to-noise ratios using processed conversational speech that was uniformly stretched to match the duration of the clear speech, or by using processed clear speech that was uniformly compressed to match the duration of the conversational speech. Experiment II measured speech intelligibility using only processed conversational speech that was nonuniformly stretched by proportionally increasing silent gaps between phonetic segments to match the duration of the clear speech. Experiment III measured "chimerized" speech intelligibility using processed speech that contained either the clear speech envelope with conversational speech fine structure or the conversational speech envelope with clear speech fine structure. If a chimera containing the clear speech envelope produces the highest intelligibility, we would conclude that the envelope characteristics of clear speech are responsible for the clear speech advantage. If, in contrast, a chimera containing the clear speech fine structure cues is more intelligible, we would then reach a different conclusion that the fine structure characteristics of clear speech are responsible for the clear speech advantage.

II. EXPERIMENT I. UNIFORMLY TIME-SCALED SPEECH

A. Methods

1. Subjects

Ten normal-hearing listeners were recruited from the Undergraduate Social Science Subject Pool at the University of California, Irvine. Local Institutional Review Board approval was obtained for the experimental protocol. Informed consent was also obtained for each individual subject. None of the subjects reported any speech or hearing impairment. All subjects were native English speakers and received course credit for their participation. Five of the ten subjects were tested with original and processed fast clear speech, while the other five subjects were tested with original and processed slow conversational speech.

2. Stimuli

The original stimuli consisted of 144 sentences recorded in both clear and conversational speech styles. The sentences were modified from the original Bamford-Kowal-Bench (BKB) sentences used for British children (Bench and Bamford, 1979). A male adult talker recorded these sentences with a sampling rate of 16 KHz in a sound-treated room at the Phonetics Laboratory of the Department of Linguistics at Northwestern University (Bradlow *et al.*, 2003).

COOL EDIT PRO 2 (currently known as ADOBE AUDITION) was used to uniformly stretch or compress the original speech signal to change the speaking rate without changing the pitch. The processing algorithm was based on the pitch-synchronous overlap and add method (PSOLA) (Moulines and Laroche, 1995). First, the input waveform was decomposed into a stream of short-time signals based on pitch-synchronous marks. Second, the pitch-synchronous short-time signal was either eliminated or duplicated based on the predefined stretch factor. Third, the modified short-time signal was added to synthesize the stretched and compressed stimulus. The original pitch was preserved during processing and the duration of each voiced or silent segment in the speech was uniformly changed. Different from the earlier methods that changed sampling rate to perform time scaling (Malah, 1979), the newer algorithms used large units (i.e., pitch periods) to perform time scaling, reversed segments to avoid tonal noise in fricative consonants, and preserved the transitional properties in stop consonants. Presumably, these manipulations introduced minimal digital artifacts.

The speaking rate of clear speech was increased to match the speaking rate of conversational speech, and the speaking rate of conversational speech was decreased to match the rate of clear speech for each individual sentence. On average, the speaking rate was increased by 33% for the sped-up clear speech or decreased by 50% for the slowed-down conversational speech (the average duration for conversational speech was 1.31 s compared with 1.97 s for clear speech). Figure 1 shows spectrograms of the original clear speech (top left panel), the original conversational speech (top right), the processed slow conversational speech (left panel on the second row), and the processed fast clear speech (right panel on the second row). Note that the overall duration at the sentence level was matched, but the duration at the phonemic level was clearly not matched. Additionally, note the smeared harmonic structure and formant transitions in both types of processed speech.

All sentences were normalized to have the same overall root-mean-square (rms) level. The speech presentation level was fixed at 65 dBA. The noise level was varied to produce different signal-to-noise ratios. The speech signal was digitally mixed with a speech-spectrum-shaped noise to produce signal-to-noise ratios ranging from −15 to +10 dB.

3. Procedure

Normal-hearing subjects listened to the stimuli monaurally presented via Sennheiser HDA 200 headphones in an IAC double-walled, sound-treated booth. Sentences were presented only once for each subject over the course of the entire experiment. All subjects went through a practice session consisting of five sentences in quiet to become familiar with the test materials and procedures. To collect data, subjects were asked to type the sentences presented through the headphones. A MATLAB program recorded the subject's response and reminded the subject to double-check the spelling before accepting each answer. Speech recognition scores were automatically calculated by counting the number of correct keywords identified.

Experiment I had a total of 28 listening conditions, including the original and processed stimuli $(2)\times 2$ speech styles $\times 7$ signal-to-noise ratios from −15 to 10 dB in 5-dB steps and in quiet. Each condition had eight sentences containing three to four keywords each. The average percent-correct score from eight sentences was reported. In addition, the speech reception threshold (SRT) corresponding to the 50% correct score and the dynamic range (DR) corresponding to the dB difference between the signal-to-noise ratios producing 10% and 90% of the asymptotic performance was derived from the psychometric function (Zeng and Galvin, 1999; Liu *et al.*, 2004).

A mixed-design ANOVA was performed with speech style as a between-subjects factor and processing and signal-to-noise ratio as within-subjects factors. The processing factor examined the difference in performance between the original clear speech and the processed fast clear speech, as well as between the original conversational speech and processed slow conversational speech. A difference was significant at the 0.05 level.

B. Results and discussion

Figure 2 shows percent-correct scores as a function of signal-to-noise ratio obtained from the original clear (open circles with the solid line), original conversational (filled circles with the dotted line), uniformly time-scaled slow conversational (filled triangles with the dashed line), and uniformly time-scaled fast clear (open triangles with the dot-dashed line) speech. Table I shows three fitting parameters and two derived parameters for the perceptual data from experiment I. Several observations can be made from these data.

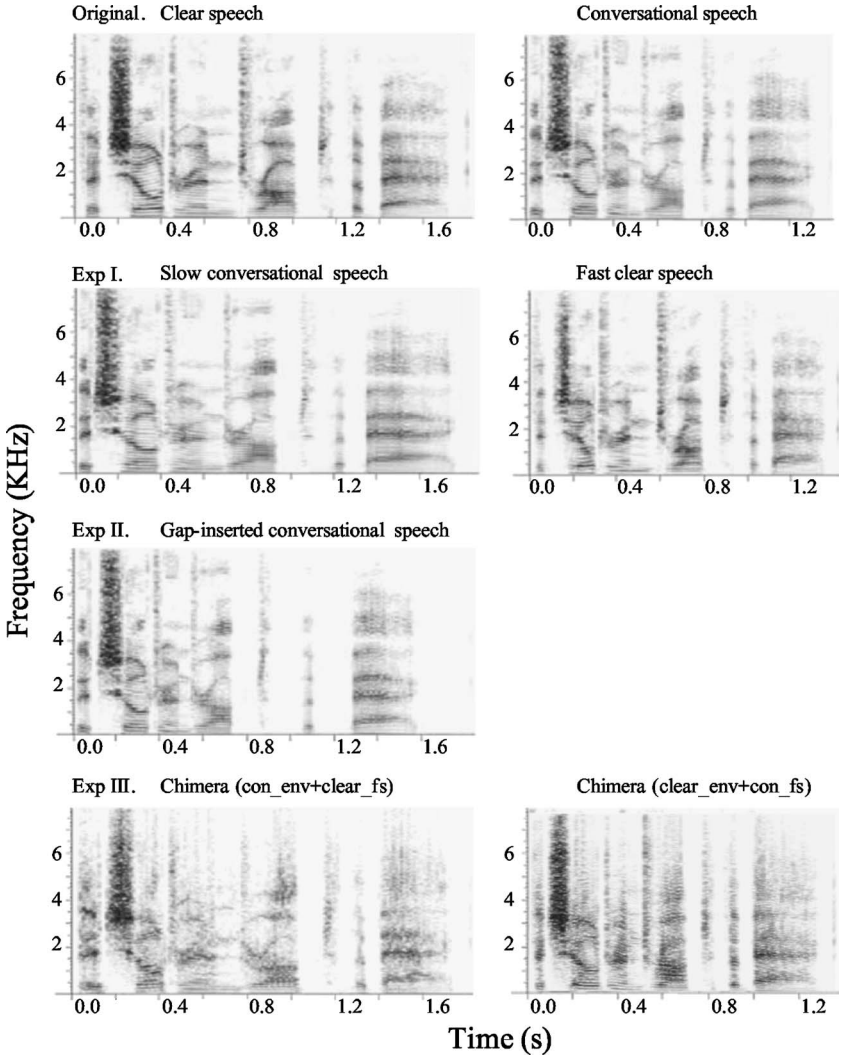

FIG. 1. Spectrograms for the sentence "The children dropped the bag" in seven stimulus conditions: clear speech (upper left), conversational speech (upper right), uniformly stretched conversational speech (second row, left), uniformly compressed clear speech (second row, right), gap-inserted conversational speech (third row), chimera containing conversational speech envelope and clear speech fine structure (bottom left), and chimera containing clear speech envelope and conversational speech fine structure (bottom right).

First, the original clear speech produced significantly better performance than the original conversational speech [$F(1,8)=15.0, p<0.05$]. The SRT was −8.8 dB for clear speech and −6.7 dB for conversational speech. Second, the processed slow conversational speech produced marginally better performance than the processed fast clear speech [$F(1,8)=4.9, p=0.06$]. The SRT was −5.9 dB for slow conversational speech and −4.0 dB for fast clear speech. Better performance with slow conversational speech suggests that either lowering speaking rate improved speech performance or time compression produced more processing artifacts than time expansion. We shall consider the latter in Sec. V. Third, the original clear speech produced significantly better performance than fast clear speech [$F(1,4)=28.4,\ p<0.05$)], but the original conversational speech produced similar performance to slow conversational speech [$F(1,4)=2.1, p>0.05$). No significant interactions were found. This result further implicates possibly more processing artifacts with time compression than time expansion. Finally, the fact that none of the processed speech produced better performance than the original speech suggests that digital processing artifacts are still a confounding factor in these newer signal-processing algorithms.

III. EXPERIMENT II. NONUNIFORMLY STRETCHED SPEECH

A. Methods

1. Subjects

Fifteen subjects were recruited to participate in this experiment using the same human subject protocol as experiment I. A within-subjects design was implemented, in which all subjects listened to the original clear, the original conversational, and the silent-gap-inserted conversational speech.

2. Stimuli

The same BKB sentences were used in this experiment as in the previous experiment. For the silent-gap-inserted conversational speech, the speaking rate was nonuniformly decreased by proportionally increasing silent gaps between phonetic segments in the conversational speech. To avoid the possibility that the silent interval between a vowel and a voiced stop consonant was inadvertently increased (Picheny *et al.*, 1986), silent gaps shorter than 10 ms were kept intact. No phonetic segments in the original conversational speech were altered; only the duration of the silent gap between these segments was proportionally increased by a predeter-

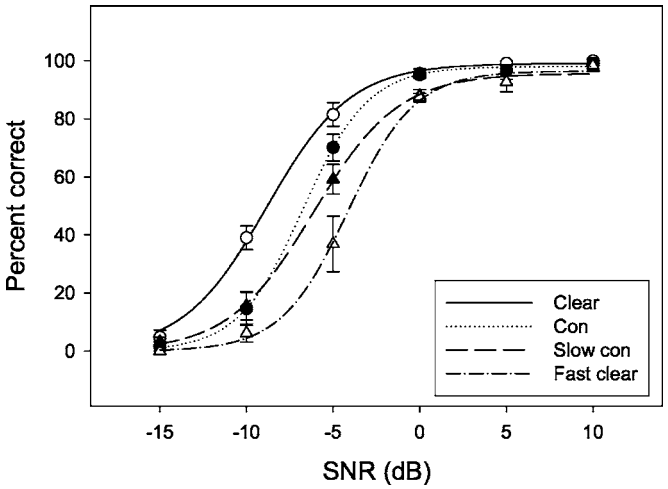

FIG. 2. Percent-correct scores as a function of signal-to-noise ratios for the original clear (open circles), original conversational (filled circles), uniformly time-scaled slow conversational (filled triangles), and fast clear speech (open triangles). Lines represent the best fitting sigmoidal psychometric functions for clear speech (solid line), conversational speech (dotted line), processed slow conversational speech (dashed line), and processed fast clear speech (dotted dashed line).

mined ratio to match the overall duration of the stretched conversational speech to that of the original clear speech. Finally, different from the 5-ms linear ramp used in the Uchanski *et al.* study, no additional ramping was used in the present study.

The left panel on the third row in Fig. 1 shows the spectrogram of the nonuniformly stretched conversational speech. Note that the gap-inserted conversational speech had the same duration as the original clear speech, but contained no apparent processing artifacts, such as smeared harmonic structure and formant transitions.

Each sentence was normalized to have the same overall root-mean-square (rms) level. Because increasing the silent intervals did not add any energy, the overall rms level in the processed speech had to be increased by an average of 1.8 dB to match the original speech overall rms level. The effect of this rms level normalization on speech intelligibility will be examined in Sec. V.

3. Procedure

The same protocol as experiment I was used in experiment II. Experiment II had a total of 15 listening conditions, including three stimulus types (original clear stimuli, original conversational stimuli, and the gap-inserted stimuli) presented at five signal-to-noise ratios (−15 to +5 dB in 5-dB steps). Each condition used eight sentences for each subject in the test. Different sentences were presented in each condition with the sentence presentation order being randomized. A within-subjects ANOVA was performed to examine the main effect of speech style and signal-to-noise ratio.

B. Results and discussion

Figure 3 shows percent-correct scores as a function of signal-to-noise ratio obtained from the original clear speech (open circles with the solid line), the original conversational speech (filled circles with the dotted line), and the gap-inserted conversational speech (filled triangles with the dashed line). A within-subjects ANOVA shows a significant main effect for both speech style $[F(2,28)=12.6, p<0.05]$ and signal-to-noise ratio $[F(4,56)=795.8, p<0.05]$. The interaction between speech style and signal-to-noise ratio was significant $[F(8,112)=2.4, p<0.05]$. The percent-correct scores at −5 dB SNR were 81.0%, 71.6%, and 62.0% for the original clear, gap-inserted conversational, and original conversational speech, respectively. The corresponding SRT values were −8.7, −7.5, and −6.2 dB (Table I). The result from experiment II appears to suggest that speaking rate accounts for roughly 50% of the clear speech advantage. We shall return to this point in Sec. V.

IV. EXPERIMENT III. CHIMERIC SPEECH

A. Methods

1. Subjects

Forty subjects were recruited to participate in experiment III using the same human subject protocol as in experiments I and II. The subjects were equally divided into four groups with each group being tested with the original clear speech, the original conversational speech, the clear speech

TABLE I. Comparison of parameters derived from the psychometric function in experiments I, II, and III. The asymptotic performance level "*S*," intercept "*a*," slope, speech-reception-threshold (SRT), and dynamic range (dB) were defined by Eqs. (1), (2), (3), and (4) in Liu *et al.* (2004).

Expt.	Stimuli	*S* (%)	*a* (dB)	Slope (%/dB)	SRT (dB)	DR (dB)
I	Clear	99.1	−8.8	10.2	−8.8	10.5
	Con	98.1	−6.7	13.0	−6.7	8.3
	Fast-clear	96.5	−4.1	12.8	−4.0	8.5
	Slow-con	95.7	−6.1	9.8	−5.9	10.7
II	Clear	100.0	−8.7	10.6	−8.7	10.3
	Con	100.0	−6.2	10.2	−6.2	10.7
	Gap-con	98.4	−7.6	9.6	−7.5	11.4
III	Clear	98.2	−8.7	13.0	−8.7	8.3
	Con	100.0	−5.3	9.4	−5.3	11.7
	Clear_env+con_fs	96.1	−4.4	15.0	−4.3	7.1
	Con_env+clear_fs	100.0	−5.0	8.2	−5.0	13.3

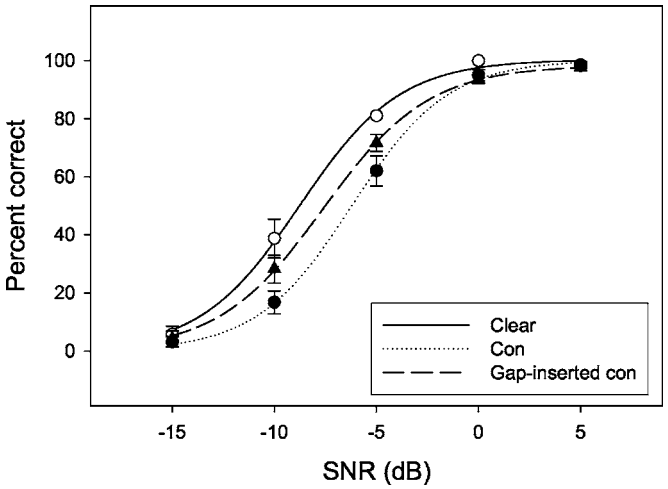

FIG. 3. Percent-correct scores as a function of signal-to-noise ratios for the original clear (open circles), gap-inserted conversational (filled triangles), and original conversational (filled circles) speech. Lines represent best-fit sigmoidal psychometric functions for clear speech (solid line), conversational speech (dotted line), and gap-inserted conversational speech (dashed line).

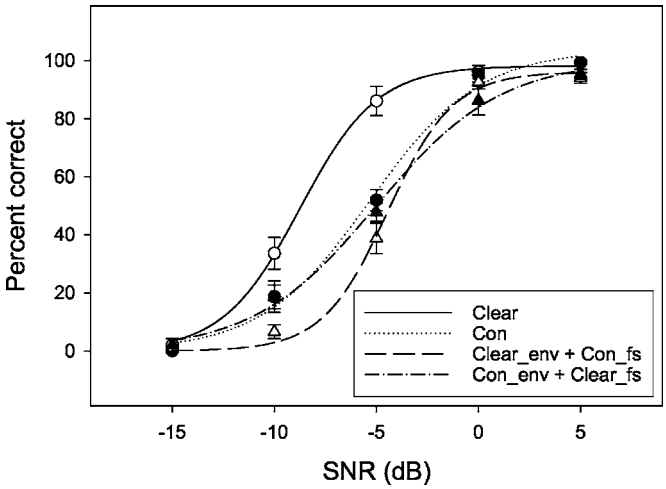

FIG. 4. Percen-correct scores as a function of signal-to-noise ratios for the original clear (open circles), original conversational (filled circles), chimera of clear speech envelope and conversational speech fine structure (open triangles), and chimera of clear speech fine structure and conversational speech envelope (filled triangles). Lines represent the best-fit sigmoidal psychometric function for clear speech (solid line), conversational speech (dotted line), chimera of clear speech envelope and conversational speech fine structure (dashed line), and chimera of clear speech fine structure and conversational speech envelope (dotted dashed line).

envelope and conversational speech fine-structure chimera, and the conversational speech envelope and clear speech fine-structure chimera, respectively.

2. Stimuli

The same BKB sentences were used in this experiment, which chimerized clear and conversational speech to create two types of new stimuli that contained either the clear speech envelope and conversational speech fine structure, or the conversational speech envelope and clear speech fine structure (Smith *et al.*, 2002). To create these stimuli, both the clear and conversational speech stimuli were spectrally divided into 16 logarithmically spaced filters spanning a frequency range of 80 to 8000 Hz (Greenwood, 1990). The number of bandlimited filters was chosen to avoid cochlear filtering with a low number of filters and filter ringing with a high number of filters (Zeng *et al.*, 2004). The bandpassed signal was then decomposed into its envelope and fine structure via the Hilbert transform. The bandlimited conversational speech envelope was nonuniformly stretched to align each segment in the original conversational speech to that in the original clear speech. The nonuniformly stretched conversational envelope was then used to amplitude modulate the clear speech fine structure. Similarly, nonuniform compression was used to match the clear speech envelope to the original conversational fine structure. Finally, the chimerized bandlimited signals were summed to form the chimerized speech.

The bottom-left panel in Fig. 1 shows the conversational speech envelope and clear speech fine structure chimera ("con_env+clear_fs"), and the bottom-right panel shows the clear speech envelope and conversational speech fine structure chimera ("clear_env+con_fs"). Because the temporal envelope was adjusted to match the duration between clear and conversational speech, the temporal fine structure determines both the overall sentence duration and individual phoneme duration in the chimera. For example, the "con_env +clear_fs" chimera (bottom left) has the same relative and overall durations as the original clear speech. Note also the slight spectral smearing in the chimerized speech. If, at a given SNR, the chimera containing the clear speech envelope produces the highest intelligibility, we would conclude that the envelope characteristics of clear speech underlie the superior intelligibility of clear speech. If, in contrast, the chimera containing the clear speech fine structure cues is more intelligible, we would reach a different conclusion that the fine structure characteristics of clear speech are responsible for its superior intelligibility.

3. Procedure

The experimental protocol used in experiments I and II was also used in experiment III. Experiment III had a total of 24 conditions, including 2 original speech stimuli and two chimeras (4 types of stimuli) $\times 5$ signal-to-noise ratios from -15 to 10 dB in 5-dB steps and in quiet (6 signal-to-noise ratios). Each condition used eight sentences for each subject in the test. A mixed ANOVA design was performed with stimulus type being the between-subjects factor and signal-to-noise ratio being the within-subjects factor.

B. Results and discussion

Figure 4 shows percent-correct scores as a function of signal-to-noise ratio for the original clear speech (open circles with the solid line), the original conversational speech (filled circles with the dotted line), the clear speech envelope and conversational speech fine structure chimera ("clear_env+con_fs," open triangles with the dashed line), and the conversational speech envelope and clear speech fine structure chimera ("con_env+clear_fs," filled triangles with the dot-dashed line). Table I shows fitted and derived parameters from the psychometric functions (experiment III). Both

stimulus type [$F(3,36)=17.8, p<0.05$] and signal-to-noise ratio [$F(5,180)=916.3, p<0.05$] were significant factors. The interaction between these two factors was also significant [$F(15,102)=8.9, p<0.05$]. Several observations can be made from the obtained data.

First, the original clear speech produced significantly better performance than all the other three stimuli, including the original conversational speech (a *posthoc* Bonferroni test, $p<0.05$). The SRT value was −8.7 dB for the original clear speech, as opposed to −5.3, −4.3, and −5.0 dB for the original conversational speech, the clear_env+con_fs, and the con_env+clear_fs chimera, respectively. The generally poorer performance with "auditory chimera" suggests the presence of processing artifacts.

Second, the significant interaction between stimulus type and signal-to-noise ratio occurred between the chimerized speech stimuli. At low SNRs (−10 and −5 dB), the con_env+clear_fs chimera produced an approximately 10-percentage-point better performance than the clear_env +con_fs chimera, implying a significant role of the fine structure cue in clear speech. At high SNRs (e.g., 0 dB), the reverse was true with the clear_env+con_fs chimera producing 6-percentage-point better performance than the con_env+clear_fs chimera, implying a significant role of the envelope cue in clear speech. Although confounded by processing artifacts, results from experiment III supported the hypothesis that the temporal envelope is critical for speech recognition in quiet and the temporal fine structure is critical for speech recognition in noise.

V. GENERAL DISCUSSION

A. Summary and comparison

Table I summarizes three fitting parameters and two derived parameters for the perceptual data from experiments I, II, and III [see Eqs. (1)–(4) in Liu *et al.*, 2004]. First, the original clear speech always produced the same or higher asymptotic performance (S), lower speech reception thresholds (SRT), and a steeper slope than the original conversational speech. The only exception was for experiment I, in which the conversational speech produced a steeper slope. Second, the SRT value was essentially identical for the original clear speech (−8.8, −8.7, and −8.7 dB) but varied greatly for the conversational speech (−6.7, −6.2, and −5.3 dB) in the present three experiments, which used the same materials and the same procedure but different subjects. For comparison, these SRT values were closely matched to the −8.5- and −6.3-dB SRT values found in the Liu *et al.* (2004) study, which used the same materials and the same procedure and an additional independent group of normal-hearing subjects. These SRT values suggest that acoustic cues in clear speech are less susceptible to individual variability than conversational speech.

We can also use the intelligibility difference in percentage points, which is equal to the product of the slope and the SRT difference between the conditions to quantify the clear speech and signal-processing effects. Except for the gap-inserted conversational speech producing 13-percentage-point higher intelligibility than the original conversation speech, all processed speech stimuli produced lower intelligibility than the original speech. The original clear speech produced intelligibility 29 percentage points higher than the uniformly stretched conversational speech, while the original conversational speech produced intelligibility 35 percentage points higher than the uniformly compressed clear speech. Similarly, Picheny *et al.* (1989) found a 30-percentage-point difference between the original and compressed clear speech and a 13-percentage-point difference between the original and the expanded conversational speech. These results suggest the presence of digital signal-processing artifacts as a confounding factor in the evaluation of the role of speaking rate in clear speech perception.

B. Signal-processing artifacts

To identify the source of processing artifacts, reversibility was tested in experiment I (Picheny *et al.*, 1989). We used the same COOL EDIT program to first compress the clear speech and then stretch the processed stimulus back to its original duration. The recovered clear speech had apparent audible processing artifacts and significantly lower intelligibility than the original clear speech. On the other hand, the recovered conversational speech, which underwent the expansion process first and the compression process second, had no audible artifacts and essentially the same intelligibility as the original conversational speech. The reversibility test revealed that the processing algorithm introduced more processing artifacts during compression than during expansion, and additionally that compression followed by expansion is irreversible while expansion followed by compression is reversible. Close examination of the "fast clear speech" spectrogram (right panel on the second row in Fig. 1) already shows a less accurate representation of formant transitions compared with the original clear speech. A processing artifact in time compression is a result of deleting segments that introduce discontinuities in fast changes, such as frequency transitions. Therefore, the compressed clear speech produced worse performance than the original clear speech, while the stretched conversational speech produced similar performance to the original conversational speech.

The chimerized speech may have introduced different types of processing artifacts than the uniformly time-scaled speech. The chimera method first extracted the bandlimited temporal envelope and fine structure from two sentences of different durations. The envelope had to be compressed (in clear speech) or stretched (in conversational speech) to match the duration of the fine structure, which remained intact. There were at least three sources of processing artifacts. The first artifact stemmed from alterations in modulation frequencies, which were introduced by digital resampling in temporal envelopes. The second artifact was introduced by the segment mismatch between one sentence's temporal envelope and another sentence's fine structure. The third artifact was due to the bandpass filtering in the analysis-synthesis process, which was generally irreversible. Clearly, these processing artifacts degraded performance and confounded the interpretation of the present results.

C. Speaking rate

While both uniform time scaling and chimerizing introduced processing artifacts, inserting silent gaps to decrease the conversational speech rate did not introduce any artifacts. At first glance, the results from experiment II seem to indicate that longer pauses between speech segments improved the perception of conversational speech by 1.3 dB in terms of the SRT measure. However, one may question whether this improvement is truly a result of the decreased speaking rate in the processed conversational speech.

Recall from Sec. II A 2 in experiment I that the average overall duration was 1.31 s for conversational speech and 1.97 s for clear speech, indicating that, on average, 0.66 s of silent gaps had to be inserted in the conversational speech to match the duration of the clear speech. As described in Sec. II A, a normalization procedure was employed to equalize the overall rms for all processed and original sentences. This normalization procedure increased the overall rms level by 1.8 dB for the gap-inserted conversational speech. Because the inserted silent gaps did not contribute to the overall rms level, the short-term rms level had to be increased proportionally by 1.8 dB for all phonetic segments. If we assume that the listener used a short-term window (tens to hundreds of milliseconds), instead of a 1- or 2-s window, to calculate the sentence-level rms level, then the effective short-term signal-to-noise ratio would be 1.8 dB higher than suggested by the overall rms level. Therefore, it is possible that the observed 1.3-dB improvement in SRT was a result of the rms level normalization employed at the sentence level. If this short-term rms level hypothesis holds true, then inserting silent gaps in conversational speech would not necessarily improve intelligibility.

Because longer silent gaps or pauses were consistently observed in clear speech, the above examination on the role of the short-term rms level brings about an important question: to what extent is the so-called clear speech advantage a result of the increased short-term signal-to-noise ratio? To answer this question, we removed all pauses in the original clear and original conversational speech and calculated their rms levels. For male talker materials used in the present study, we found that the pause-removed clear speech had a 0.2 dB higher overall rms level than the pause-removed conversational speech. Clearly, this 0.2-dB difference cannot account for the observed 3-dB clear speech advantage, suggesting that acoustic cues other than speaking rate contribute significantly to the clear speech advantage.

D. Temporal envelope and fine structure

Previous studies have emphasized the importance of the temporal envelope in speech recognition (Van Tasell *et al.*, 1987; Rosen, 1992; Drullman *et al.*, 1994; Shannon *et al.*, 1995), but recent results have suggested a complementary role of the temporal fine structure in speech recognition in noise (Nie *et al.*, 2005; Zeng *et al.*, 2005b). Although auditory chimera introduced digital processing artifacts (Smith *et al.*, 2002; Zeng *et al.*, 2004), results from experiment III suggest that this idea can be extended and applied to clear speech perception. At high signal-to-noise ratios, the chimera with the clear speech envelope and the conversational speech fine structure produced higher intelligibility than the chimera with conversational speech envelope and clear speech fine structure. On the other hand, the reverse was true at low signal-to-noise ratios. As far as clear speech perception is concerned, the present result suggests that the temporal envelope and fine structure contribute complementarily to the clear speech advantage. The temporal envelope contributes to the clear speech advantage in quiet, while the temporal fine structure contributes to the clear speech advantage in noise.

VI. CONCLUSIONS

The present study used three methods to evaluate temporal properties in clear speech perception. The three methods included (1) uniform time scaling to increase the clear speech rate or decrease the conversational speech rate; (2) nonuniform time scaling to decrease the conversational speech rate by increasing pauses between phonetic segments in conversational speech; and (3) "auditory chimera" with clear speech temporal envelope and conversational speech fine-structure or vice versa (Smith *et al.*, 2002). Based on acoustic analysis and perceptual data, we reached the following conclusions:

(1) Consistent with previous studies, the present study found a consistent clear speech advantage corresponding to a 2–3-dB difference in the speech reception threshold between clear and conversational speech.
(2) While both uniform time compression and stretching introduced processing artifacts, time compression was found to be more detrimental than time stretching in terms of processing reversibility and the degree of performance degradation.
(3) Increasing silent gaps in conversational speech decreased the speaking rate without introducing any processing artifacts. Perceptual results showed a 1.3-dB advantage in SRT for the gap-inserted conversational speech, accounting for roughly half of the overall clear speech advantage. Acoustic analysis indicated that this improvement in SRT might be a result of an increased short-term signal-to-noise ratio due to the rms level normalization at the sentence level.
(4) Although auditory chimera introduced digital processing artifacts, perceptual results from the chimerized clear and conversational speech suggested a complementary role of temporal envelope and fine structure in speech perception: the temporal envelope contributes more to the clear speech advantage at high signal-to-noise ratios, while the temporal fine structure contributes more at low signal-to-noise ratios.

ACKNOWLEDGMENTS

We thank Tiffany Chua, Elsa Del Rio, Paul Meneses, and Frank Z. Yu for stimulus processing and data collection. We thank Ann R. Bradlow for providing speech materials. Abby Copeland, Alex Francis, and two anonymous reviewers provided helpful comments on an earlier draft of this paper. This work was funded by the National Institutes of Health,

Department of Health and Human Services (2 RO1-DC02267).

Assmann, P. F., and Katz, W. F. (**2005**). "Synthesis fidelity and time-varying spectral change in vowels," J. Acoust. Soc. Am. **117**, 886–895.

Beasley, D. S., Schwimmer, S., and Rintelmann, W. F. (**1972**). "Intelligibility of time-compressed CNC monosyllables," J. Speech Hear. Res. **15**, 340–350.

Bench, J., and Bamford, J. (**1979**). *Speech-hearing Tests and the Spoken Language of Hearing-impaired Children* (Academic, London).

Bradlow, A. R., and Bent, T. (**2002**). "The clear speech effect for non-native listeners," J. Acoust. Soc. Am. **112**, 272–284.

Bradlow, A. R., Kraus, N., and Hayes, E. (**2003**). "Speaking clearly for children with learning disabilities: sentence perception in noise," J. Speech Lang. Hear. Res. **46**, 80–97.

Chen, F. (**1980**). *Acoustic Characteristics of Clear and Conversational Speech at Segmental Level* (Massachusetts Institute of Technology, Cambridge, MA).

Dorman, M. F., and Loizou, P. C. (**1996**). "Relative spectral change and formant transitions as cues to labial and alveolar place of articulation," J. Acoust. Soc. Am. **100**, 3825–3830.

Drullman, R. (**1995**). "Temporal envelope and fine structure cues for speech intelligibility," J. Acoust. Soc. Am. **97**, 585–592.

Drullman, R., Festen, J. M., and Plomp, R. (**1994**). "Effect of reducing slow temporal modulations on speech reception," J. Acoust. Soc. Am. **95**, 2670–2680.

Dudley, H. (**1939**). "The vocoder," Bell Lab. Rec. **17**, 122–126.

Fairbanks, G., Everitt, W. L., and Jerger, R. P. (**1954**). "Method for time or frequency compression-expansion of speech," IRE Trans. Audio **2**, 7–12.

Fairbanks, G., Guttman, N., and Miron, M. S. (**1957**). "Effects of time compression upon the comprehension of connected speech," J. Speech Hear Disord. **22**, 10–19.

Ferguson, S. H., and Kewley-Port, D. (**2002**). "Vowel intelligibility in clear and conversational speech for normal-hearing and hearing-impaired listeners," J. Acoust. Soc. Am. **112**, 259–271.

Fu, Q. J. (**2002**). "Temporal processing and speech recognition in cochlear implant users," NeuroReport **13**, 1635–1639.

Gagne, J., Rochette, A., and Charest, M. (**2002**). "Auditory, visual, and audiovisual clear speech," Speech Commun. **37**, 213–230.

Gagne, J., Querengesser, C., Folkeard, P., Munhall, K., and Mastern, V. (**1995**). "Auditory, visual and audiovisual speech intelligibility for sentence-length stimuli: An investigation of conversational and clear speech," The Volta Review **97**, 33–51.

Gordon-Salant, S., and Fitzgibbons, P. J. (**1995**). "Recognition of multiply degraded speech by young and elderly listeners," J. Speech Hear. Res. **38**, 1150–1156.

Gordon-Salant, S., and Fitzgibbons, P. J. (**1997**). "Selected cognitive factors and speech recognition performance among young and elderly listeners," J. Speech Lang. Hear. Res. **40**, 423–431.

Greenwood, D. D. (**1990**). "A cochlear frequency-position function for several species—29 years later," J. Acoust. Soc. Am. **87**, 2592–2605.

Helfer, K. S. (**1997**). "Auditory and auditory-visual perception of clear and conversational speech," J. Speech Lang. Hear. Res. **40**, 432–443.

Houtgast, T., and Steeneken, H. J. (**1985**). "A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria," J. Acoust. Soc. Am. **77**, 1071–1077.

Kawahara, H., Masuda-Katsuse, I., and de Cheveigne, A. (**1999**). "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based *F*0 extraction: Possible role of a repetitive structure in sounds," Crystallogr. Rep. **27**, 187–207.

Kong, Y. Y., Stickney, G. S., and Zeng, F. G. (**2005**). "Speech and melody recognition in binaurally combined acoustic and electric hearing," J. Acoust. Soc. Am. **117**, 1351–1361.

Krause, J. C., and Braida, L. D. (**2002**). "Investigating alternative forms of clear speech: the effects of speaking rate and speaking mode on intelligibility," J. Acoust. Soc. Am. **112**, 2165–2172.

Krause, J. C., and Braida, L. D. (**2004**). "Acoustic properties of naturally produced clear speech at normal speaking rates," J. Acoust. Soc. Am. **115**, 362–378.

Kurdziel, S., Noffsinger, D., and Olsen, W. (**1976**). "Performance by cortical lesion patients on 40 and 60% time-compressed materials," J. Am. Aud Soc. **2**, 3–7.

Liu, C., and Kewley-Port, D. (**2004**). "Vowel formant discrimination for high-fidelity speech," J. Acoust. Soc. Am. **116**, 1224–1233.

Liu, S., Del Rio, E., Bradlow, A. R., and Zeng, F. G. (**2004**). "Clear speech perception in acoustic and electric hearing," J. Acoust. Soc. Am. **116**, 2374–2383.

Malah, D. (**1979**). "Time-domain algorithms for harmonic bandwidth reduction and time scaling of speech signals," IEEE Trans. Acoust., Speech, Signal Process. **27**, 121–133.

Moulines, E., and Laroche, J. (**1995**). "Non-parametric techniques for pitch-scale and time-scale modification of speech," Crystallogr. Rep. **16**, 175–205.

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

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

Payton, K. L., Uchanski, R. M., and Braida, L. D. (**1994**). "Intelligibility of conversational and clear speech in noise and reverberation for listeners with normal and impaired hearing," J. Acoust. Soc. Am. **95**, 1581–1592.

Picheny, M. A., Durlach, N. I., and Braida, L. D. (**1985**). "Speaking clearly for the hard of hearing. I. Intelligibility differences between clear and conversational speech," J. Speech Hear. Res. **28**, 96–103.

Picheny, M. A., Durlach, N. I., and Braida, L. D. (**1986**). "Speaking clearly for the hard of hearing. II. Acoustic characteristics of clear and conversational speech," J. Speech Hear. Res. **29**, 434–446.

Picheny, M. A., Durlach, N. I., and Braida, L. D. (**1989**). "Speaking clearly for the hard of hearing. III. An attempt to determine the contribution of speaking rate to differences in intelligibility between clear and conversational speech," J. Speech Hear. Res. **32**, 600–603.

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.

Rosen, S. (**1992**). "Temporal information in speech: Acoustic, auditory and linguistic aspects," Philos. Trans. R. Soc. London, Ser. B **336**, 367–373.

Schum, D. J. (**1996**). "Intelligibility of clear and conversational speech of young and elderly talkers," J. Am. Acad. Audiol **7**, 212–218.

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

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

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

Uchanski, R. M., Choi, S. S., Braida, L. D., Reed, C. M., and Durlach, N. I. (**1996**). "Speaking clearly for the hard of hearing IV: Further studies of the role of speaking rate," J. Speech Hear. Res. **39**, 494–509.

Van Tasell, D. J., Soli, S. D., Kirby, V. M., and Widin, G. P. (**1987**). "Speech waveform envelope cues for consonant recognition," J. Acoust. Soc. Am. **82**, 1152–1161.

Walley, A. C., and Carrell, T. D. (**1983**). "Onset spectra and formant transitions in the adult's and child's perception of place of articulation in stop consonants," J. Acoust. Soc. Am. **73**, 1011–1022.

Zeng, F. G., and Galvin, J. J. (**1999**). "Amplitude mapping and phoneme recognition in cochlear implant listeners," Ear Hear. **20**, 60–74.

Zeng, F. G., Kong, Y. Y., Michalewski, H. J., and Starr, A. (**2005a**). "Perceptual consequences of disrupted auditory nerve activity," J. Neurophysiol. **93**, 3050–3063.

Zeng, F. G., Oba, S., Garde, S., Sininger, Y., and Starr, A. (**1999**). "Temporal and speech processing deficits in auditory neuropathy," NeuroReport **10**, 3429–3435.

Zeng, F. G., Nie, K., 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., Stickney, G. S., Kong, Y. Y., Vongphoe, M., Bhargave, A., Wei, C., and Cao, K. (**2005b**). "Speech recognition with amplitude and frequency modulations," Proc. Natl. Acad. Sci. U.S.A. **102**, 2293–2298.