JARO 19: 467–478 (2018)
DOI: 10.1007/s10162-018-0672-x

Research Article

JARO
Journal of the Association for Research in Otolaryngology

Factors Affecting Speech Reception in Background Noise with a Vocoder Implementation of the FAST Algorithm

Shaikat Hossain[1] and Raymond L. Goldsworthy[1]

[1]*Department of Otolaryngology, University of Southern California, Los Angeles, CA, USA*

Received: 22 August 2017; Accepted: 23 April 2018; Online publication: 9 May 2018

ABSTRACT

Speech segregation in background noise remains a difficult task for individuals with hearing loss. Several signal processing strategies have been developed to improve the efficacy of hearing assistive technologies in complex listening environments. The present study measured speech reception thresholds in normal-hearing listeners attending to a vocoder based on the Fundamental Asynchronous Stimulus Timing algorithm (FAST: Smith et al. 2014), which triggers pulses based on the amplitudes of channel magnitudes in order to preserve envelope timing cues, with two different reconstruction bandwidths (narrowband and broadband) to control the degree of spectrotemporal resolution. Five types of background noise were used including same male talker, female talker, time-reversed male talker, time-reversed female talker, and speech-shaped noise to probe the contributions of different types of speech segregation cues and to elucidate how degradation affects speech reception across these conditions. Maskers were spatialized using head-related transfer functions in order to create co-located and spatially separated conditions. Results indicate that benefits arising from voicing and spatial cues can be preserved using the FAST algorithm but are reduced with a reduction in spectral resolution.

Keywords: speech comprehension, vocoder, cocktail party, cochlear implants, FAST

INTRODUCTION

In complex listening environments, multiple acoustic and linguistic cues allow a listener to segregate different sound sources. The "cocktail party" problem (Cherry 1953) refers to commonly encountered situations where a listener must attend to a target talker in the presence of competing sound sources. In such situations, listeners use different cues to derive a release from masking. Three types of cues which affect such release from masking are semantic, voicing, and spatial cues (Moore 2012).

Semantic cues consist of differences in the semantic content of two concurrent streams of speech and are a subset of cues which contribute to informational masking. Informational masking is a term coined by Pollack (1975) which refers to a form of masking in which auditory detection or discrimination is degraded due to the signal being embedded in sounds with a similar semantic context (Leek et al. 1991; Durlach et al. 2003). This form of masking is thought to result from central auditory processes and is distinct yet complementary to energetic masking, which occurs due to the spectrotemporal overlap of target and masker energy at the auditory periphery. Semantic masking has traditionally been quantified as the improvement in speech intelligibility when a masker with similar semantic content is substituted with either time-reversed speech or noise with similar spectrotemporal characteristics to speech but with no semantic content (e.g., speech-shaped noise) (Watson 2005; Yost 2006; Kidd Jr et al. 2007).

Voicing cues consist of talker-specific differences such as differences in fundamental frequency and formant frequencies that arise from vocal fold and vocal tract physiology, respectively. These types of voicing cues are important for the perception of voice

Correspondence to: Shaikat Hossain · Department of Otolaryngology · University of Southern California · Los Angeles, CA, USA. email: shaikath@usc.edu

gender (Smith and Patterson 2005; Hillenbrand and Clark 2009; Skuk and Schweinberger 2014). Typically, males have lower average fundamental and formant frequencies than females due to differences in the length, mass, and tension of the vocal folds and the length and overall shape of the vocal tract (Fitch and Giedd 1999; Smith and Patterson 2005). In a cocktail party scenario, the ability to discriminate between different voices leads to improvements in speech intelligibility when the talkers differ in fundamental and/or formant frequencies (Brokx and Nooteboom 1982; Brungart 2001a, b; Darwin et al. 2003, Başkent and Gaudrain 2016). Furthermore, voice discrimination has been found to be correlated with the recognition of vowels and consonants spoken by adult male and female talkers (Li and Fu 2011).

Spatial cues for localization in the horizontal plane include interaural level and timing differences (ILDs and ITDs, respectively). These binaural cues facilitate spatial release from masking (SRM), an improvement in speech intelligibility amidst competing talkers when the target speech is moved from co-located with the masker to spatially separated (Hirsh 1948; Hirsh 1950; Cherry 1953; Arbogast et al. 2002; Marrone et al. 2008). With respect to the contributions of bottom-up cues, this SRM benefit can be broken down in terms of three binaural components: the better ear effect, binaural summation, and the squelch effect. The better ear effect is the benefit derived from a monaural comparison of the signal-to-noise ratio (SNR) at each ear, with the ear facing the target speaker resulting in a more favorable SNR than the contralateral ear (Zurek 1993). Binaural summation occurs when both ears are activated by a sound presented from the front, where the signals are summed together and, as a result, easier to hear due to the processing of redundant information across ears (Blauert 1997). Lastly, additional benefit is derived from contrasting spectrotemporal ILDs and ITDs between the target speech and the masker, leading to a form of interaural correlation where cues that are highly correlated are perceptually grouped. This phenomenon is referred to as "binaural unmasking" or "binaural squelch" (Blauert 1997; Bronkhorst 2000).

The tradeoff between semantic, voicing, and spatial cues underlies listeners' abilities to comprehend speech in cocktail party listening scenarios. For normal-hearing listeners, there can be a saturation of benefits derived from multiple speech segregation cues. When the target and masker are of opposite genders, there is little additional benefit from semantic cues (Brungart 2001a, b; Brungart and Simpson 2002). Similarly, differences in spatial location have been found to reduce the effects of semantic distraction (Freyman et al. 2001; Gallun et al. 2005; Kidd Jr. et al. 2005) and spatial attention (Carlile and Corkhill 2015).

The pattern of interaction between different speech segregation cues changes with auditory resolution. Previous studies have shown that degrading the spectral and temporal fine structure cues affects patterns of interaction between speech segregation cues. Freyman et al. (2008) investigated the contributions of semantic and spatial cues to SRM benefits with normal-hearing listeners attending to speech degraded using vocoder processing and found that there were no SRM benefits for nonsense sentences when presented against two-talker maskers but that there was a large SRM benefit (~20 dB SNR) when single words excised from the nonsense sentences were used instead. Other studies using vocoder processing investigated the contributions of informational and voicing cues demonstrated that vocoder processing led to a reduction in voice gender cues and a subsequent increase in susceptibility to informational masking (Qin and Oxenham 2003; Stickney et al. 2004). The diminished benefit from voicing cues is likely due to reduced spectral resolution and a poor representation of pitch. This notion is supported by the finding that transmission of spectral envelope cues was linked to the number of spectral channels (Gaudrain and Başkent 2015). Fuller et al. (2014) performed a cue-weighting analysis for gender recognition and found that both fundamental frequency and spectral envelope cues related to vocal tract length were reduced in weighting for normal-hearing listeners attending to vocoder-processed speech.

Swaminathan et al. (2016) probed the relative contributions of different acoustic stimulus properties on SRM benefits in normal-hearing listeners. They used a noise vocoder to retain the envelope components while replacing the temporal fine structure with noise carriers. Binaural cues were preserved in the envelopes and selectively retained or eliminated in the fine structure through correlation of noise carriers in specific frequency regions. They found a decrease in speech reception thresholds with increasing spatial separation of target and masker which was greater for conditions where stimuli had correlated low-frequency temporal fine structure, as compared to conditions with uncorrelated temporal fine structure. Findings from their study indicate that speech reception thresholds (SRTs) with 32-channel vocoders using correlated temporal fine structure were similar to SRTs for unprocessed speech. However, the SRM was substantially reduced when the speech was presented through eight broad vocoder channels even with correlated TFS. This demonstrates the importance of spectral resolution for the transmission of temporal fine structure to ultimately facilitate SRM.

In addition to their utility as a tool to investigate the relative contributions of acoustic envelope and fine structure cues, vocoders have been used with normal-hearing listeners as a method to understand how cochlear implant signal processing affects hearing (Shannon et al. 1995; Dorman et al. 1998; Qin and Oxenham 2003; Stickney et al. 2004; Poissant et al. 2006). Presently, there is great interest for improving signal processing to coordinate bilateral cochlear implants (Kan and Litovsky, 2015). Several recent cochlear implant signal processing strategies have attempted to encode TFS cues in a perceptually meaningful way (van Hoesel and Tyler 2003; Vandali et al. 2005). One recent approach which alternatively seeks to encode envelope details is the Fundamental Asynchronous Stimulus Timing (FAST) algorithm, which triggers timing of pulses along the electrode array based on the temporal maxima of channel envelopes instead of using fixed high rate carrier pulses like conventional strategies such as continuous interleaved sampling (Smith et al. 2014). By triggering the pulses using this sparse approach, FAST is thought to achieve better precision in terms of encoding envelope details and may therefore be suitable for the encoding of envelope ITDs which could improve spatial hearing for bilateral cochlear implant users who rely solely on envelope cues given existing limitations in the transmission of temporal fine structure information in CI devices. Furthermore, it may potentially be more feasible to encode envelope details than TFS due to unwanted channel interactions resulting from the electroneural interface.

The present study examines a vocoder implementation of the FAST algorithm which we use to degrade spectral resolution and eliminate temporal fine structure information while precisely encoding the temporal envelope in each channel. This explicit encoding of temporal envelopes while simultaneously degrading spectral resolution is not possible with noise-band vocoders, which degrade transmission of temporal envelope cues (Kates 2011; Moon et al. 2014). The FAST algorithm is perceptually relevant in investigating whether envelope-based encoding may be sufficient for the transmission of binaural speech segregation cues. We analyze the patterns of interaction between semantic, voicing, and spatial cues in unprocessed and vocoder-processed conditions. In addition, the effect of spectral resolution is examined through the inclusion of narrowband and broadband vocoder reconstruction methods to simulate the effects of channel interactions in cochlear implants. The goal of the present study is to elucidate the interactions between speech segregation cues under spectrotemporal degradation to inform the development of sound encoding algorithms for hearing assistive technologies.

METHODS

Participants

Twelve normal-hearing listeners, who ranged in age between 19 and 22 years of age, were recruited for participation. All listeners were undergraduate students at the University of Southern California, native speakers of American English, and had normal audiometric pure tone thresholds (< 20 dB HL) between 0.25 and 8 kHz. Informed consent was obtained in compliance with an approved Institutional Review Board protocol from the University of Southern California Health Sciences Review Board.

Stimuli

Stimuli used in this study were sentences from the Coordinate Response Measure (CRM) database (Bolia et al. 2000). These sentences are of the form "Ready (name) go to (color) (number) now." This database includes sentences spoken by four male and four female talkers. For this experiment, the target speech sentences were always spoken by the same male talker. The masker stimuli consisted of five different conditions: sentences spoken by the same male talker, sentences spoken by the same male talker but time reversed, sentences spoken by a female talker, sentences spoken by a female talker but time reversed, and stationary speech-shaped noise. The speech-shaped noise was synthesized by taking the average of the log-magnitude spectra of all the sentences in the CRM corpus and was used to design a finite impulse response filter which was used to impose the averaged spectral envelope onto Gaussian white noise.

Signal Processing

Non-individualized head-related transfer functions were used to spatialize the target and masker stimuli to create two conditions: a co-located condition where target and masker were both located in front of the listener at 0° azimuth and a spatially separated condition where the target was located at 0° azimuth but the masker was located at 90° azimuth. After spatial processing, the stimuli were vocoder processed to create three different processing conditions: unprocessed, FAST vocoder with narrowband reconstruction filters, and FAST vocoder with broadband reconstruction filters.

The FAST vocoder is based on convolution with filterbank impulse responses to acoustically reconstruct pulsatile sequences associated with cochlear implant signal processing. For the FAST algorithm, pulsatile stimulation was synchronized to temporal

maxima of channel envelopes (Fig. 1). Pulsatile stimulation patterns were then filtered through a reconstruction filterbank to provide an acoustic representation of the determined pulsatile pattern. The average channel power from the analysis filterbank was used to scale the average channel power of the reconstruction filterbank to ensure no across channel power fluctuations were introduced by the processing. The vocoder used a 16-channel filterbank with logarithmically spaced center frequencies between 250 and 4000 Hz. Each of the individual filters was implemented as a 256th-order finite impulse response (FIR) filter constructed using the Hanning window method. The bandwidths of the filters in the analysis filterbanks were defined so that the 3 dB crossover points occur midway between center frequencies (logarithmically spaced). Note that this spacing of filters corresponds to 1/4th-octave wide filters. Local temporal maxima of the channel envelopes were used to trigger pulses while all other values were set to zero. The resulting pulsatile sequence was then filtered through the reconstruction filterbank. For the reconstruction filterbank, two different bandwidths were used for narrowband and broadband reconstruction. The narrowband reconstruction filterbank was identical to the analysis filterbank with 1/4th-octave filters. The broadband reconstruction filterbank contained filters with 1-octave wide filters. The two filterbanks were otherwise identical. The outputs of the reconstruction filterbank were scaled (using the channel power measured at the analysis filterbank) and then summed across channels to produce the vocoder output.

Procedure

Participants were seated in front of a computer in a sound-attenuating booth and asked to complete a closed-set word recognition task while listening through Sennheiser HD 280 Pro headphones. A graphical user interface allowed participants to select their responses indicating the color and number in the target sentence that they heard. SRTs were measured for 30 conditions consisting of every combination of five masker types (same-talker, same-talker time-reversed, different-talker, different-talker time-reversed, and speech-shaped noise), two spatial locations for the masker (0° and 90°), and three processing conditions (unprocessed, FAST narrowband, and FAST broadband). For each trial in the procedure, a target sentence was randomly selected from the CRM corpus that was always spoken by the same male talker. The target sentence was spatially processed and combined with the masker and then vocoder processed in accordance with the condition. The processed stimuli were presented to both ears at 65 dB SPL. Sentences were scored as correct when the subject identified both the color and number of the sentence correctly. The initial SNR of the procedure was 12 dB SNR, which was decreased/increased by 2 dB after each trial that was scored correct/incorrect. The procedure continued for eight reversals and the average of SNR values from the last four reversals was taken as the SRT for the condition. Conditions were randomized in order across participants. Participants were given breaks between conditions in order to avoid fatigue.

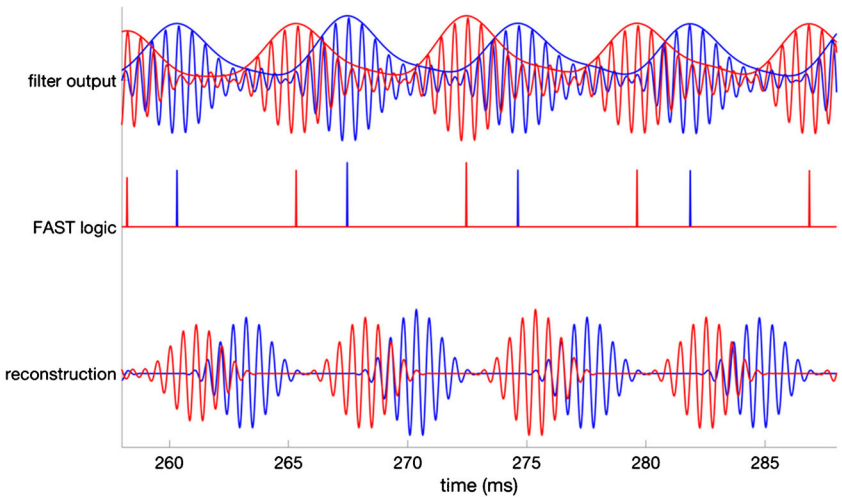

Fig. 1. Fundamental Asynchronous Stimulus Timing (FAST) algorithm. Pulses are triggered at the peaks of the envelopes in each channel. The top panel plots the acoustic output for left (blue) and right (red) channels prior to FAST processing. The middle panel shows the logic behind how FAST triggers pulses to the peaks of envelopes. The bottom panel plots the resulting output of the convolution of the impulse responses shown in the middle panel with corresponding reconstruction band-pass filters for the channel. The interaural timing difference between left and right outputs is visible in the final output

RESULTS

SRTs were measured for 30 acoustic conditions consisting of all combinations of five masker types (same-talker, same-talker time-reversed, different-talker, different-talker time-reversed, and speech-shaped noise), two spatial locations for the masker (0° and 90°), and three processing conditions (unprocessed, FAST narrowband, and FAST broadband). Measured SRTs across subjects are summarized in Fig. 2 for these 30 conditions. A three-factor within-subjects analysis of variance (ANOVA) was conducted on the measured SRTs. The factors were masker type, masker location, and vocoder processing. All main effects were significant: masker type ($F_{4,248} = 39.84$, $p < 0.01$), masker location ($F_{1,248} = 93.66$, $p < 0.01$), and vocoder processing ($F_{2,248} = 272.18$, $p < 0.01$). There was a significant interaction between masker type and vocoder processing ($F_{8,248} = 5.34$, $p < 0.01$). The interaction between masker type and masker location was also significant ($F_{4,248} = 4.72$, $p < 0.01$). The interaction between masker location and vocoder processing just missed at the .05 level ($F_{2,248} = 2.6$, $p = 0.076$) but relatively weak in comparison with the other interactions.

The preceding paragraph summarized general trends in the measured SRTs. A primary goal of this study is to elucidate how degradation of auditory resolution as produced by vocoder processing differentially affects speech reception across conditions. To that end, Fig. 3 compares SRTs in terms of the segregation cues investigated: semantic, voicing, and spatial cues. SRTs are plotted for one set of conditions versus a corresponding other set of conditions. For example, the left subplot plots SRTs for time-reversed maskers versus SRTs for the semantic maskers. Figure 4 plots the spatial release from masking as a function of the inclusion of semantic and voicing cues to illustrate how SRM is diminished with reduced spectrotemporal resolution and demonstrate the limited benefit provided when either or both of the other cues are present.

Considering semantic cues, the general trends and the effects of degrading spectral resolution through vocoder processing can be observed by tracing SRTs and the corresponding masking release for the different masker types. On one extreme, the condition with the most cues available for stream segregation is time-reversed speech (i.e., no semantic distractors), spoken by a different talker, and presented at 90° azimuth. In the left subpanels, this condition is compared to the corresponding condition but with a semantic masker such that the semantic content serves as a distractor (plotted as red, upside-down triangles). For this comparison, the average SRTs across subjects were −15.8 dB for the time-reversed

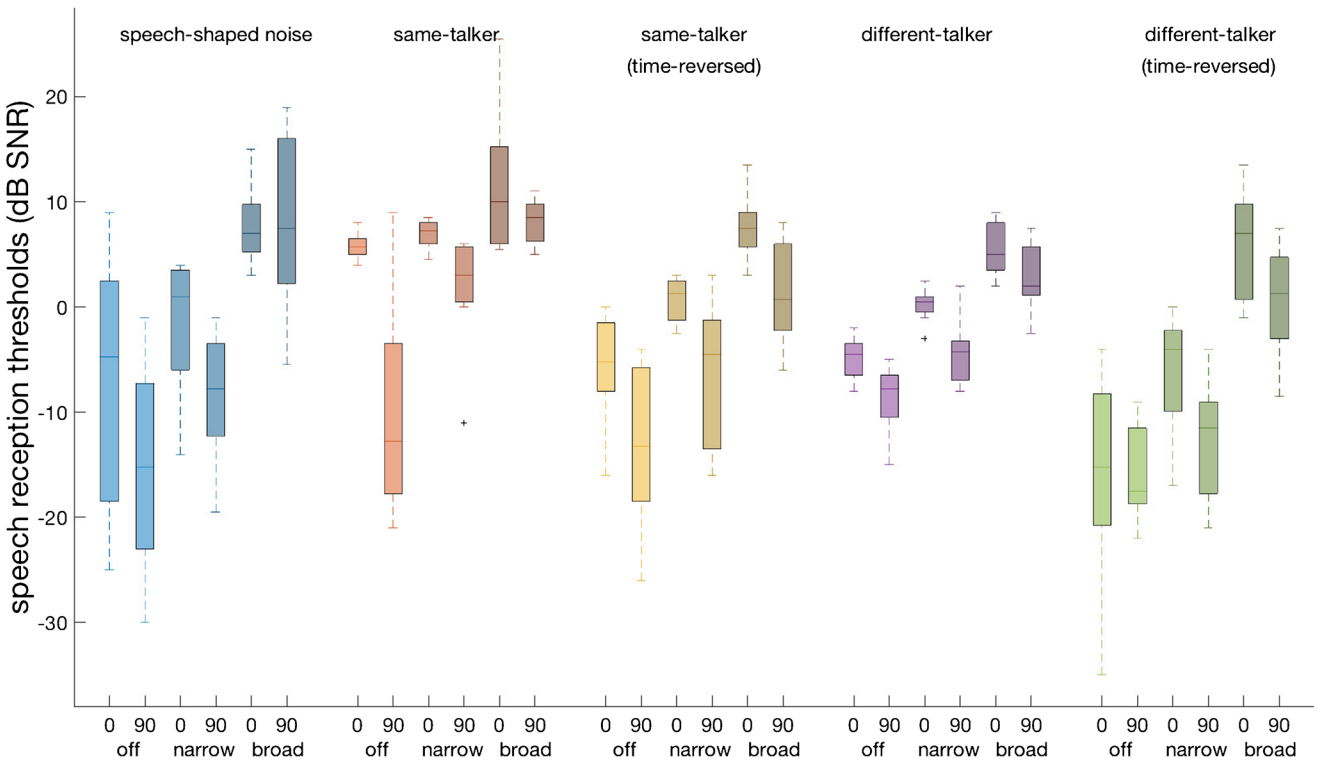

Fig. 2. Speech reception thresholds averaged across subjects for the 30 conditions tested. There was a progressive elevation in SRTs going from unprocessed to narrow and to broad reconstruction respectively, with the same-talker being the most effective masker. Off corresponds to unprocessed speech

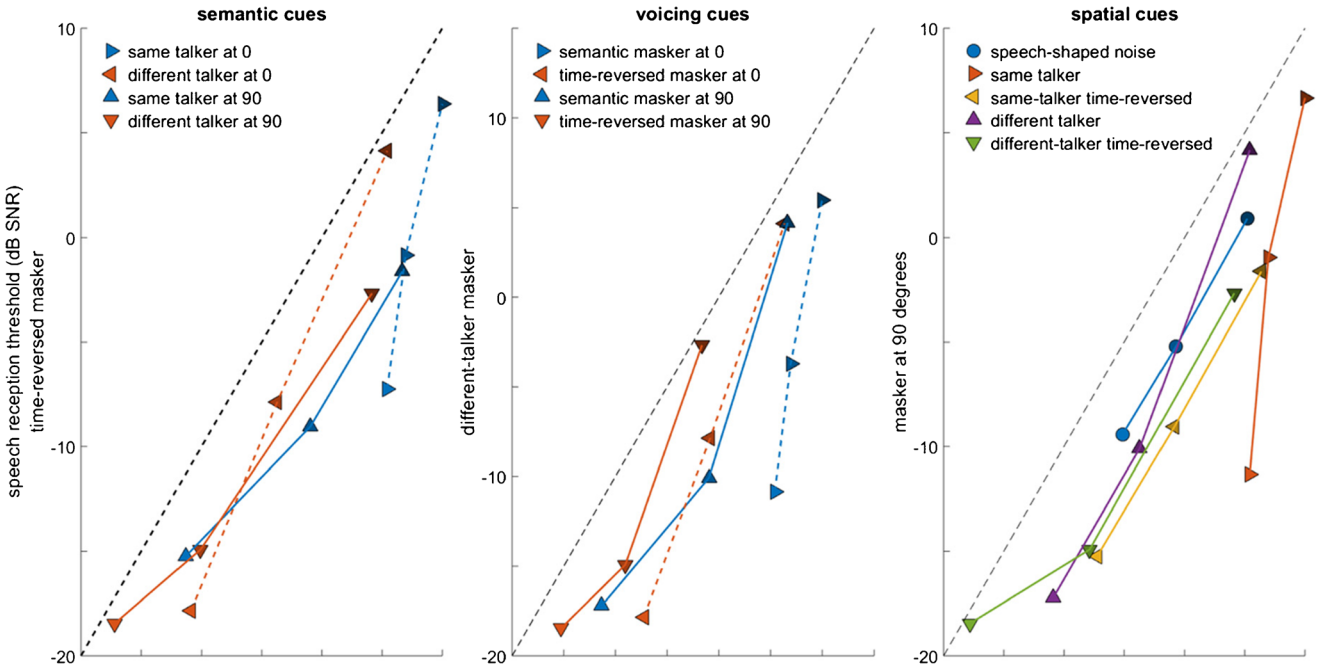

Fig. 3. Semantic (left panels), different-talker (middle panels), and spatial (right panels) release from masking plotted by comparing the SRTs of the different experimental conditions. Lines connecting symbols indicate unprocessed, narrowband, and broadband conditions going from left to right for a given masker type

masker and – 15.3 dB for the semantic masker. In other words, adding semantic distractors only decreased SRTs on average by 0.5 dB. Apparently, voicing and spatial cues for this comparison were sufficient to allow stream segregation regardless of semantic distraction. However, as spectral resolution was degraded by vocoder processing, the corresponding masking difference attributed to semantic distraction increases to 4.3 and 7.5 dB for vocoder processing with narrow and broad reconstruction, respectively. The salience of voicing and spatial cues is presumably weakened thereby increasing semantic distraction.

On the other extreme, the condition with minimal cues available for stream segregation was the semantic distractor spoken by the same talker and presented at 0° azimuth. For that condition, the target and masker sentences are simply different sentences spoken by

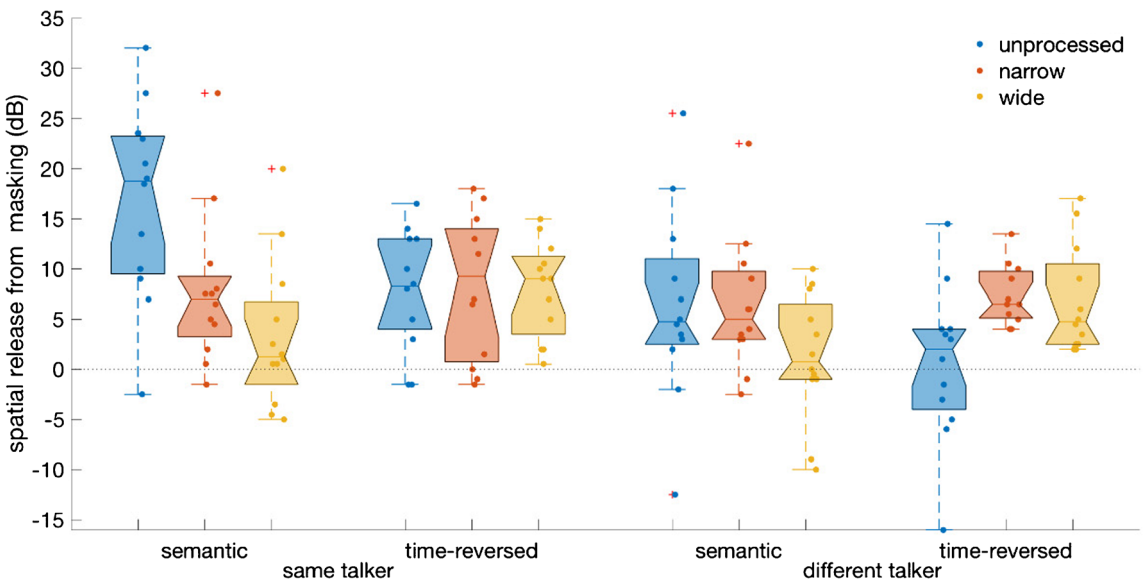

Fig. 4. Spatial release from masking (SRM) as a function of inclusion of semantic and voicing cues across the different processing conditions. SRM is reduced with spectrotemporal degradation and is limited when semantic or voicing cues are also present

the same talker. In that case, subjects require an SRT greater than 0 dB to perform the task at which point they use intensity cues to decide which sentence is the target. This condition is represented in Fig. 3 as blue, right-pointing triangles. The average SRT for subjects listening with no vocoder processing for this comparison was 5.8 dB, but when the masker was time-reversed, the average SRT improved to − 5.7 dB for an 11.5-dB masking difference associated with semantic masking. Tracing this masking difference through the different vocoder conditions, the masking difference decreased to 6.3 and 4.1 dB for vocoder processing with narrowband and broadband reconstruction, respectively. These two extreme comparisons highlight the importance of careful selection of masker conditions when investigating the effects of hearing loss: for the comparison with multiple available stream segregation cues, masking *difference* associated with semantic distraction improved with reduction of spectral resolution; but for the comparison with no available stream segregation cues, masking *difference* decreases with reduction of spectral resolution.

Similar insight can be gained from tracing other conditions in terms of overall SRTs and the corresponding masking release for voicing and spatial cues. Considering voicing cues, the extreme conditions are again the time-reversed masker presented at 90° azimuth (red, down-pointing triangles) and the semantic masker presented at 0° azimuth (blue, right-pointing triangles). For the comparison with multiple cues available for stream segregation, the addition of a voicing cue (e.g., comparing different-talker and same-talker masker conditions) provides little additional benefit for the comparison with no vocoder processing. Specifically, with no vocoder processing average SRTs for the time-reversed masker presented at 90° azimuth was − 13.1 dB for the same-talker masker, which improved to − 15.8 dB for the different-talker masker, indicating a small 2.7-dB masking difference for this condition. This masking difference improves to 6.3 dB for the conditions degraded by vocoder processing with narrow reconstruction, indicating that with the reduction of spectral resolution, subjects receive additional benefit from the voicing cue. However, for the conditions degraded by vocoder processing with broad reconstruction, the corresponding masking release derived from the voicing cue decreases to 0.8 dB indicating that this degree of vocoder processing deteriorates the encoding of the voicing cue to the extent that it is no longer salient.

Considering spatial cues, the effects of spectral resolution on the masker conditions with minimal cues for stream segregation are examined. This condition is the semantic masker spoken by the same talker. For that condition, the average SRT when the masker is presented at 0° azimuth was 5.8 dB, which improved to − 10.0 dB when the masker was presented at 90° azimuth. Thus, when no other cues are available, the masking difference associated with the spatial cue was 15.8 dB, which decreased to 5.3 and 3.6 dB for the conditions using vocoder processing with narrowband and broadband reconstruction, respectively. In a similar manner, the relative effects of spectral resolution can be examined in Fig. 3 by tracing comparisons when different combinations of cues for stream segregation are available.

To understand these interactions on a higher level, a second within-subjects ANOVA was conducted on the measured SRTs to quantify interactions between the acoustic conditions as organized by acoustic cue. Specifically, this ANOVA excludes the speech-shaped noise condition to allow the noise types to be organized in a factorial design with semantic cues (forward and time-reversed speech), voicing cues (same and different talker), spatial cues (0° and 90°), and vocoder processing as main factors. All four factors were significant: spatial location ($F_{1,206} = 87.4$, $p < 0.01$), voicing cues ($F_{1,206} = 75.26$, $p < 0.01$), semantic cues ($F_{1,206} = 190.31$, $p < 0.01$), and vocoder processing ($F_{2,206} = 226.06$, $p < 0.01$). There were significant interactions between voicing cues and masker angle ($F_{1,206} = 11.77$, $p < 0.01$) as well as between voicing cues and vocoder processing ($F_{2,206} = 11.92$, $p < 0.01$) and between voicing cues and semantic cues ($F_{1,206} = 7.05$, $p < 0.01$). The interaction between voicing cues and vocoder processing is readily observed in Fig. 3 subplot D, which shows a reduction in masking release from voicing cues in the broadband vocoder condition for all conditions tested.

Additional analyses were conducted to clarify how the reduction of spectral resolution as produced by the vocoder processing affects how acoustic cues contribute to the variance observed in measured SRTs. This perceptual cue-weighting analysis was conducted using a balanced factorial ANOVA as described in the preceding paragraph, but conducted on each level of vocoder processing to quantify the contributions of the acoustic cues. Coefficients of determination were calculated by dividing the sum of squares for the particular cue by the total sum of squares for each level of vocoder. Figure 5 plots the percentage of variance accounted for by each type of speech segregation cue. Semantic cues explained less of the variance than voicing and spatial cues. For the unprocessed condition, semantic cues account for 8.0 %, whereas voicing and spatial cues each accounted for 13.6 % of the total variance. For vocoder processing using narrow reconstruction: semantic, voicing, and spatial cues accounted for 13.9, 25.9, and 15.5 % of the total variance, respectively. Relatively then, listeners placed more weight

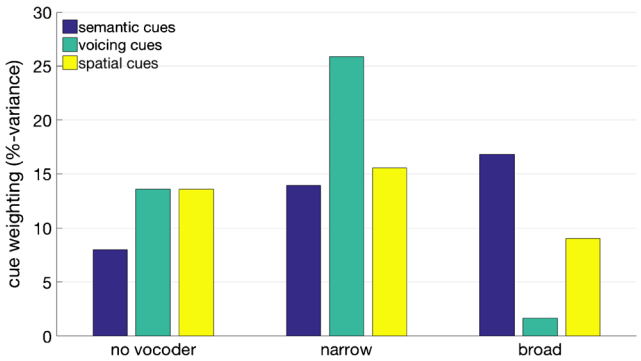

Fig. 5. Percentage of variance accounted for by each of the speech segregation cues across the three processing conditions

upon voicing cues when listening to speech degraded by this vocoder implementation. However, for vocoder processing using broad reconstruction: semantic, voicing, and spatial cues accounted for 16.8, 1.7, and 9.0 % of the total variance, respectively. This indicates a gross reduction in voicing cues available for this specific vocoder implementation using broadband reconstruction filters, presumably due to across channel smearing of temporal envelopes.

DISCUSSION

The present study investigated the interplay between different speech segregation cues that affect how listeners function in cocktail party listening situations. Our findings demonstrate a tradeoff between semantic, voicing, and spatial cues that is differentially affected by auditory resolution. For listeners with normal hearing, optimal performance is predictably achieved when background noise contains no semantic distractors, is spoken by a distinctly different talker, and comes from a different spatial location than the target speech. Less predictable are the interactions between selectively removing these cues, and less predictable still are the effects of hearing loss upon these interactions. Our discussion summarizes how these segregation cues interact, how hearing loss may affect these interactions, and how our findings are relevant to the design of signal processing for cochlear implants and hearing aids.

Interaction of Sound Segregation Cues

Listeners with normal hearing, without distortions introduced by vocoder processing, derive benefits from any combination of two segregation cues but derive little to no additional benefit from a third cue. For example, no additional benefit was derived from having masking speech located at 90° compared to having it collocated with the target speech at 0° when the masking speech was from a different talker and contained no semantic distractors. In other words, the stream segregation cues were sufficiently abundant that normal-hearing listeners received no additional benefit from the spatial cue. This finding is in line with previous studies which found that large benefits could be derived from voicing cues in reducing semantic masking, depending on the degree of similarity between target and masker (Freyman et al. 1999; Darwin and Hukin 2000; Brungart 2001a), and that these benefits are greatest when targets and maskers are spoken by the opposite gender (Brungart 2001b; Balakrishnan and Freyman 2008).

Similarly, our results indicate only a marginal benefit (~ 2 dB SNR) for normal-hearing listeners for masking speech spoken by a distinctly different talker when the masking speech has a different spatial location from the target speech and contains no semantic distractors. Again, the abundance of segregation cues led to a saturation of benefits. The large benefit derived from a difference in spatial location of target from masker is in accordance with previous findings where a perceived difference in spatial location led to large improvements in speech recognition scores when target and masker were sentences with similar semantic content (Kidd Jr. et al. 1998; Freyman et al. 1999, Arbogast et al. 2002). More recently, Carlile and Corkhill (2015) measured speech intelligibility using a spatial release from masking paradigm where they used a masker which was an unintelligible speech-like stimulus but with within-channel modulations which were similar to those in intelligible speech. They found that these stimulus components contributed to semantic masking and that the extent of this masking was dependent on the degree of spatial separation of target and masker. Their findings were novel in that they highlighted bottom-up contributions to semantic masking.

When considering semantic distractors, our results indicate no significant speech reception differences between semantically meaningful and meaningless competing speech when the competing speech has a

different spatial location and spoken by a distinctly different talker for normal-hearing listeners. It should be kept in mind that the competing speech materials used in the present study had exceptionally high levels of semantic distraction. The speech sentences shared the same basic structure, for example, the target "ready Charlie go to blue seven now" and the masker "ready Ringo go to white four now." Consequently, while the similarity across the speech corpus is an aid to our experimental design in terms of maximize derived benefits of semantic cues while controlling for voicing and spatial cues, a limitation of the present study is the translation of such controlled stimuli to the types of speech communication one would encounter in real-world listening environments. In such situations, multiple segregation cues often co-vary and the semantic content of speech streams is much less restricted.

How Hearing Loss May Affect These Interactions

The pattern of interaction between the different segregation cues changes progressively as spectrotemporal resolution is degraded using vocoder processing. The vocoder processing with narrowband and broadband reconstruction used in our study reduced spectral resolution, simulating the degree of current spread in cochlear implant devices. By progressively reducing spectral resolution, there is an observable increase in the reliance on multiple segregation cues, and for speech reception benefits to be derived from having all three cues available, as indicated by the greater release from semantic masking. This is verified by the increase in the computed perceptual weights (see Fig. 4). In addition, the degree of semantic masking was reduced when introducing a difference in voicing cues by switching the male masker to a female masker and when the masker was moved away from the target. Our findings are in accordance with previous findings that spatial separation can lead to reduced semantic masking (Kidd Jr. et al. 1998; Freyman et al. 1999; Arbogast et al. 2002; Carlile and Corkhill 2015) and that differences in voicing characteristics can lead to reduced semantic masking (Freyman et al. 1999; Darwin and Hukin 2000; Brungart 2001a), particularly when targets and maskers are spoken by talkers with opposing gender (Brungart 2001b; Balakrishnan and Freyman 2008).

Several previous studies have investigated semantic masking in cochlear implant users and in normal-hearing listeners attending to spectrally degraded speech (Qin and Oxenham 2003; Stickney et al. 2004). For normal-hearing listeners, it has been found that the inherent temporal envelope fluctuations in noise are largely responsible for the masking of speech (Dubbelboer and Houtgast 2008; Jorgensen et al. 2013; Stone et al. 2011; Stone et al. 2012; Stone and Moore 2014). Given that cochlear implants typically do not convey temporal fine structure, users are particularly susceptible to the effects of masking by noise. Whereas normal-hearing listeners can "glimpse" temporal peaks of the target within the valleys of the masker signal (Cooke 2006), cochlear implant users do not receive the same release from masking resulting from an amplitude modulated masker (Oxenham and Kreft 2014; Goldsworthy 2015). This is surprisingly in contrast to the finding that normal-hearing listeners are even able to show glimpsing benefits even when listening through vocoder simulations to modulated maskers. Taken together, these findings indicate that modulation energy is a more significant predictor for masking for normal-hearing listeners, whereas overall noise energy is a more significant predictor of masking for cochlear implant users. This fundamental difference between the normal hearing and cochlear implant users could be partially attributed to the indirect effects of poor spectral resolution which lead to smoothing of temporal envelopes which is further exacerbated by the lack of temporal fine structure cues in existing cochlear implant signal processing strategies. In acoustic hearing loss, spectral smearing occurs prior to the extraction of temporal envelopes by inner hair cells. This is not the case with electric hearing, where a reduction in spectral resolution leads to the flattening of the modulation spectrum, thereby limiting the benefits derived from temporal fine structure. Given that sensitivity to temporal fine structure is a hallmark of normal hearing, next-generation cochlear implant signal processing strategies should attempt to convey temporal fine structure in a salient manner.

Considerations for the Design of Cochlear Implants and Hearing Aid Signal Processing

The results from our study indicate that the factors which affect speech segregation can interact in interesting ways. If future studies are focused on real-world outcomes for CI and hearing aid users, then their methods should consider how multiple cues can be used to derive an advantage amongst competing sound sources by using realistic stimuli (i.e., containing multiple segregation cues). However, for studies that are focused on a new algorithm that attempts to improve either the salience of voicing or spatial cues (or both), it is also important to probe the segregation cues available to ascertain how much benefit is derived from each individual cue by means of careful experimental design. Our strategy in the present study aimed to strike a balance between both

approaches in evaluating the effectiveness of the FAST algorithm in a cocktail party environment to determine how the different segregation cues interact in real-world listening situations. The findings from our study indicate that the FAST algorithm can preserve envelope details which facilitate the transmission of voicing and spatial cues. However, such cues are most robust with sufficient spectral resolution (such as in our narrow reconstruction condition), highlighting the importance of reducing channel interactions from neighboring electrodes in CI devices. Our findings support previous research by Swaminathan et al. (2016) demonstrating the importance of low-frequency fine timing information to SRM benefits for NH listeners attending to sentences processed through noise vocoders with either uncorrelated or correlated noise carriers by selectively retaining TFS cues with the correlated noise carrier and eliminating them in the uncorrelated noise carrier. While the FAST algorithm does not preserve TFS information, we have found that preserving envelope details can yield similar benefits to speech unmasking.

Previous studies have focused on more effectively encoding TFS information or envelope details into CI signal processing to enhance spatial cues. Churchill et al. (2014) found that bilateral CI users exhibited sensitivity to TFS information when bilateral stimulation was synchronized and pitch-matched across both ears by measuring ITD discrimination and lateralization of speech using low-rate stimulation provided on multiple electrodes. Zirn et al. (2016) found that the TFS encoding strategy FS4 improved interaural phase discrimination compared to the high definition continuous interleaved sampling (HDCIS) strategy but that this improvement did not translate to improvements in binaural intelligibility level differences with a speech-shaped noise masker. Part of this lack of improvement with FS4 might be related to the fact that fine timing information is only transmitted on the four most apical electrodes using this strategy. In contrast, the FAST algorithm encodes the periodicity of amplitude envelopes across all channels. This may explain why the present study was able to find significant SRM benefits.

Previous studies have also focused on encoding TFS information to enhance voicing cues. Vandali et al. (2016) which found that pitch coding with the OPAL strategy led to improved lexical tone recognition and speech perception in noise for Mandarin speakers. Ping et al. (2017) similarly found a small improvement lexical tone recognition and speech in quiet performance with the C-tone strategy. However, it is worth mentioning that future studies that evaluate the effectiveness of TFS encoding algorithms should incorporate experimental designs which avoid ceiling effects which are encountered in experiments which test speech reception in quiet or which use speech-shaped noise maskers. These types of experiments may underestimate the effectiveness of TFS encoding strategies which are particularly well-suited for improving speech segregation competing talkers, a scenario which is typically more difficult for hearing impaired listeners. The choice of stimuli and experimental design of the present study demonstrated the effectiveness of the FAST algorithm in preserving speech segregation cues in such difficult types of situations.

CONCLUSION

Speech reception in the presence of background noise was evaluated using a number of different maskers processed through a vocoder based on the FAST signal processing algorithm. The results of the present study indicate that speech segregation cues can be preserved with sufficient spectral resolution. These findings have important implications for improving signal processing in cochlear implants and hearing assistive technologies which must prioritize the preservation of spectrotemporal cues through the utilization of fine timing encoding strategies and reduce undesirable channel interactions in order to facilitate speech segregation abilities in hearing impaired individuals.

REFERENCES

ARBOGAST TL, MASON CR, KIDD G (2002) The effect of spatial separation on informational and energetic masking of speech. J Acoust Soc Am 112:2086–2098

BALAKRISHNAN U, FREYMAN RL (2008) Speech detection in spatial and non-spatial speech maskers. J Acoust Soc Am 123:2680–2691

BAŞKENT D, GAUDRAIN E (2016) Musician advantage for speech-on-speech perception. J Acoust Soc Am 139:EL51–EL56

BLAUERT J (1997) Spatial hearing: the psychophysics of human sound localization. MIT Press, Cambridge

BOLIA RS ET AL (2000) A speech corpus for multitalker communications research. J Acoust Soc Am 107:1065–1066

BROKX JPL, NOOTEBOOM SG (1982) Intonation and the perceptual separation of simultaneous voices. J Phon 10:23–36

BRONKHORST AW (2000) The cocktail party phenomenon: a review of research on speech intelligibility in multiple-talker conditions. Act Acust U Acust 86(1):117–128

BRUNGART DS (2001A) Evaluation of speech intelligibility with the coordinate response measure. J Acoust Soc Am 109:2276–2279

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

BRUNGART D, SIMPSON B (2002) Within-ear and across-ear interference in a cocktail-party listening task. J Acoust Soc Am 112:2985–2995

Carlile S, Corkhill C (2015) Selective spatial attention modulates bottom-up informational masking of speech. Sci Rep 5(8662):1–7

Cherry EC (1953) Some experiments on the recognition of speech, with one and with two ears. J Acoust Soc Am 25(5):975–979

Churchill T et al (2014) Spatial hearing benefits demonstrated with presentation of acoustic temporal fine structure cues in bilateral cochlear implant listeners. J Acoust Soc Am 136:1246–1256

Cooke M (2006) A glimpsing model of speech perception in noise. J Acoust Soc Am 119(3):1562–1573

Darwin C, Hukin R (2000) Effectiveness of spatial cues, prosody, and talker characteristics in selective attention. J Acoust Soc Am 107:970–977

Darwin C, Brungart D, Simpson B (2003) Effects of fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers. J Acoust Soc Am 114:2913–2922

Dorman MF et al (1998) The recognition of sentences in noise by normal-hearing listeners using simulations of cochlear-implant signal processors with 6–20 channels. J Acoust Soc Am 104:3583–3585

Dubbelboer F, Houtgast T (2008) The concept of signal-to-noise ratio in the modulation domain and speech intelligibility. J Acoust Soc Am 124:3937–3946

Durlach NI, Mason CR, Kidd Jr. G, Arbogast TL, Colburn HS, Shinn-Cunningham B (2003) Note on informational masking. J Acoust Soc Am in press

Fitch WT, Giedd J (1999) Morphology and development of the human vocal tract: a study using magnetic resonance imaging. J Acoust Soc Am 106:1511–1522

Freyman RL et al (1999) The role of perceived spatial separation in the unmasking of speech. J Acoust Soc Am 106(6):3578–3588

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

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

Fuller CD et al (2014) Gender categorization is abnormal in cochlear implant users. J Assoc Res Otolaryngol 15:1037–1048

Gallun FJ, Mason CR, Kidd G (2005) Binaural release from informational masking in a speech recognition task. J Acoust Soc Am 118:1614–1625

Gaudrain E, Başkent D (2015) Factors limiting vocal-tract length discrimination in cochlear implant simulations. J Acoust Soc Am 137:1298–1308

Goldsworthy R (2015) Correlations between pitch and phoneme perception in cochlear implant users and their normal hearing peers. J Assoc Res Otolaryngol 16(6):797–809

Hillenbrand JM, Clark MJ (2009) The role of F0 and formant frequencies in distinguishing the voices of men and women. Atten Percept Psychophys 71(5), pp. 16

Hirsh IJ (1948) The influence of interaural phase on interaural summation and inhibition. J Acoust Soc Am 20:536–544

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

van Hoesel RJ, Tyler RS (2003) Speech perception, localization, and lateralization with bilateral cochlear implants. J Acoust Soc Am 113:1617–1630

Jorgensen S, Ewert SD, Dau T (2013) A multi-resolution envelope-power based model for speech intelligibility. J Acoust Soc Am 134(1):436–446

Kan A, Litovsky R (2015) Binaural hearing with electrical stimulation. Hear Res 322:127–137. https://doi.org/10.1016/j.heares.2014.08.005

Kates JM (2011) Spectro-temporal envelope changes caused by temporal fine structure modification. J Acoust Soc Am 129(6):3981–3990

Kidd G Jr et al. (2007) Informational masking. Springer handbook of auditory research 29: auditory perception of sound sources, edited by W. Yost (Springer, New York), pp. 143–190

Kidd G Jr et al (1998) Release from masking due to spatial separation of sources in the identification of nonspeech auditory patterns. J Acoust Soc Am 104:422–431

Kidd G Jr, Mason C, Gallun F (2005) Combining energetic and informational masking for speech identification. J Acoust Soc Am 118:982–992

Leek M, Brown ME, Dorman MF (1991) Informational masking and auditory attention. Percept Psychophys 50:205–214

Li T, Fu QJ (2011) Voice gender discrimination provides a measure of more than pitch-related perception in cochlear implant users. Int J Audiol 50:498–502

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

Moon IJ, Won J-H, Park M-H, Ives DT, Nie K, Heinz MG, Lorenzi C, Rubinstein JT (2014) Optimal combination of neural temporal envelope and fine structure cues to explain speech identification in background noise. J Neurosci 34:12145–12154

Moore BCJ (2012) An introduction to the psychology of hearing. 6. The Netherlands, Brill

Oxenham AJ, Kreft HA (2014) Speech perception in tones and noise via cochlear implants reveals influence of spectral resolution on temporal processing. Trends Hear 18:1–14

Ping L et al (2017) Implementation and preliminary evaluation of 'C-tone': a novel algorithm to improve lexical tone recognition in Mandarin-speaking cochlear implant users. Cochlear Implants Int 18(5):240–249

Poissant SF, Whitmal NA III, Freyman RL (2006) Effects of reverberation and masking on speech intelligibility in cochlear implant simulations. J Acoust Soc Am 119:1606–1615

Pollack I (1975) Auditory informational masking. J Acoust Soc Am 57:S5

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

Shannon R et al (1995) Speech recognition with primarily temporal cues. Science 270:303–304

Skuk VG, Schweinberger SR (2014) Influences of fundamental frequency, formant frequencies, aperiodicity and spectrum level on the perception of voice gender. J Speech Lang Hear Res 57(1):285–296

Smith DR, Patterson RD (2005) The interaction of glottal-pulse rate and vocal-tract length in judgements of speaker size, sex, and age. J Acoust Soc Am 118:3177–3186

Smith ZM et al (2014) Hearing better with interaural time differences and bilateral cochlear implants. J Acoust Soc Am 135(4):2190–2191

Stickney G et al (2004) Cochlear implant speech recognition with speech maskers. J Acoust Soc Am 116(2):1081–1091

Stone MA, Moore BCJ (2014) On the near non-existence of "pure" energetic masking release for speech. J Acoust Soc Am 135(4):1967–1977

Stone MA et al (2011) The importance for speech intelligibility of random fluctuations in "steady" background noise. J Acoust Soc Am 130(5):2874–2881

Stone MA, Fullgrabe C, Moore BCJ (2012) Notionally steady background noise acts primarily as a modulation masker of speech. J Acoust Soc Am 132(1):317–326

Swaminathan J et al (2016) Role of binaural temporal fine structure and envelope cues in cocktail-party listening. J Neurosci 36(31):8250–8257

VANDALI AE ET AL (2005) Pitch ranking ability of cochlear implant recipients: a comparison of sound-processing strategies. J Acoust Soc Am 117(5):3126–3138

VANDALI AE, DAWSON PW, ARORA K (2016) Results using the OPAL strategy in Mandarin speaking cochlear implant recipients. Int J Audiol Jun 22, pp. 1–12

WATSON CS (2005) Some comments on informational masking. Acta Acoust 91:502–512

YOST B (2006) Informational masking: what is it?, in paper presented at the 2006 Computational and Systems Neuroscience (Cosyne) meeting

ZIRN S ET AL (2016) Perception of interaural phase differences with envelope and fine structure coding strategies in bilateral cochlear implant users. Trends Hear 20:2331216516665608

ZUREK PM (1993) Binaural advantages and directional effects in speech intelligibility. Acoustical factors affecting hearing aid performance, edited by G.A. Studebaker & I. Hochberg, pp. 255-275