Effect of temporal modulation rate on the intelligibility of phase-based speech

Fei Chen[a)]
Division of Speech and Hearing Sciences, The University of Hong Kong, Prince Philip Dental Hospital, 34 Hospital Road, Hong Kong
feichen1@hku.hk

Tian Guan[b)]
Research Centre of Biomedical Engineering, Graduate School at Shenzhen, Tsinghua University, Shenzhen 518055, China
guantian@sz.tsinghua.edu.cn

Abstract: This study investigated the effect of temporal modulation rate on the intelligibility of speech synthesized with primarily phase information using two methods: Phase-based vocoded speech (preserving phase cues and discarding envelope cues) and Hilbert fine-structure stimuli (summing up the multi-channel Hilbert fine-structure waveforms). Listening experiments with normal-hearing participants showed that the intelligibility of the two types of phase-based speech was significantly improved when synthesized using a high temporal modulation rate (or short frame) compared to that synthesized using the whole speech segment. This intelligibility advantage appears to be attributed to better preservation of the temporal envelope cues in phase-based speech.

PACS numbers: 43.71.Gv, 43.71.Es [SGS]
Date Received: July 18, 2013 **Date Accepted:** October 14, 2013

1. Introduction

Amplitude and phase are two properties carrying important information for speech perception. The relative importance of amplitude to speech recognition has been extensively investigated through, for instance, envelope-based vocoder simulation studies, which preserve envelope cues and eliminate temporal fine-structure (FS) (or phase) cues by replacing them with a sinusoidal or band-limited noise carrier (e.g., Shannon *et al.*, 1995; Dorman *et al.*, 1997). A relatively large number of channels and a high low-pass (LP) cut-off frequency (or temporal modulation rate) to extract the envelope signal were found to favor better recognition of the envelope-based vocoded speech (e.g., Shannon *et al.*, 1995; Xu *et al.*, 2005).

Recently many studies used the Hilbert-transform-derived FS (HFS) stimuli to study the impact of phase information on speech intelligibility (e.g., Smith *et al.*, 2002; Zeng *et al.*, 2004; Lorenzi *et al.*, 2006; Gilbert and Lorenzi, 2006; Moore, 2008). The Hilbert transform decomposes a band-passed signal into its envelope and FS (or frequency modulation, which is the first derivative of the phase information preserved in the fine-structure waveform) components (Smith *et al.*, 2002). While the envelope captures the slowly varying modulations of amplitude in time, the FS component captures the rapid oscillations occurring at a rate close to the center frequency of the band. Studies showed that listeners could understand, with high accuracy, speech synthesized to contain only HFS information (e.g., Smith *et al.*, 2002; Gilbert and Lorenzi, 2006). In addition, studies attempted to investigate how the intelligibility of phase-based speech (e.g., the HFS stimuli) was affected by the properties used in speech synthesis, e.g.,

[a)]Author to whom correspondence should be addressed.
[b)]Also at Department of Biomedical Engineering, Medical School, Tsinghua University, Beijing 100084, China.

number of channels, bandwidth of analysis filters, spectral resolution, and temporal resolution (e.g., Smith *et al.*, 2002; Gilbert and Lorenzi, 2006; Kazama *et al.*, 2010). Smith *et al.* (2002) found that the number of channels in speech synthesis affected the intelligibility of the HFS stimuli, i.e., the larger the number of channels, the less intelligible the HFS stimuli. Gilbert and Lorenzi (2006) suggested that the HFS stimuli contained envelope cues that could be recovered by auditory filtering, and those recovered envelope cues were affected by the bandwidth of analysis filters to synthesize the HFS stimuli. Kazama *et al.* (2010) assessed the roles of spectral resolution and temporal resolution on the significance of phase information in the short-time Fourier transform (STFT) spectrum for speech intelligibility. Their speech intelligibility data showed the significance of phase spectrum for long (>256 ms) and for very short (<4 ms) windows.

Despite the number of studies examining the effect of the number of channels on the intelligibility of phase-based speech (e.g., the HFS stimuli), most of those studies used the whole speech segment to synthesize the phase-based speech [e.g., the HFS stimuli in Smith *et al.* (2002) and Gilbert and Lorenzi (2006)]. The limitation of the STFT-based speech synthesis in Kazama *et al.* (2010) is that the effect of temporal resolution (or temporal modulation rate) is entangled with that of spectral resolution when synthesizing the phase-based speech with the inverse STFT. That is, the better the temporal resolution is, the worse the spectral resolution is. Hence our understanding of the influence of temporal modulation rate (or frame duration in speech synthesis) to the intelligibility of phase-based speech is still limited. This motivates the present study to evaluate the effect of temporal modulation rate on the role of phase information to sentence intelligibility via two listening experiments. Because vocoder simulation is widely used to assess the roles of speech properties on speech intelligibility, experiment 1 uses a phase-based vocoder simulation to study the significance of temporal modulation rate of phase information to sentence intelligibility. Different from the traditional envelope-based vocoder (e.g., Shannon *et al.*, 1995; Dorman *et al.*, 1997), the phase-based vocoder in experiment 1 preserves temporal phase cues and eliminates envelope cues (see more in Sec. 2.2). Note that many existing phase-based vocoder techniques are implemented based on the STFT spectrum (e.g., Dolson, 1986). As this study aims to examine how temporal modulation rate will affect the intelligibility of phase-based speech with fixed spectral resolution, a non-STFT based implementation of a phase-based vocoder is used in experiment 1. Experiment 2 investigates the effect of temporal modulation rate on the intelligibility of the HFS stimuli.

Previous studies suggested the importance of preserving the narrow-band envelope for the intelligibility of phase-based HFS speech (e.g., Gilbert and Lorenzi, 2006). They found that the temporal envelope cues recovered by auditory filtering or the correlation between the narrow-band envelopes of the original speech and the synthesized phase-based signal predicted well the effects of the properties used in speech synthesis, e.g., bandwidth of analysis filters (Gilbert and Lorenzi, 2006) and segment length (Kazama *et al.*, 2010). Motivated by this, we hypothesize that the effect of temporal modulation rate on the intelligibility of phase-based speech may be attributed to the amount of envelope cues preserved in the phase-based speech. To verify this hypothesis, the present work will use an envelope-based objective intelligibility metric [i.e., the normalized covariance metric (NCM) (Chen and Loizou, 2011)] to assess the degree to which temporal envelope cues could be recovered from phase-based speech and to model the intelligibility of phase-based speech. A large NCM value indicates a better preservation of the envelope cues in phase-based speech relative to the original unprocessed speech, and predicts a high intelligibility score in listening experiments (Chen and Loizou, 2011).

In short, the aim of the present work is twofold: (1) To investigate the effect of temporal modulation rate on the intelligibility of two types of phase-based speech (i.e., phase-based vocoded speech and HFS stimuli) and (2) to examine the hypothesis that the intelligibility advantage (if any) of phase-based speech synthesized with a high temporal modulation rate could be attributed to better preservation of the temporal envelope cues in phase-based speech.

2. Experiment 1: Effect of temporal modulation rate on the intelligibility of phase-based vocoded speech

2.1 Subjects and materials

Eight normal-hearing (NH) (i.e., pure tone thresholds better than 20 dB hearing level at octave frequencies from 125 to 8000 Hz in both ears) listeners participated in this experiment. All subjects were native-speakers of Mandarin Chinese and were paid for their participation. The speech material consisted of sentences taken from the Mandarin Hearing in Noise Test (MHINT) database (Wong *et al.*, 2007). There were a total of 24 lists in the MHINT corpus. Each MHINT list had 10 sentences, and each sentence contained 10 keywords. All the sentences were spoken by a male native-Mandarin speaker (with fundamental frequency ranging from 75 to 180 Hz) and recorded at a sampling rate of 16 kHz.

2.2 Signal processing

To synthesize the phase-based vocoded speech, signals were first processed through a pre-emphasis (high-pass) filter (2000 Hz cut-off) with a 3 dB/octave roll-off and then band-passed into N ($N = 1$, 2, 4, 8, 16, 32, or 64 in this study) frequency bands between 80 and 6000 Hz using sixth-order Butterworth analysis filters. The cut-off frequencies of the N band-pass analysis filters were computed according to the cochlear frequency-position mapping function (Greenwood, 1990). Sinusoids were generated with amplitudes equal to one, frequencies equal to the center frequencies of the band-pass analysis filters, and phases estimated from the fast Fourier transform of every T ms ($T = 1$, 2, or 4 ms in this study) of non-overlapping frames (McAulay and Quatieri, 1995). The sinusoids of each band were finally summed up, and the level of the synthesized speech was adjusted to have the same root-mean-square (RMS) level as the original speech.

2.3 Procedure

The experiment was performed in a sound-proof room, and stimuli were played to listeners monaurally through a Sennheiser HD 250 Linear II circumaural headphone at a comfortable listening level. Before the test, each subject participated in a 5-min training session to listen to a set of phase-based vocoded speech materials and to familiarize themselves with the testing procedure. During the test, subjects were asked to repeat the sentences they heard, and each keyword in the sentences was scored as correct or incorrect. Each subject participated in a total of 18 testing conditions [six numbers of channels (i.e., $N = 1$, 2, 4, 8, 16, and 32) $\times$ three frame durations (i.e., $T = 1$, 2, and 4 ms)]. One list of MHINT sentences (i.e., 10 sentences) was used per condition, and none of the lists were repeated across the conditions. The order of the testing conditions was randomized across subjects. Subjects were given a 5-min break every 30 min during the test. The percentage intelligibility score was calculated by dividing the number of keywords correctly identified by the total number of keywords in a testing condition. As each testing condition contained 100 keywords, the percentage intelligibility score also equaled the number of correctly recognized keywords.

2.4 The envelope-based speech intelligibility metric

In computing the NCM metric, signals were first band-pass filtered into a number of bands, and then the temporal envelope of each band was extracted by using the Hilbert transform. The normalized covariance between the envelopes of the original unprocessed signal and the phase-based signal was computed, mapped to an apparent signal-to-noise ratio value, and converted to a transmission index (TI) value for each frequency band. The weighted average of the TI values for all bands was finally computed to derive the NCM measure. This study split the signals into 16 analysis bands spanning the signal bandwidth and used the ANSI weights (ANSI, 1997) for TI values in computing the NCM metric (Chen and Loizou, 2011).

2.5 Results and discussion

Figure 1(a) shows the mean recognition scores of the phase-based vocoded speech as a function of the number of channels and frame duration used in speech synthesis. Statistical significance was determined by using the percentage correct score as the dependent variable, and the number of channels and frame duration as the two within-subjects factors. The scores were first converted to rational arcsine units (RAU) using the rationalized arcsine transform (Studebaker, 1985). Two-way analysis of variance (ANOVA) with repeated measures indicated a significant effect [$F(5,35) = 280.1$, $p < 0.0005$] of the number of channels, frame duration [$F(2,14) = 361.6$, $p < 0.0005$], and a significant interaction [$F(10,70) = 33.7$, $p < 0.0005$] between the number of channels and frame duration.

Results in Fig. 1(a) clearly show the favorable effects of a large number of channels and a high temporal modulation rate (or short frame) on the intelligibility of phase-based vocoded speech. When the number of channels and frame duration are set to 1 and 4 ms, respectively, in vocoder simulation, the phase-based vocoded speech carries little intelligibility information (i.e., intelligibility score = 0.0%). The intelligibility score improves to 55.8% as the number of channels increases to 32 (at 4 ms frame duration). Furthermore, when a frame duration of 1 ms is used in speech synthesis, the intelligibility score increases 97.9% (at 32 channels) in Fig. 1(a) relative to one channel. To some extent, these results are consistent with those regarding the effects of the number of channels and temporal modulation rate on the intelligibility of envelope-based vocoded speech (e.g., Shannon *et al.*, 1995; Dorman *et al.*, 1997; Xu *et al.*, 2005). In previous studies, the identification of envelope-based vocoded sentences improved markedly as the number of channels increased (Shannon *et al.*, 1995; Dorman *et al.*, 1997). The temporal modulation rate adjustment in the envelope-based vocoder simulation was implemented by altering the LP cutoff frequency to extract the envelope signals. Using a high LP cutoff frequency significantly improved the intelligibility of envelope-based vocoded speech (e.g., Xu *et al.*, 2005). The results of the present study together with those of Xu *et al.* (2005) suggest that the number of channels and the temporal modulation rate have a consistent influence contributing to the intelligibility of both envelope- and phase-based vocoded speech. Note that the present findings [i.e., the favorable effect of high temporal modulation rate (or short frame)] differ from those reported by Kazama *et al.* (2010), in which the intelligibility of the STFT- and phase-based speech improved when they were synthesized with either a long window (>256 ms) or a short window (<4 ms). This may be attributed to the use of two different mechanisms to synthesize the phase-based speech in the two studies. Kazama *et al.* (2010) reported the hybrid influence of temporal modulation rate and spectral resolution on the intelligibility of the STFT- and phase-based speech. However, with the same number of channels in the phase-based vocoder simulation, the present work predicts a negative influence on intelligibility by using a long frame to synthesize the phase-based vocoded speech.

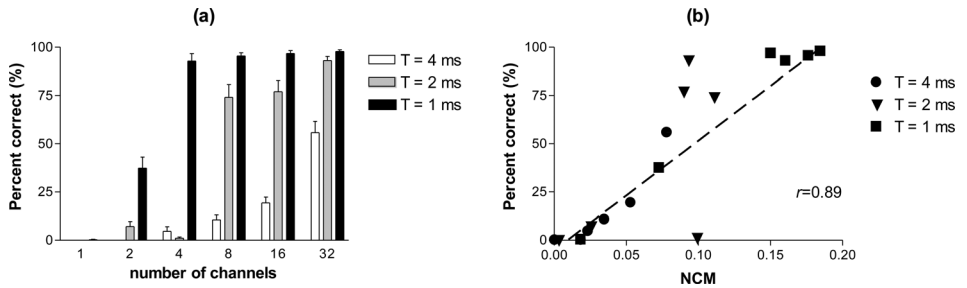

Fig. 1. (a) Mean recognition scores of the phase-based vocoded speech as a function of the number of channels and frame duration used in speech synthesis and (b) scatter plot of the intelligibility scores against the predicted NCM values. The error bars denote ±1 standard error of the mean.

Figure 1(b) shows the scatter plot of the individual subjective intelligibility scores against the predicted NCM values of the 18 testing conditions in Fig. 1(a). A linear function was used for mapping the NCM values to the intelligibility scores in Fig. 1(b), and the Pearson's correlation coefficient between the NCM values and the intelligibility scores is $r = 0.89$. A high correlation between intelligibility scores and NCM values indicates that the recovered envelope cues from phase-based speech may account for much of the variance of the intelligibility of phase-based vocoded speech. This high correlation coefficient also suggests that the increased intelligibility of phase-based vocoded speech when synthesized with a high temporal modulation rate (or short frame) may be attributed to better preservation of the temporal envelope cues in phase-based vocoded speech compared to the condition when synthesized with a low temporal modulation rate (or long frame).

3. Experiment 2: Effect of temporal modulation rate on the intelligibility of the Hilbert fine-structure stimuli

3.1 Subjects and materials

The same eight NH, native-Mandarin listeners participated in this experiment. The speech material consisted of sentences taken from the MHINT database.

3.2 Signal processing

The signal processing condition in this experiment followed that used by Lorenzi *et al.* (2006) to investigate the role of the Hilbert fine-structure to speech intelligibility. Signals were first split into N frequency bands comparable to the process of creating the phase-based vocoded speech in experiment 1. The Hilbert transform was applied to the band-passed signals to obtain the HFS waveforms. The envelope components were discarded, while the N-channel HFS components were weighted to have the same RMS value as the band-passed signal, summed up, and finally adjusted to the RMS level of the original speech signal.

Note that Smith *et al.* (2002) and Lorenzi *et al.* (2006) did not synthesize the HFS stimuli on a segment basis. In other words, the segment length in their studies equaled that of the whole speech signal. The current experiment assessed the influence of segment length used in synthesizing the HFS stimuli to the intelligibility of these materials. The signal processing in this experiment was conducted for every T ms of non-overlapping speech segment. Finally, the concatenated signal of all T ms HFS stimuli was presented to listeners for recognition. We selected three segment lengths, i.e., T = length of the original speech signal, 100 ms, and 50 ms. Previous studies found that the intelligibility of the HFS stimuli decreased when a large number of channels of analysis filters was used in speech synthesis (e.g., Smith *et al.*, 2002). Our pilot study showed that within the range of 1–16 channels, the HFS stimuli were notably intelligible. Hence to examine the effect of segment length on the intelligibility of the HFS stimuli, we used a large number of channels (i.e., $N = 32$ and 64) in this experiment.

3.3 Procedure

The experimental procedure was the same as that used in experiment 1. Each subject participated in a total of six testing conditions [= two numbers of channels (i.e., $N = 32$ and 64) × three segment lengths (i.e., T = full length, 100 ms, and 50 ms)]. One list of MHINT sentences was used per condition, and none of the lists were repeated across the conditions. The order of the testing conditions was randomized across subjects.

3.4 Results and discussion

Figure 2(a) shows the mean speech recognition scores for HFS stimuli as a function of the number of channels and segment length used in speech synthesis. It shows that when the number of channels is set to $N = 32$ or 64, the HFS stimuli (full segment) carry little intelligibility information, i.e., sentence recognition score = 0.0%. This finding is consistent with that reported in Smith *et al.* (2002). Statistical significance was determined by using the percentage correct score as the dependent variable, and the number of channels and

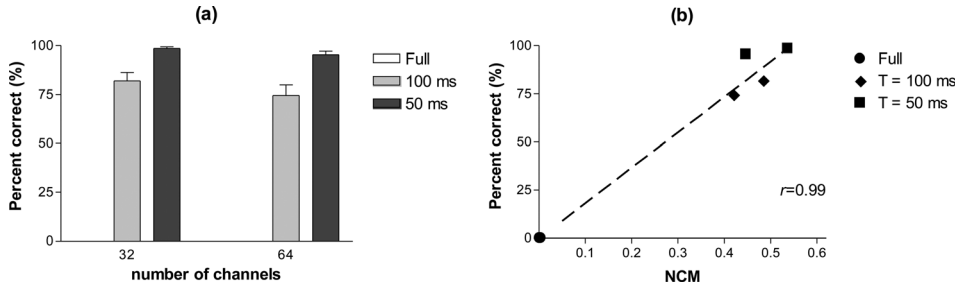

Fig. 2. (a) Mean recognition scores of the HFS stimuli as a function of the number of channels and segment length used in speech synthesis and (b) scatter plot of the intelligibility scores against the predicted NCM values. The error bars denote ±1 standard error of the mean.

segment length as the two within-subjects factors. The scores were first converted to RAU using the rationalized arcsine transform (Studebaker, 1985). Two-way ANOVA with repeated measures indicated a significant effect [$F(1,7) = 181.7$, $p < 0.0005$] of the number of channels, segment length [$F(2,14) = 314.3$, $p < 0.0005$], and a significant interaction [$F(2,14) = 9.9$, $p = 0.002$] between the number of channels and segment length. The significant interaction appears to be due to the floor effect of intelligibility scores (i.e., 0.0%) when the HFS stimuli are synthesized with the whole speech segment at either 32 or 64 channels as shown in Fig. 2(a).

The outcomes from the present experiment are consistent with those in experiment 1, i.e., both show significantly improved intelligibility of phase-based speech as a result of speech synthesis using a high temporal modulation rate (i.e., short frame or short segment). For instance, with 64 channels, the intelligibility of the HFS stimuli is 0.0%, 74.5%, and 95.5% when synthesized with segment lengths equal to that of the whole speech (i.e., condition "full"), 100 ms, and 50 ms, respectively, as shown in Fig. 2(a). The experimental results here suggest that the low intelligibility of the HFS stimuli synthesized with a large number of channels of analysis filters may be attributed to the reduced envelope cues (to be recovered by auditory filters) *and* the low temporal modulation rate in speech synthesis. For instance, the HFS stimuli in Smith *et al.* (2002) and Gilbert and Lorenzi (2006) were synthesized with the lowest temporal modulation rate, i.e., a segment length equal to that of the whole speech stimulus. When the temporal modulation rate is increased to synthesize the HFS stimuli, the intelligibility is significantly improved, as observed in Fig. 2(a). This indicates that the use of a high temporal modulation rate in synthesizing the HFS stimuli can compensate for reduced envelope cues preserved in the HFS stimuli due to the use of a large number of analysis filters (or channels) to synthesize the HFS stimuli.

Figure 2(b) shows the scatter plot of the intelligibility scores against the predicted NCM values of the six testing conditions in Fig. 2(a). A linear function was used for mapping the NCM values to the intelligibility scores in Fig. 2(b), and the Pearson's correlation coefficient between the NCM values and the intelligibility scores is $r = 0.99$, suggesting that the increased intelligibility of the HFS stimuli when synthesized with a high temporal modulation rate may be attributed to better preservation of the temporal envelope cues in the HFS stimuli. Note that, as this experiment only consists of six testing conditions, a bimodal distribution is observed in intelligibility scores. Further study is warranted to investigate the distribution of intelligibility scores when tested with more conditions.

Previous studies showed that phase-based HFS stimuli processed with one to two analysis bands and the whole speech segment were highly intelligible (e.g., Smith *et al.*, 2002); however, Fig. 1(a) shows that phase-based vocoded speech synthesized with one to two vocoded channels is unintelligible (i.e., intelligibility score 0.0%). This difference may be attributed to the different degrees of envelope cues contained in the two types of phase-based speech. Many studies suggested that, when processed with one to two analysis bands, HFS stimuli carried a large amount

of envelope cues (at the output of the auditory filters) favorable for speech perception (e.g., Zeng *et al.*, 2004; Gilbert and Lorenzi, 2006). However, phase-based vocoded speech processed with one to two vocoded channels may contain fewer envelope cues as shown by the small NCM value (i.e., close to zero) in Fig. 1(b). Further study is needed to investigate the effect of analysis filters (e.g., number of channels) on the envelope cues preserved in phase-based vocoded speech.

4. Conclusions

The present studies extend previous findings on the intelligibility of phase-based speech (e.g., Smith *et al.*, 2002; Gilbert and Lorenzi, 2006; Kazama *et al.*, 2010). The present results indicate that at a fixed number of channels of analysis filters in speech synthesis, the use of a high temporal modulation rate (or short frame) can bring substantial gains in improving the intelligibility of the two types of phase-based speech, i.e., phase-based vocoded speech and HFS stimuli. The use of a high temporal modulation rate can compensate for the loss of envelope cues for the HFS stimuli processed with a large number of channels of analysis filters. Consistent with previous findings, the intelligibility improvement in this study may be attributed to the increased amount of temporal envelope cues preserved in phase-based speech.

Acknowledgments

This research was supported by Faculty Research Fund (Faculty of Education) and Seed Funding for Basic Research, The University of Hong Kong. This work was also supported Grant No. 31271056 from National Natural Science Foundation of China.

References and links

ANSI (**1997**). ANSI S3.5, *American National Standards Methods for Calculation of the Speech Intelligibility Index* (Acoustical Society of America, New York).
Chen, F., and Loizou, P. (**2011**). "Predicting the intelligibility of vocoded speech," Ear Hear. **32**, 331–338.
Dolson, M. (**1986**). "The phase vocoder: A tutorial," Comput. Music J. **10**, 14–27.
Dorman, M., Loizou, P., and Rainey, D. (**1997**). "Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs," J. Acoust. Soc. Am. **102**, 2403–2411.
Gilbert, G., and Lorenzi, C. (**2006**). "The ability of listeners to use recovered envelope cues from speech fine structure," J. Acoust. Soc. Am. **119**, 2438–2444.
Greenwood, D. D. (**1990**). "A cochlear frequency-position function for several species—29 years later," J. Acoust. Soc. Am. **87**, 2592–2605.
Kazama, M., Gotoh, S., Tohyama, M., and Houtgast, T. (**2010**). "On the significance of phase in the short term Fourier spectrum for speech intelligibility," J. Acoust. Soc. Am. **127**, 1432–1439.
Lorenzi, C., Gilbert, G., Carn, H., Garnier, S., and Moore, B. C. (**2006**). "Speech perception problems of the hearing impaired reflect inability to use temporal fine structure," Proc. Natl. Acad. Sci. U.S.A. **103**, 18866–18869.
McAulay, R., and Quatieri, T. (**1995**). "Sinusoidal coding," in *Speech Coding and Synthesis*, edited by W. Kleijn and K. Paliwal (Elsevier Science, New York).
Moore, B. C. (**2008**). "The role of temporal fine structure processing in pitch perception, masking, speech perception for normal-hearing hearing-impaired people," J. Assoc. Res. Otolaryngol. **9**, 399–406.
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 **416**, 87–90.
Studebaker, G. A. (**1985**). "A 'rationalized' arcsine transform," J. Speech Hear. Res. **28**, 455–462.
Wong, L. L., Soli, S. D., Liu, S., Han, N., and Huang, M. W. (**2007**). "Development of the Mandarin hearing in noise test (MHINT)," Ear Hear. **28**, 70S–74S.
Xu, L., Thompson, C. S., and Pfingst, B. E. (**2005**). "Relative contributions of spectral and temporal cues for phoneme recognition," J. Acoust. Soc. Am. **117**, 3255–3267.
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.