Systems/Circuits

Role of Binaural Temporal Fine Structure and Envelope Cues in Cocktail-Party Listening

Jayaganesh Swaminathan, Christine R. Mason, Timothy M. Streeter, Virginia Best, Elin Roverud, and Gerald Kidd, Jr
Department of Speech, Language and Hearing Sciences, Boston University, Boston, Massachusetts 02215

While conversing in a crowded social setting, a listener is often required to follow a target speech signal amid multiple competing speech signals (the so-called "cocktail party" problem). In such situations, separation of the target speech signal in azimuth from the interfering masker signals can lead to an improvement in target intelligibility, an effect known as spatial release from masking (SRM). This study assessed the contributions of two stimulus properties that vary with separation of sound sources, binaural envelope (ENV) and temporal fine structure (TFS), to SRM in normal-hearing (NH) human listeners. Target speech was presented from the front and speech maskers were either colocated with or symmetrically separated from the target in azimuth. The target and maskers were presented either as natural speech or as "noise-vocoded" speech in which the intelligibility was conveyed only by the speech ENVs from several frequency bands; the speech TFS within each band was replaced with noise carriers. The experiments were designed to preserve the spatial cues in the speech ENVs while retaining/eliminating them from the TFS. This was achieved by using the same/different noise carriers in the two ears. A phenomenological auditory-nerve model was used to verify that the interaural correlations in TFS differed across conditions, whereas the ENVs retained a high degree of correlation, as intended. Overall, the results from this study revealed that binaural TFS cues, especially for frequency regions below 1500 Hz, are critical for achieving SRM in NH listeners. Potential implications for studying SRM in hearing-impaired listeners are discussed.

Key words: auditory-nerve; cocktail-party problem; envelope; temporal fine structure; spatial release from masking; speech

Significance Statement

Acoustic signals received by the auditory system pass first through an array of physiologically based band-pass filters. Conceptually, at the output of each filter, there are two principal forms of temporal information: slowly varying fluctuations in the envelope (ENV) and rapidly varying fluctuations in the temporal fine structure (TFS). The importance of these two types of information in everyday listening (e.g., conversing in a noisy social situation; the "cocktail-party" problem) has not been established. This study assessed the contributions of binaural ENV and TFS cues for understanding speech in multiple-talker situations. Results suggest that, whereas the ENV cues are important for speech intelligibility, binaural TFS cues are critical for perceptually segregating the different talkers and thus for solving the cocktail party problem.

Introduction

In everyday listening situations such as conversing in a crowded social setting, a listener is often required to follow a target speech signal in the presence of multiple competing masking speech signals. This circumstance is commonly referred to as the "cocktail party" problem (Cherry, 1953) and has been studied extensively over the past several decades (for review, see Carlile, 2014; Bronkhorst, 2015). When the speech maskers are spatially separated from the target, the listener may gain a considerable advantage in perceptually segregating and selectively attending to the target source relative to the case where all the sounds arise from the same location (Freyman et al., 1999; Brungart, 2001; Freyman et al., 2001; Hawley et al., 2004), an effect known as "spatial release from masking" (SRM; Hirsh, 1950; Marrone et al., 2008b). SRM is an important advantage due to binaural hearing that aids speech understanding for normal-hearing (NH) listen-

Received Dec. 10, 2015; revised May 21, 2016; accepted June 19, 2016.

Author contributions: J.S., C.R.M., and G.K. designed research; J.S., C.R.M., and T.M.S. performed research; J.S., V.B., E.R., and G.K. analyzed data; J.S. wrote the paper.

This work was supported by the National Institutes of Health, National Institute on Deafness and Other Communication Disorders (Grants R01-DC04545, R01-DC013286, P30-DC04663, and R01-DC000100) and by the Air Force Office of Scientific Research (Grant FA9550-12-1-0171). We thank H. Steven Colburn, William Hartmann, and Louis D. Braida for helpful comments and discussions on this work and Lorraine Delhorne for helping with subject recruitment and data collection.

The authors declare no competing financial interests.

Correspondence should be addressed to Jayaganesh Swaminathan, PhD, Department of Speech, Language and Hearing Sciences, Boston University, 635 Commonwealth Avenue, 320, Boston University, Boston, MA 02215. E-mail: jswamy@bu.edu.

Present address: Starkey Hearing Research Center, 2150 Shattuck Avenue, Berkeley, CA 94704-1362. E-mail: jayaganesh_swaminathan@starkey.com.

DOI:10.1523/JNEUROSCI.4421-15.2016
 0270-6474/16/368250-08$15.00/0

ers in a "cocktail-party" like environment. However, solving this complex listening task can be extremely challenging for listeners with sensorineural hearing loss (SNHL) even with hearing aids (Marrone et al., 2008a) or cochlear implants (Loizou et al., 2009). It also appears to be a challenge for some listeners with clinically normal auditory thresholds (Ruggles et al., 2011; Swaminathan et al., 2015) especially older listeners (e.g., Gallun et al., 2013). Therefore, there is considerable theoretical and translational interest in understanding the contributions of specific binaural speech cues that aid listeners in achieving SRM.

Acoustic signals received by the auditory system pass first through an array of physiologically based band-pass filters. Conceptually, at the output of each filter, there are two principal forms of temporal information: fluctuations in the envelope (ENV), which are the relatively slow variations in amplitude over time, and fluctuations in the temporal fine structure (TFS), which are the rapid variations of the waveform with rate close to the center frequency of the filter. Although the relative roles of ENV and TFS cues for monaural speech perception have been studied extensively and are reasonably well understood (Drullman, 1995; Smith et al., 2002; Zeng et al., 2005; Gilbert and Lorenzi, 2006; Swaminathan and Heinz, 2012), the contributions of ENV and TFS to solving binaural tasks (and SRM) currently are not clearly understood.

Part of the difficulty in assessing the contributions of TFS and ENV cues for spatial hearing with complex broadband stimuli such as speech arises due to cochlea-generated interactions between TFS and ENV (Ghitza, 2001). When broadband speech is filtered through a set of narrow-band filters (such as cochlear filters), the TFS component of the broadband speech gets converted into (recovered) ENVs (Ghitza, 2001; Heinz and Swaminathan, 2009). Such interactions between the ENV and TFS components of band-pass-filtered speech signals limit the ability to retain/eliminate the spatial cues in ENV and/or TFS selectively to study their relative roles for SRM. In this study, we used "noise vocoding" (Dudley, 1939; Flanagan and Golden, 1966; Shannon et al., 1995) to evaluate systematically the role of ENV and TFS cues in SRM with speech stimuli. With this approach, within each frequency band, the speech ENV component was retained, but the speech TFS component was replaced with noise carriers, thereby eliminating the interactions between the speech TFS and ENV components. Furthermore, with this processing, intelligibility is conveyed only by the speech ENVs.

The target and masker sentences were first subjected to multichannel noise vocoder processing and then spatialized such that the target speech was presented from the front (0° azimuth) and the speech maskers were either colocated with or symmetrically separated from the target in azimuth. In the spatially separated conditions, the maskers were placed symmetrically to reduce the usefulness of long-term head shadow differences (Marrone et al., 2008b; Jones and Litovsky, 2011). Interaural correlations in TFS were varied by using either the same or different noise carriers in the two ears. Because the spatial cues were applied after creating the vocoded speech, it was assumed that the binaural cues were largely preserved in the speech ENVs and were selectively retained (same) or eliminated (different) in TFS. Interaural correlations in TFS and ENV were quantified from a phenomenological auditory-nerve model to verify that the correlation in TFS was varied across different processing conditions, as intended, while maintaining high correlations in speech ENVs.

Materials and Methods

Subjects. A total of 10 young NH subjects (9 females and 1 male) between 19 and 22 years of age participated in this study. All the listeners were native speakers of American English and had normal hearing with audiometric pure tone thresholds of <20 dB HL from 0.25 to 8 kHz. Informed consent was obtained in compliance with an approved Institutional Review Board protocol from the Boston University Human Research Protection Program. All subjects were paid for their participation in the study.

Speech stimuli. The target and masker were comprised of five-word sentences that were syntactically correct but not necessarily semantically meaningful. The stimuli were taken from a corpus of monosyllabic words recorded for the laboratory by Sensimetrics Corporation. The sentences had the structure <name> <verb> <number> <adjective> <object> and there were eight possible words in each category. On each trial, the listener heard three sentences spoken by three different randomly chosen female talkers from seven available talkers. One sentence was designated as the target and always contained the <name> "Jane" with other keywords being randomly selected from the available choices (e.g., "Jane bought two red shoes"). The two masker sentences contained randomly selected names (excluding "Jane") and key words that differed from the target and from each other.

Procedure. The stimuli were presented via Sennheiser HD 280 headphones to listeners seated in a double-walled sound-attenuating chamber (Industrial Acoustics). The digital signals were generated on a PC outside of the booth and then routed either through separate channels of Tucker-Davis Technologies System II hardware or through a RME HDSP 9632 24-bit soundcard (ASIO). Target and maskers were spatialized using KEMAR head-related transfer functions (HRTFs). The HRTFs were obtained using tone sweeps recorded in a single-walled Industrial Acoustics sound booth (12 feet × 13 feet × 7.5 feet). The choice of spatial conditions was made based on the results from Swaminathan et al. (2013) in which, for natural speech, the effect of angular separation between the target and the maskers on target intelligibility was studied systematically. In that study, the results for maskers placed at ±30° and ±45° did not differ from the results for the maskers at ±90°. Therefore, the ±30° and ±45° conditions were not tested here. That left three spatial conditions: one in which the target and both the maskers were colocated at 0° azimuth and two in which the target was presented at 0° and the maskers were placed symmetrically at either ±15° or ±90°.

On a given run, the two maskers were each fixed in level at 55 dB SPL and the level of the target was varied adaptively using a one-down one-up procedure that tracked the 50% correct point on the psychometric function (giving a threshold target-to-masker ratio, TMR). The target level was varied in 6 dB steps initially and then in 3 dB steps after the third reversal. Each run consisted of at least 25 trials and at least nine reversals. Subjects were instructed to identify the keywords originating from the front uttered by the target talker, the female saying "Jane." The possible responses were displayed graphically on a computer screen. Subjects reported the perceived target keywords using the computer mouse to select the buttons showing the keywords. Correct answer feedback was provided during testing. For the purposes of the tracking procedure, responses were counted as correct if the listener successfully identified three of the four keywords (excluding <name>).

Stimulus processing. The target and masker sentences either were presented as produced naturally or were presented after noise vocoding (Shannon et al., 1995). The noise-vocoded signals were created to retain the speech ENV cues, but not the speech TFS cues, which were replaced with noise carriers within frequency bands. The vocoded versions of the sentences were created using MATLAB (The MathWorks). To create the vocoded speech, each sentence was initially band-pass filtered into eight or 32 contiguous bands of equal bandwidth on a logarithmic frequency scale spanning 80 to 8000 Hz. The band-pass filters were created using the auditory chimera package as described in Smith et al. (2002). The Hilbert transform was applied in each band and the ENV component within the band was extracted as the magnitude of the Hilbert analytic signal. The ENV signal within each band was further low-pass filtered below 300 Hz with a fourth-order Butterworth filter (48 dB/oct rolloff).

The filtered ENV signals were used to modulate narrow bands of noise with the same bandwidth as the analysis bands. Finally, these modulated signals were band-pass filtered through the original analysis bands to attenuate any spectral splatter and then were summed across all bands to create the ENV-vocoded speech stimulus. Independent noise tokens were used for the target and masker sentences.

The vocoding was implemented before spatializing the stimuli. Different vocoded speech-processing conditions were created to retain the spatial cues in the speech ENVs while retaining/eliminating them from interaural TFS within selected frequency channels. This was achieved by using correlated/uncorrelated noise carriers in the two ears for each of the three sentences (target and maskers). Four speech-processing conditions were used in this experiment: (1) correlated TFS (Corr TFS), in which the noise carriers were the same in the two ears and thus the spatial cues were preserved in both the speech ENV and the TFS from the correlated noise carriers; (2) uncorrelated TFS (Uncorr TFS), in which the noise carriers were uncorrelated between the two ears preserving the spatial cues only in the speech ENV and not in the TFS; (3) low-pass correlated TFS (LP Corr TFS), in which the noise carriers were the same in the two ears for frequencies below 1500 Hz but differed above 1500 Hz so the spatial cues were preserved in both the ENV and TFS below 1500 Hz and only the ENVs above 1500 Hz; and (4) low-pass uncorrelated TFS (LP Uncorr TFS), in which the noise carriers were the same in the two ears only for frequencies above 1500 Hz preserving the spatial cues in both the ENV and TFS above 1500 Hz but only in the ENV below 1500 Hz.

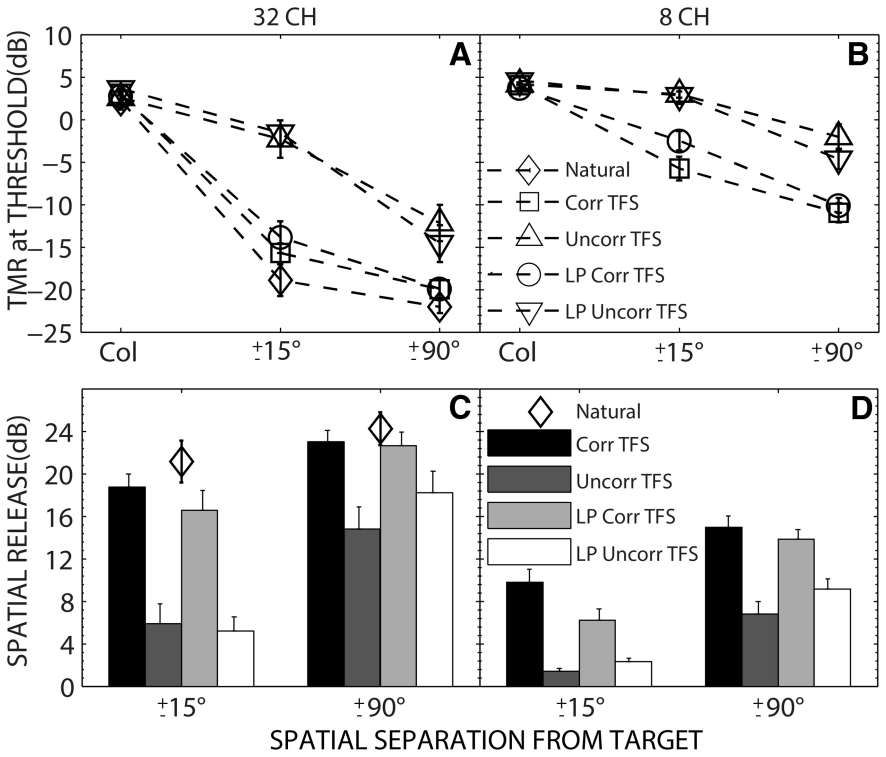

Figure 1. ***A***, ***B***, Group mean thresholds as a function of spatial separation of the two masker talkers from the target for speech processed to retain ENV cues with speech TFS replaced with noise carriers that were either interaurally correlated over the entire frequency region (Corr TFS), uncorrelated over the entire frequency region (Uncorr TFS), uncorrelated ≥1500 Hz (HP Uncorr TFS), or uncorrelated ≤1500 Hz (LP Uncorr TFS) over 32 (***A***) or eight (***B***) frequency bands. ***C***, ***D***, SRM as a function of the horizontal separation of the two masker talkers from the target for the conditions shown in ***A*** and ***B***. Thresholds and SRMs for natural speech is also shown in ***A*** and ***C***, respectively. Error bars indicate ±1 SEM.

In the first session, each listener was tested with the natural speech condition, which consisted of 3 spatial configurations (0, ±15°, and ±90°) × 6 runs for a total of 18 runs that were presented in a random order. After the first session, each listener was tested in 4 speech-processing conditions (Corr TFS, Uncorr TFS, LP Corr TFS and LP Uncorr TFS) × 2 vocoder conditions (32 and 8 channels) × 3 spatial configurations (0, ±15°, and ±90°) × 6 runs for a total of 144 runs, which were completed in multiple sessions. Each session presented a speech-processing and vocoder conditions across different spatial configurations. The ordering of the spatial configurations within a speech-processing and vocoder conditions was randomized. The ordering of speech-processing and vocoder conditions was randomized across subjects.

Results

Effect of varying interaural correlations in TFS on spatial release from masking

Figure 1*A* shows the group mean TMRs at threshold in dB as a function of spatial separation (0, ±15°, and ±90°) between target and maskers for each speech-processing condition. The different speech-
processing conditions include natural speech and ENV-vocoded speech created over 32 spectral channels with varying amounts of interaural correlations in TFS. TMR at threshold was calculated as the level of the target at adaptive threshold minus the fixed masker level (55 dB SPL). Figure 1*C* shows the mean SRM for the conditions shown in Figure 1*A*. SRM, for each listener, was calculated as the difference in thresholds between colocated and separated conditions. The group mean thresholds and SRM values are also provided in Table 1.

Table 1. Group mean thresholds and SRM for data shown in Figure 1

	32 Channels	8 Channels	Natural speech
Thresholds (dB)[a]			
Col	3.1, 2.5, 2.8, 3.7	4.0, 4.2, 3.7, 4.5	2.2
±15°	−15.7, −2.3, −13.8, −1.6	−5.8, 2.9, −2.5, 2.8	−18.9
±90°	−19.9, −12.2, −19.9, 14.6	−10.9, −2.0, −10.1, −4.6	−22.0
SRM (dB)[a]			
±15°	18.7, 5.9, 16.6, 5.2	9.8, 1.4, 6.2, 2.3	21.1
±90°	22.9, 14.8, 22.6, 18.2	14.9, 6.8, 13.8, 9.2	24.2

[a]Corr TFS, Uncorr TFS, LP Corr TFS, LP Uncorr TFS.

The colocated and separated thresholds were nearly identical for natural and 32-channel vocoded speech with correlated TFS (Corr TFS). This is consistent with considerable past work showing that speech intelligibility does not depend on preserving the natural speech TFS, only the ENVs (Shannon et al., 1995). Second, the main difference in the results from the four processing conditions was the decrease in thresholds with spatial separation, which was significantly greater when the stimuli had correlated low-frequency TFS (Corr TFS and LP Corr TFS). Absence of interaural low-frequency TFS cues (Uncorr TFS and LP Uncorr TFS) resulted in relatively higher spatially separated thresholds. As a result, the SRM was substantially larger for the Corr TFS and LP Corr TFS conditions compared with Uncorr TFS and LP Uncorr TFS conditions.

The highest mean thresholds were found when the target and masker talkers were colocated regardless of the processing condi-

tion. For natural speech and for all processing conditions, thresholds improved (decreased) with an increase in the amount of spatial separation between the target and the maskers. However, there were differences in the amount of improvement with spatial separation across different conditions. For natural speech and for processing conditions that had interaurally correlated low-frequency TFS (Corr TFS and LP Corr TFS), there was a steep improvement in thresholds from colocated to ±15° and just a small improvement from ±15° to ±90°. Mean SRM across these three conditions improved by ~4 dB from ±15° to ±90° (from 19 dB to 23 dB). In contrast, for processing conditions with uncorrelated low-frequency TFS (Uncorr TFS and LP Uncorr TFS), the thresholds improved only marginally from colocated to ±15° and steeply from ±15° to ±90°. Mean SRM across these two conditions improved by ~11 dB from ±15° to ±90° (from 6 dB to 17 dB). A two-way repeated measures ANOVA on the thresholds shown in Figure 1*A* found significant main effects of speech-processing condition ($F_{(4,36)} = 28.7$, $p < 0.001$, partial $\eta^2 = 0.761$), spatial separation ($F_{(2,18)} = 226.0$, $p < 0.001$, partial $\eta^2 = 0.962$), and a significant interaction ($F_{(8,72)} = 31.64$, $p < 0.001$, partial $\eta^2 = 0.624$).

Figure 1, *B* and *D*, shows group mean thresholds and SRM for ENV speech processed over eight spectral channels with different processing conditions. Overall, the separated thresholds were elevated and the SRM was reduced in the coarsely vocoded, eight-channel condition compared with the 32-channel condition, although similar trends were observed across processing conditions. A two-way repeated measures ANOVA on the thresholds shown in Figure 1*B* found significant main effects of speech-processing condition ($F_{(3,27)} = 19.6$, $p < 0.001$, partial $\eta^2 = 0.685$), spatial separation ($F_{(2,18)} = 121.1$, $p < 0.001$, partial $\eta^2 = 0.931$), and a significant interaction ($F_{(6,54)} = 16.2$, $p < 0.001$, partial $\eta^2 = 0.643$). Mean SRM across the conditions with correlated low-frequency TFS (Corr TFS and LP Corr TFS) was ~8 dB at ±15° and improved to ~14 dB at ±90°. For processing conditions with uncorrelated low-frequency TFS (Uncorr TFS and LP Uncorr TFS), the mean SRM was ~2 dB at ±15° and improved to ~8 dB at ±90°.

Predicted auditory-nerve fiber responses with interaurally correlated and uncorrelated TFS

A disadvantage of using noise-band vocoders is that within each frequency band, the intrinsic fluctuations from the ENV of the noise-band carrier can interfere with and cause disruption of the original speech ENVs (Whitmal et al., 2007). This can subsequently affect the binaural cues available in the speech ENVs when modulated with uncorrelated noise carriers (TFS) in the two ears. Furthermore, the relationship between "acoustic" TFS (extracted using signal processing techniques such as vocoders) and neural TFS ("phase-locked" information in the auditory nerve, AN) is not obvious and needs to be approached with caution (Shamma and Lorenzi, 2013). In this study, we quantified the interaural correlations in TFS and ENV from the output of a phenomenological AN model to determine whether the correlations in TFS varied across different processing conditions, as intended, while maintaining high correlations in speech ENVs and if the variations in interaural correlation in acoustic TFS and ENV were preserved in the neural coding of TFS and ENV at the level of the AN.

Neural cross-correlation coefficients, ρ_{TFS} and ρ_{ENV}, were used to quantify the similarity between TFS or ENV components of different spike-train responses (Heinz and Swaminathan, 2009; Swaminathan, 2010; Swaminathan and Heinz, 2012). Specifically, ρ_{TFS} and ρ_{ENV} provide metrics ranging from 0 to 1 that represent the degree of similarity between responses under two different conditions (e.g., vocoded speech created with noise carriers A and B). These metrics quantify the amount of coding in common between the two responses (cross-correlograms) relative to the amount of coding within each individual response (auto-correlograms). Figure 2 illustrates the use of neural cross-correlational coefficients (ρ_{TFS} and ρ_{ENV}) to compute interaural correlation in TFS and ENV for the noise-vocoded speech. The details involved in calculating ρ_{TFS} and ρ_{ENV} have been described previously (Heinz and Swaminathan, 2009; Swaminathan and Heinz, 2012) and are briefly summarized below.

The correlograms are temporal representations of stimulus-related periodicities as coded by individual AN fibers. Shuffled auto correlograms are computed by tallying spike intervals within and across spike trains, yielding a more robust characterization of temporal responses (Fig. 2*B*) than classic all-order interval histograms (Ruggero, 1973). Normalized shuffled auto correlograms are plotted as a function of time delay and are much like auto-correlation functions (Fig. 2*C*, thick lines). For the example shown in Figure 2, spike trains were obtained first from a model AN fiber with a characteristic frequency (CF) of 744 Hz in response to a 32-channel ENV-vocoded speech stimulus created with two independent noise carriers, A and B. For each noise carrier, spike trains were obtained for positive and negative stimulus polarity presentations (A^+/A^- and B^+/B^-).

Shuffled auto-correlograms (Fig. 2*C*,*D*, thick lines) were generated by tallying coincidences from spike trains obtained in response to the same stimuli (e.g., A^+/A^+ or A^-/A^-). TFS and ENV coding can be separated by comparing the responses to a stimulus and its polarity-inverted pair (e.g., A^+ with A^-) (Joris and Yin, 1992; Joris, 2003; 2006). Polarity inversion acts to invert the TFS, but does not affect ENV. Cross-polarity correlograms are computed by comparing spikes from A^+ and A^- (Fig. 2*C*,*D*, thin lines). To emphasize TFS coding, difcors were computed as the difference between the shuffled auto correlogram (original ENV, original TFS; thick line in Fig. 2*C*,*D*) and the cross-polarity correlogram (original ENV and inverted TFS; thin line in Fig. 2*C*,*D*), where the difcor peak height quantifies the strength of TFS coding. To quantify ENV coding, sumcors were computed as the average of the shuffled auto correlogram and the cross-polarity correlogram. The ENV correlograms were further low-pass filtered at 64 Hz to retain the slow phonemic ENV cues that have been suggested to be important for speech intelligibility (Rosen, 1992; Swaminathan and Heinz, 2012).

For shuffled cross-correlograms (Fig. 2*E*), spike coincidences were counted from pairs of spike trains derived from two different stimuli (e.g., vocoded speech with noise carrier A and noise carrier B presented to right ear). To compute "interaural" correlations for the conditions tested in this study, it was assumed that the pair of spike trains arrived from two different ears, 744 Hz fiber in the left ear responding to vocoded speech created with noise carrier A and a 744 Hz fiber in the right ear responding to vocoded speech created with noise carrier B. Counting the spike coincidences between responses from such a pair predicts the output of a simple binaural coincidence detector, which receives inputs from two ears without any cochlear disparities (Fig. 2*A*). Cross-stimulus correlograms (e.g., A^+/B^+, thick line in Fig. 2*E*) and cross-stimulus, cross-polarity correlograms (e.g., A^+/B^-, Fig. 2*E*, thin line) were computed to facilitate the separation of TFS and ENV cross-correlations by using difcors and sumcors, respectively.

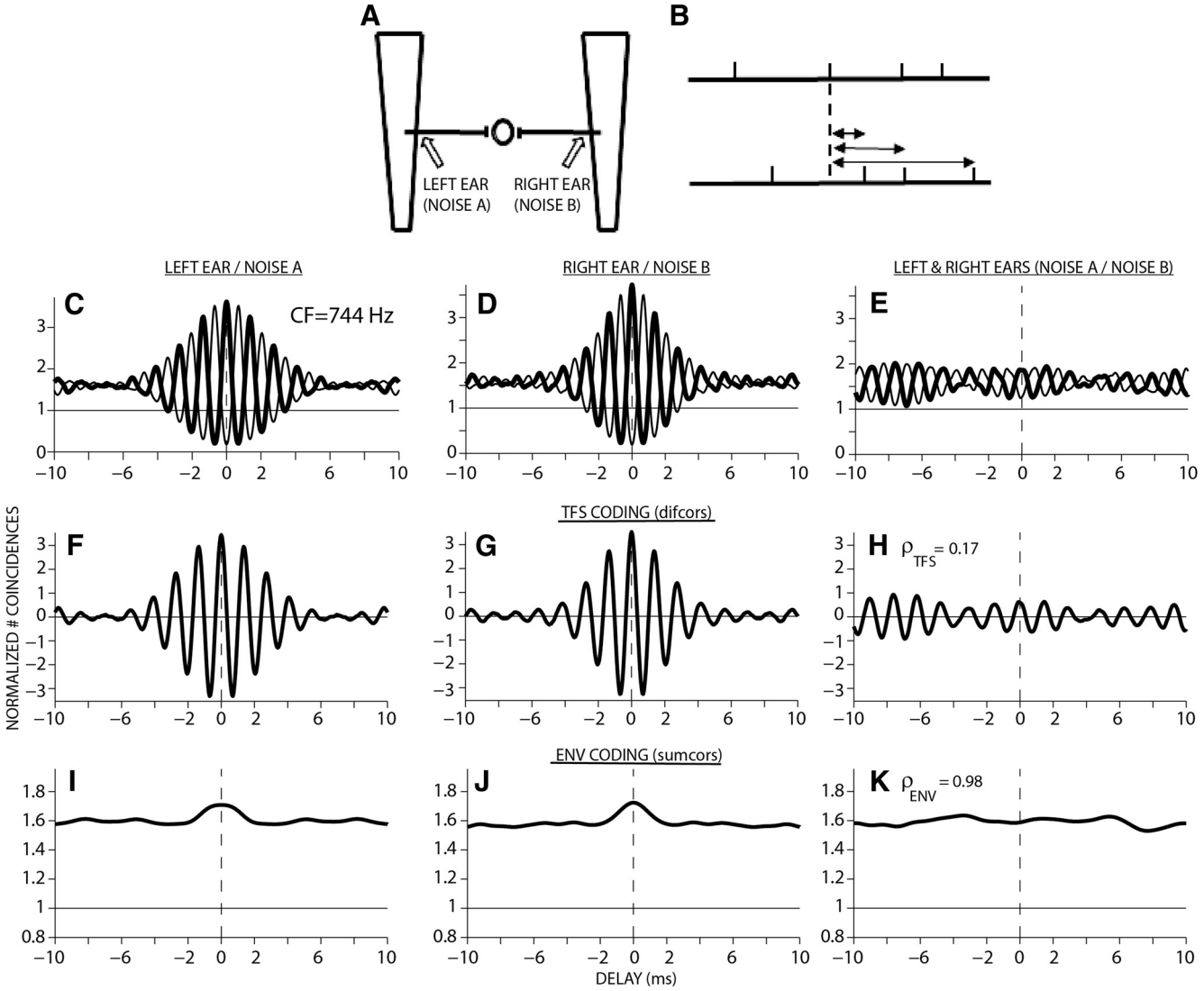

Figure 2. Correlogram analyses used to quantify the (interaural) correlations in neural coding of ENV and TFS with noise-vocoded speech. ***A***, To compute interaural correlations, pairs of spike trains are treated as if derived from two different ears. Counting of spike coincidences between responses from such a pair predicts the output of a simple binaural coincidence detector, which would receive inputs from a physiological site in the two ears. Trapezoidal shapes represent uncoiled basilar membranes. ***B***, Construction of correlograms from two sets of spike trains. The delays at which coincidences are obtained between spikes in these conditions are tallied in a histogram. Columns 1 and 2 (***C***, ***D***, ***F***, ***G***, ***I***, ***J***) show temporal coding of ENV-vocoded speech created with noise carriers A and B, respectively; column 3 (***E***, ***H***, ***K***) illustrates the similarity in temporal coding between these two conditions (ρ_{TFS} and ρ_{ENV}). Fiber CF = 744 Hz.

Neural cross-correlation coefficients (Heinz and Swaminathan, 2009) ranging between 0 and 1 were computed by comparing the degree of response similarity across stimuli and ears (Fig. 2*H*,*K*) to the degree of temporal coding for each stimulus presented to each ear individually (Fig. 2*F*,*G*,*I*,*J*). The cross-correlation coefficient for TFS was computed from the difcor peak heights as follows:

$$\rho_{\mathrm{TFS}} = \frac{\mathrm{difcor}_{AB}}{\sqrt{\mathrm{difcor}_{A} \times \mathrm{difcor}_{B}}}$$

Likewise, the neural cross-correlation coefficient for ENV was computed from the sumcor peak heights as follows:

$$\rho_{\mathrm{ENV}} = \frac{(\mathrm{sumcor}_{AB} - 1)}{\sqrt{(\mathrm{sumcor}_{A} - 1) \times (\mathrm{sumcor}_{B} - 1)}}$$

For the single-fiber responses illustrated in Figure 2, the interaural correlation in TFS for ENV-vocoded speech created with two independent noise carriers was very low ($\rho_{\mathrm{TFS}} = 0.17$) and the correlation in ENVs was close to 1 ($\rho_{\mathrm{ENV}} = 0.98$). These neural metrics provide a quantitative physiological framework with which to compare/relate the variations in acoustic TFS and ENVs to variations in neural TFS and ENVs as reflected in AN phase-locked responses.

The interaural correlations in neural ENV and TFS were computed for the different speech-processing conditions of the psychophysical experiment (Corr TFS, LP Corr TFS, Uncorr TFS, and LP Uncorr TFS) for ENV-vocoded speech created over 32 channels. ENV-vocoded speech for a subset of 40 words from the corpus was created for the different speech-processing conditions. Stimuli were resampled to 100 kHz before presentation to the AN model to obtain the spike times. Thirty-two high-spontaneous-rate AN fibers with CFs matching the center frequencies of the analysis filters used to create the vocoded speech were selected. A total of 15360 sets of neural cross-correlation coefficients were computed (32 AN fibers × 40 words × 4 speech

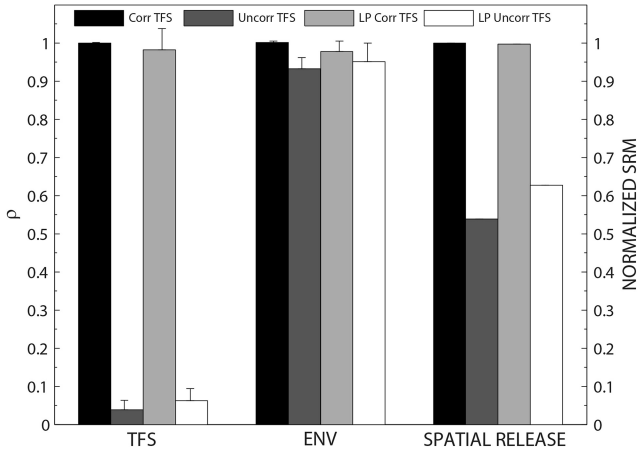

Figure 3. Mean interaural correlations in neural ENV and TFS across different speech-processing conditions for 32-channel ENV-vocoded speech. The normalized SRM derived from measured SRM from NH listeners at ±90° for these processing conditions is also included for comparison.

conditions × 3 repetitions). TFS coding of fibers with CF > 2000 Hz were not included due to roll-off in phase locking.

Figure 3 shows the mean interaural correlations in neural ENV and TFS across different speech-processing conditions. For comparing trends with the psychophysical results, the normalized SRM at ±90° for the 32-channel vocoder for different processing conditions is also included in Figure 3. Across different processing conditions, the colocated thresholds were similar and the SRM was primarily influenced by differences in the separated thresholds (Fig. 1*A*). The normalized SRM for each processing condition, which varied from 0 to 1, was computed as follows: (mean measured SRM for a specific condition)/(maximum mean measured SRM across all conditions).

For the different processing conditions used in this study, the interaural correlations (ρ) for TFS were close to 1 for Corr TFS and LP Corr TFS and close to the noise floor for Uncorr TFS and LP Uncorr TFS. The interaural correlations in ENVs were close to 1 for all processing conditions. The important points from this analysis are as follows. First, the model output supports the conclusion that the acoustic manipulations of TFS in the noise-vocoded masker speech were preserved in the AN responses without disrupting the high interaural correlations in speech ENVs. Second, because the spatial cues were applied after creating the vocoded speech, this information was retained in the neural TFS for the Corr TFS and LP Corr TFS and eliminated from the TFS for Uncorr TFS and LP Uncorr TFS. The spatial cues were retained in the neural ENVs across all processing conditions. Finally, the trends in SRM across different processing conditions matched with the trends observed in the interaural correlations in neural TFS; that is, the normalized SRM was highest for Corr TFS and LP Corr TFS and lowest for Uncorr TFS and LP Uncorr TFS.

Discussion

This study assessed the contributions of binaural ENV and TFS cues for spatial release from speech-on-speech masking. A key feature of the design was varying the strength of the spatial information available to the listener by manipulating the interaural TFS while preserving the ENVs across the different conditions. This goal was achieved by creating different noise band vocoder conditions in which the interaural correlation in TFS was varied by presenting the same versus different noise carriers to the two ears. Because the spatial cues were applied after creating the vocoded speech, the binaural cues were largely preserved in the speech ENVs and selectively retained/eliminated in the TFS. A phenomenological auditory-nerve model was used to verify that the interaural correlations in TFS differed across conditions, whereas the ENVs retained a high degree of correlation, as intended.

For vocoded speech created with 32 narrow channels, replacing speech TFS with (interaurally correlated) noise carriers yielded comparable SRM to that observed with natural speech. These results suggest that speech ENV cues combined with correlated noise carriers are sufficient to produce spatial benefits with competing speech sounds when the ENV cues are provided over 32 independent channels. This large benefit is due to the lower thresholds obtained in the spatially separated conditions. Therefore, the spatial cues conveyed by correlated noise carriers appears to be sufficient for the listener to use spatial separation as a means for selecting one speech source among competing speech sources even at very low TMRs. Overall, this suggests that preserving speech TFS information is not necessary as long as the TFS information is interaurally correlated and that there are a relatively high number of independent channels conveying ENV information to support speech intelligibility.

However, consistent with Best et al. (2012), the thresholds were higher in the spatially separated conditions for the more coarsely vocoded speech (i.e., 8 channels). The underlying reasons for this observation are not fully clear. A large number of channels (e.g., 32) yielded performance very similar to natural speech. As the number of channels decreases, the quality of the vocoded signal is degraded although the unmasked intelligibility is maintained (at least for these closed-set materials). Progressively broadening the vocoder channels can lead to an increased overlap of spectrotemporal acoustic cues between the target and masker, thereby increasing energetic masking. Previous studies have shown that SRM decreases as the amount/proportion of energetic masking increases due primarily to an elevation in thresholds in spatially separated conditions as observed here for the more coarsely vocoded conditions (Arbogast et al., 2005; Marrone et al., 2008b, a, Best et al., 2012). Furthermore, the vocoder processing over fewer broader channels can severely reduce or eliminate pitch and other voice difference cues between different talkers, increasing target/masker confusability, and, potentially, informational masking (Qin and Oxenham, 2003; Freyman et al., 2008; Garadat et al., 2009).

Results from this study also revealed that disrupting the interaural correlation in TFS, especially in the low-frequency regions (≤ 1500 Hz), resulted in elevated (poorer) spatially separated thresholds and a reduction in SRM. Even with 32-channel vocoded speech, the thresholds with uncorrelated TFS were poorer than with correlated TFS by at least 10 dB. For these conditions, the results from physiology-based modeling showed high interaural correlations of speech ENVs. Therefore, when considered together, the findings from the psychophysical experiment and neural modeling suggest that, whereas the ENV cues are important for speech intelligibility, low-frequency TFS is a primary cue that aids in spatial release with speech-on-speech masking. This result is in general agreement with the finding that interaural-level differences conveyed by speech ENVs are not sufficient for restoring SRM in simulated cochlear implant listening conditions (Ihlefeld and Litovsky, 2012). This result is also in agreement with the finding that SRM largely depends on receiving interaural time differences from low frequencies (Kidd et al., 2010).

The underlying reasons for the reduction in SRM with disruptions in interaural TFS are not clear. Across all four processing conditions, the peripheral/neural overlap of target and maskers (i.e., energetic masking) was similar. However, it is likely that disruptions in interaural TFS rendered the maskers spatially diffuse, thereby limiting the release from informational masking normally afforded by spatial separation. Indeed, it has been shown that lateralization of speech in quiet is affected by reducing the interaural correlation in TFS (Drennan et al., 2007). Further experiments are warranted to investigate how disruptions of interaural TFS can affect release from IM, thereby influencing SRM.

Although this study was not intended to simulate any specific physiological mechanism causing hearing loss and their associated effects on SRM, it is nonetheless possible to speculate about the potential implications of these findings for studying (and improving) SRM in listeners with hearing loss. First of all, the main implication here is that, even when the speech TFS was replaced with correlated noise carriers, a sufficient number of independent channels conveying ENV information can provide considerable benefit in achieving spatial release with speech on speech masking. As few as eight channels yielded SRMs of 15 dB in this study. A well known consequence of SNHL is reduced frequency selectivity that results from the broadening of the peripheral auditory filters (Liberman and Dodds, 1984; Glasberg and Moore, 1986; Patuzzi et al., 1989; Ruggero and Rich, 1991). Due to this broadened tuning, there are fewer peripheral channels providing independent information. It has been shown that hearing impaired (HI) listeners with symmetric binaural hearing (as measured by audiograms) often demonstrate reduced SRM compared with NH listeners primarily due to increased thresholds in spatially separated conditions (Arbogast et al., 2005; Marrone et al., 2008b; Best et al., 2012) similar to the findings here. If the hearing loss is indeed perfectly symmetric in the two ears, resulting in similar coding of TFS, the results from this study suggest that a promising avenue to explore for improving SRM for HI subjects is to try to restore ENV cues over several independent channels (e.g., 32 channels). Alternately, any TFS coding deficit that can arise from noise-induced hearing loss (Henry and Heinz, 2012), from aging (Gallun et al., 2013), or because of "hidden hearing loss" (Plack et al., 2014) can possibly result in disruptions to interaural correlations in TFS. The results from this study suggest that such TFS coding deficits can affect SRM even when the ENVs are provided over several narrow channels. In such situations, synchronizing the TFS across the two ears and restoring appropriate spatial cues will be critical for improving SRM in listeners with hearing loss.

References

Arbogast TL, Mason CR, Kidd G Jr (2005) The effect of spatial separation on informational masking of speech in normal-hearing and hearing-impaired listeners. J Acoust Soc Am 117:2169–2180. CrossRef Medline

Best V, Marrone N, Mason CR, Kidd G Jr (2012) The influence of non-spatial factors on measures of spatial release from masking. J Acoust Soc Am 131:3103–3110. CrossRef Medline

Bronkhorst AW (2015) The cocktail-party problem revisited: early processing and selection of multi-talker speech. Atten Percept Psychophys 77: 1465–1487. CrossRef Medline

Brungart DS (2001) Informational and energetic masking effects in the perception of two simultaneous talkers. J Acoust Soc Am 109:1101–1109. CrossRef Medline

Carlile S (2014) Active listening: Speech intelligibility in noisy environments. Acoustics Australia 42:90–96.

Cherry EC (1953) Some experiments on the recognition of speech, with one and with two ears. J Acoust Soc Am 25:975–979. CrossRef

Drennan WR, Won JH, Dasika VK, Rubinstein JT (2007) Effects of temporal fine structure on the lateralization of speech and on speech understanding in noise. J Assoc Res Otolaryngol 8:373–383. CrossRef Medline

Drullman R (1995) Temporal envelope and fine structure cues for speech intelligibility. J Acoust Soc Am 97:585–592. CrossRef Medline

Dudley HW (1939) The vocoder. Bell Labs Rec 18.

Flanagan JL, Golden RM (1966) Phase vocoder. Bell Systems Technical J 1493–1509.

Freyman RL, Helfer KS, McCall DD, Clifton RK (1999) The role of perceived spatial separation in the unmasking of speech. J Acoust Soc Am 106:3578–3588. CrossRef Medline

Freyman RL, Balakrishnan U, Helfer KS (2001) Spatial release from informational masking in speech recognition. J Acoust Soc Am 109:2112–2122. CrossRef Medline

Freyman RL, Balakrishnan U, Helfer KS (2008) Spatial release from masking with noise-vocoded speech. J Acoust Soc Am 124:1627–1637. CrossRef Medline

Gallun FJ, Diedesch AC, Kampel SD, Jakien KM (2013) Independent impacts of age and hearing loss on spatial release in a complex auditory environment. Front Neurosci 7:252. CrossRef Medline

Garadat SN, Litovsky RY, Yu G, Zeng FG (2009) Role of binaural hearing in speech intelligibility and spatial release from masking using vocoded speech. J Acoust Soc Am 126:2522–2535. CrossRef Medline

Ghitza O (2001) On the upper cutoff frequency of the auditory critical-band envelope detectors in the context of speech perception. J Acoust Soc Am 110:1628–1640. CrossRef Medline

Gilbert G, Lorenzi C (2006) The ability of listeners to use recovered envelope cues from speech fine structure. J Acoust Soc Am 119:2438–2444. CrossRef Medline

Glasberg BR, Moore BC (1986) Auditory filter shapes in subjects with unilateral and bilateral cochlear impairments. J Acoust Soc Am 79:1020–1033. CrossRef Medline

Hawley ML, Litovsky RY, Culling JF (2004) The benefit of binaural hearing in a cocktail party: effect of location and type of interferer. J Acoust Soc Am 115:833–843. CrossRef Medline

Heinz MG, Swaminathan J (2009) Quantifying envelope and fine-structure coding in auditory nerve responses to chimaeric speech. J Assoc Res Otolaryngol 10:407–423. CrossRef Medline

Henry KS, Heinz MG (2012) Diminished temporal coding with sensorineural hearing loss emerges in background noise. Nat Neurosci 15: 1362–1364. CrossRef Medline

Hirsh IJ (1950) The relation between localization and intelligibility. J Acoust Soc Am 22:196–200. CrossRef

Ihlefeld A, Litovsky RY (2012) Interaural level differences do not suffice for restoring spatial release from masking in simulated cochlear implant listening. PLoS One 7:e45296. CrossRef Medline

Jones GL, Litovsky RY (2011) A cocktail party model of spatial release from masking by both noise and speech interferers. J Acoust Soc Am 130: 1463–1474. CrossRef Medline

Joris PX (2003) Interaural time sensitivity dominated by cochlea-induced envelope patterns. J Neurosci 23:6345–6350. Medline

Joris PX (2006) A dogged pursuit of coincidence. J Neurophysiol 96: 969–972. CrossRef Medline

Joris PX, Yin TC (1992) Responses to amplitude-modulated tones in the auditory nerve of the cat. J Acoust Soc Am 91:215–232. CrossRef Medline

Kidd G Jr, Mason CR, Best V, Marrone N (2010) Stimulus factors influencing spatial release from speech-on-speech masking. J Acoust Soc Am 128:1965–1978. CrossRef Medline

Liberman MC, Dodds LW (1984) Single-neuron labeling and chronic cochlear pathology. III. Stereocilia damage and alterations of threshold tuning curves. Hear Res 16:55–74. CrossRef Medline

Loizou PC, Hu Y, Litovsky R, Yu G, Peters R, Lake J, Roland P (2009) Speech recognition by bilateral cochlear implant users in a cocktail-party setting. J Acoust Soc Am 125:372–383. CrossRef Medline

Marrone N, Mason CR, Kidd G Jr (2008a) The effects of hearing loss and age on the benefit of spatial separation between multiple talkers in reverberant rooms. J Acoust Soc Am 124:3064–3075. CrossRef Medline

Marrone N, Mason CR, Kidd G (2008b) Tuning in the spatial dimension: evidence from a masked speech identification task. J Acoust Soc Am 124: 1146–1158. CrossRef Medline

Patuzzi RB, Yates GK, Johnstone BM (1989) Outer hair cell receptor current and sensorineural hearing loss. Hear Res 42:47–72. CrossRef Medline

Plack CJ, Barker D, Prendergast G (2014) Perceptual consequences of "hidden" hearing loss. Trends Hear 18.

Qin MK, Oxenham AJ (2003) Effects of simulated cochlear-implant processing on speech reception in fluctuating maskers. J Acoust Soc Am 114:446–454. CrossRef Medline

Rosen S (1992) Temporal information in speech: acoustic, auditory, and linguistic aspects. Philos Trans R Soc Lond B 336:367–373. CrossRef

Ruggero MA (1973) Response to noise of auditory nerve fibers in the squirrel monkey. J Neurophysiol 36:569–587. Medline

Ruggero MA, Rich NC (1991) Furosemide alters organ of Corti mechanics: Evidence for feedback of outer hair cells upon the basilar membrane. J Neurosci 11:1057–1067. Medline

Ruggles D, Bharadwaj H, Shinn-Cunningham BG (2011) Normal hearing is not enough to guarantee robust encoding of suprathreshold features important for everyday communication. Proc Natl Acad Sci U S A 108:15516–15521. CrossRef Medline

Shamma S, Lorenzi C (2013) On the balance of envelope and temporal fine structure in the encoding of speech in the early auditory system. J Acoust Soc Am 133:2818–2833. CrossRef Medline

Shannon RV, Zeng FG, Kamath V, Wygonski J, Ekelid M (1995) Speech recognition with primarily temporal cues. Science 270:303–304. CrossRef Medline

Smith ZM, Delgutte B, Oxenham AJ (2002) Chimaeric sounds reveal dichotomies in auditory perception. Nature 416:87–90. CrossRef Medline

Swaminathan J (2010) The role of envelope and temporal fine structure in the perception of noise degraded speech. PhD Dissertation, Purdue University; p. 231.

Swaminathan J, Heinz MG (2012) Psychophysiological analyses demonstrate the importance of neural envelope coding for speech perception in noise. J Neurosci 32:1747–1756. CrossRef Medline

Swaminathan J, Mason CR, Streeter TM, Best V, Kidd G (2013) Spatial release from masking for noise-vocoded speech. Proc Mtgs Acoust 19:1–8.

Swaminathan J, Mason CR, Streeter TM, Best V, Kidd G Jr, Patel AD (2015) Musical training, individual differences and the cocktail party problem. Sci Rep 5:11628. CrossRef Medline

Whitmal NA, Poissant SF, Freyman RL, Helfer KS (2007) Speech intelligibility in cochlear implant simulations: Effects of carrier type, interfering noise, and subject experience. J Acoust Soc Am 122:2376–2388. CrossRef Medline

Zeng FG, Nie K, Stickney GS, Kong YY, Vongphoe M, Bhargave A, Wei C, Cao K (2005) Speech recognition with amplitude and frequency modulations. Proc Natl Acad Sci U S A 102:2293–2298. CrossRef Medline