Low-frequency fine-structure cues allow for the online use of lexical stress during spoken-word recognition in spectrally degraded speech

Ying-Yee Kong^{a)}

Department of Communication Sciences & Disorders, Northeastern University, 226 Forsyth Building,
360 Huntington Avenue, Boston, Massachusetts 02115, USA

Alexandra Jesse

Department of Psychological and Brain Sciences, University of Massachusetts, 135 Hicks Way, Amherst,
Massachusetts 01003, USA

(Received 23 June 2016; revised 22 November 2016; accepted 7 December 2016; published online
19 January 2017)

English listeners use suprasegmental cues to lexical stress during spoken-word recognition. Prosodic cues are, however, less salient in spectrally degraded speech, as provided by cochlear implants. The present study examined how spectral degradation with and without low-frequency fine-structure information affects normal-hearing listeners' ability to benefit from suprasegmental cues to lexical stress in online spoken-word recognition. To simulate electric hearing, an eight-channel vocoder spectrally degraded the stimuli while preserving temporal envelope information. Additional lowpass-filtered speech was presented to the opposite ear to simulate bimodal hearing. Using a visual world paradigm, listeners' eye fixations to four printed words (target, competitor, two distractors) were tracked, while hearing a word. The target and competitor overlapped segmentally in their first two syllables but mismatched suprasegmentally in their first syllables, as the initial syllable received primary stress in one word and secondary stress in the other (e.g., “'admiral,” “,admi'rati'on”). In the vocoder-only condition, listeners were unable to use lexical stress to recognize targets before segmental information disambiguated them from competitors. With additional lowpass-filtered speech, however, listeners efficiently processed prosodic information to speed up online word recognition. Low-frequency fine-structure cues in simulated bimodal hearing allowed listeners to benefit from suprasegmental cues to lexical stress during word recognition. © 2017 Acoustical Society of America.

[<http://dx.doi.org/10.1121/1.4972569>]

[TB]

Pages: 373–382

I. INTRODUCTION

As speech unfolds, listeners continuously update their evaluation of how well the incoming information matches the representations of words stored in their mental lexicon (e.g., Allopenna *et al.*, 1998; Luce and Pisoni, 1998; Salverda *et al.*, 2003; Zwitserlood, 1989). Although spoken-word recognition strongly relies on segmental information, suprasegmental cues, that is, acoustic changes in duration, fundamental frequency (F_0), and intensity, also contribute. Listeners of variable-stress languages, such as English and Dutch, use suprasegmental cues to lexical stress for word recognition (e.g., Cooper *et al.*, 2002; Friedrich *et al.*, 2004; Mattys, 2000; Jesse and McQueen, 2014; Reinisch *et al.*, 2010; Soto-Faraco *et al.*, 2001; Sulpizio and McQueen, 2012; van Donselaar *et al.*, 2005). For example, in a cross-modal fragment priming study in English, Cooper *et al.* (2002) reported that after having heard fragment primes (e.g., 'mu, 'admi) with primary lexical stress,¹ listeners' lexical decisions to printed target words were faster when these words matched the fragment primes both segmentally and

suprasegmentally (e.g., “'music” and “'admiral”) compared to when they only matched segmentally (e.g., “mu'seum” and “,admi'rati'on”).

Distinguishing different degrees of lexical stress is useful for word recognition in these languages because it reduces the number of embedded words and allows words to become perceptually unique earlier (Cutler and Pasveer, 2006; van Heuven and Hagman, 1988). Recent work on English and Dutch provided evidence that suprasegmental information about lexical stress can indeed speed up the recognition of target words before later segmental cues resolve lexical competition (Jesse *et al.*, 2017; Reinisch *et al.*, 2010). Using eye tracking with a visual world paradigm, the time course of spoken-word recognition and that of lexical competition were tracked in these studies, as listeners' spontaneous fixations to a referent shown on a computer screen are indicative of the momentary degree of support for that referent's label as a lexical candidate (Allopenna *et al.*, 1998). In these two studies on lexical stress (Jesse *et al.*, in press; Reinisch *et al.*, 2010), listeners heard a target word while seeing its printed version on a display, along with a competitor word and two unrelated distractor words. Target and competitor words matched segmentally in their first two

^{a)}Electronic mail: yykong@neu.edu

syllables but differed in their degree of lexical stress on the first syllable (e.g., 'admiral vs 'admi'ration). After having heard a target word's first syllable with primary stress (e.g., 'ad), young normal-hearing English and Dutch listeners fixated more on the target word ('admiral) than on its stress-mismatching competitor (,admi'ration). Normal-hearing listeners thus utilize suprasegmental cues to lexical stress as soon as these cues become available to recognize the target word. This finding strongly indicates that prosodic information can alter the time course of spoken-word recognition (e.g., Ito and Speer, 2008; Salverda *et al.*, 2003, 2007), and that prosodic information is therefore important for keeping up with the high demands of fast and efficient word recognition (Dahan *et al.*, 2001; Norris and McQueen, 2008; Tanenhaus *et al.*, 1995).

The perception of prosody is, however, significantly affected by the quality of the incoming speech signal. One extreme situation of speech perception is listening via a cochlear implant (CI), where the speech signal is spectrally degraded. The current signal processing strategies for CIs are designed to preserve temporal envelope cues (i.e., amplitude changes over time) for speech recognition, while discarding the spectral and temporal fine-structure cues [i.e., the carrier defined by the Hilbert transform (Hilbert, 1912)] that are important for pitch perception (Smith *et al.*, 2002; Zeng *et al.*, 2005). The reduced frequency selectivity in CI listening is also due to other factors, including limited neural survival, current spread, and channel interaction (e.g., Bierer *et al.*, 2011; Fu and Nogaki, 2005; Hughes, 2008; Jones *et al.*, 2013; Pfungst *et al.*, 2011). The spectral details provided by CIs resemble that of 4–8 channels of vocoded speech to normal-hearing listeners (Fishman *et al.*, 1997; Friesen *et al.*, 2001; Stickney *et al.*, 2004). The available information can be sufficient for phoneme, word, and sentence recognition in quiet (e.g., >80% recognition accuracy for simple sentences) for normal-hearing listeners when tested with vocoded speech as well as for multichannel CI recipients (e.g., Shannon *et al.*, 1995; Fishman *et al.*, 1997).

The reduced spectral resolution, compounded with the lack of temporal fine-structure cues in the processing strategies, poses a challenge for using pitch cues for speech recognition in noise (e.g., Stickney *et al.*, 2004; Cullington and Zeng, 2008) and music perception (e.g., Kong *et al.*, 2004; McDermott, 2004) in CI listening. Also, the perception of prosody is generally poorer in CI users compared to normal-hearing listeners (e.g., Holt and McDermott, 2013; Holt *et al.*, 2016; Marx *et al.*, 2015; Meister *et al.*, 2009, 2011; Morris *et al.*, 2013; Peng *et al.*, 2008, 2009). Although the temporal envelope of speech carries information related to duration, pauses, and amplitude, the reduced spectral and temporal resolution in CI listening (e.g., Oxenham and Kreft, 2014) negatively impact the perception of prosody. For example, Morris *et al.* (2013) reported for Swedish that the discrimination of vowel length and of word stress (e.g., "kaffe" ['kafé] vs "kafé" [ka'fe:]) in a minimal pair task was significantly poorer in CI users compared to normal-hearing listeners. Most relevant for the perception of prosody is that the spectrally degraded speech significantly impairs the perception of voice pitch, such as used for the perception of

sentence-level stress and the discrimination of statements vs questions based on intonation (e.g., English: Chatterjee and Peng, 2008; See *et al.*, 2013; French: Marx *et al.*, 2015; German: Meister *et al.*, 2009). The severe deficit in pitch perception with electric hearing has been well documented. The upper limit of temporal pitch for the majority of CI users is about 300 Hz, much lower than that of normal-hearing listeners (e.g., Carlyon *et al.*, 2010; Kong *et al.*, 2009; Zeng, 2002). Recipients of a CI exhibit poor recognition of simple melodies (e.g., Dorman *et al.*, 2008; Galvin *et al.*, 2009; Kong *et al.*, 2004; Singh *et al.*, 2009; see McDermott, 2004 for review), poor lexical tone discrimination (e.g., Ciocca *et al.*, 2002; Lee *et al.*, 2002; Peng *et al.*, 2004; Wei *et al.*, 2004), and abnormal pitch perception (Zeng *et al.*, 2014).

One goal of the present study was therefore to investigate the effect of spectral degradation on listeners' ability to effectively use suprasegmental information about lexical stress during online spoken-word recognition. A noise-channel vocoder spectrally degraded the speech stimuli while largely preserving the temporal envelope cues. We hypothesized that the combined deficits in spectral resolution and temporal fine-structure encoding after channel vocoding negatively impact, or even prevent, the online use of suprasegmental cues to lexical stress for word recognition.

A second goal of this study was to examine the contribution of low-frequency fine-structure cues to the perception of lexical stress as these cues are often preserved in residual hearing in listeners with combined electric-acoustic stimulation (EAS). Complex pitch perception is achieved by neural phase locking to the resolved low-numbered harmonics (e.g., Bernstein and Oxenham, 2003; Dai, 2000; Houtsma and Smurzynski, 1990; Plomp, 1967; Ritsma, 1967). Available in the fine structure of the acoustic signal, the additional low-frequency harmonic and periodicity cues in EAS [or in simulated EAS where vocoded speech is combined with low-pass (LP)-filtered speech] have been shown to improve recognition performance of musical melodies and lexical tones compared to electric hearing (or vocoded speech) alone (e.g., Kong *et al.*, 2005; Li *et al.*, 2014). Furthermore, Spitzer *et al.* (2009) found that, compared to users of a unilateral CI, EAS users were better able to take advantage of *F0* contour information associated with metrical stress to reduce the number of errors made in segmenting continuous speech into words. Stimuli were phrases consisting of six syllables (three to five words) that alternated in stress (iambic or trochaic). These phrases were either presented with a flat *F0* contour or with their natural *F0* contour preserved. While unilateral CI users, as a group, were able to use syllable duration, dispersion of vowel formant frequencies, and amplitude envelope modulation as cues to metrical stress to guide speech segmentation, they did not show a significant difference in the number of stress-based segmentation errors they made between the flat *F0* and *F0* contour conditions. In contrast, EAS users significantly benefited from the additional *F0* cues in this word segmentation task.

Speech stimuli used in previous studies on word-level stress in CI and EAS listening contained both segmental (e.g., vowel reduction) and suprasegmental cues (Morris *et al.*, 2013; Spitzer *et al.*, 2009). It is unclear how the

perception of lexical stress would be affected if only suprasegmental cues were available to CI users (or in simulated CI listening) and the extent to which the enhanced F_0 representation in EAS listening (or in simulated EAS listening) improves the perception of lexical stress in English. Given that low-frequency residual hearing improves pitch perception in EAS compared to CI only listening (e.g., Dorman *et al.*, 2008; Kong *et al.*, 2005; Li *et al.*, 2014), we expect that listeners' ability to utilize suprasegmental cues, particularly pitch cues, for lexical stress perception would be preserved when supplemental low-frequency harmonicity cues are provided in conjunction with vocoded speech. More importantly, it is also yet to be determined whether or not lexical stress information could be exploited during word recognition in simulated EAS listening. The timely utilization of lexical stress can reduce the uncertainty in the speech signal and can allow listeners to keep up with the high demands of the fast processing of spoken words. Although the harmonic and periodicity cues are available in the low frequencies for pitch encoding, the altered signal quality may prevent the efficient and immediate use of prosodic cues.

To test these questions, the current study employed a visual world paradigm using printed words to examine how suprasegmental cues to lexical stress affect the time course of spoken-word recognition in English for degraded speech. Using the paradigm of Jesse *et al.* (2017), listeners' eye fixations on four printed words were tracked while hearing one of the words in a neutral carrier sentence. Two of these words were segmentally identical in their initial two syllables but had either primary stress or secondary stress on the first syllable (*'admi'ration* [ˌædmə'reɪʃn] vs *'admiral* [ˈædmərəl]). On critical trials, these words served as a target and stress-mismatching phonological competitor. Speech stimuli were vocoded to simulate CI listening, and LP-filtered speech was added to the opposite ear of the vocoded speech to simulate bimodal listening (vocoder + LP condition). If listeners can efficiently process suprasegmental cues to lexical stress in spectrally degraded speech to benefit online spoken-word recognition, then we should replicate our previous results (Jesse *et al.*, 2017), showing that listeners fixate more on targets with initial primary stress than their stress competitors while hearing the second syllable; that is, before hearing segmental information that distinguishes these words. However, we predict that due to the discarded temporal fine structure in vocoded speech, listeners will be less able, or even unable, to use suprasegmental cues to lexical stress during spoken-word recognition. We further predict that listeners can use suprasegmental cues to lexical stress during online word recognition, when the spectrally degraded speech is accompanied by LP-filtered speech information, simulating bimodal hearing.

II. METHODS

A. Participants

Twenty-four college students (mean age 21 yrs; 8 men) from Northeastern University participated in the study. All participants were monolingual native speakers of American

English, with no reported speech, language, hearing, or attention deficits. All had normal or corrected-to-normal vision. Participants provided informed consent and were compensated for their participation.

B. Materials

Stimuli materials were taken from our previous study showing the online use of lexical stress in spoken-word recognition by normal-hearing English listeners (Jesse *et al.*, 2017). These materials consisted of 24 test sets, in addition to eight sets created for filler trials and six sets for practice trials. Each set contained two critical and two non-critical words, that were all semantically and morphologically unrelated to each other. For example, in one set, the four words were *'admiral*, *'admi'ration*, *con'verter*, and *con'vergence*. The first two syllables of the critical pairs (e.g., *'admiral*–*'admi'ration*) were segmentally identical, but differed suprasegmentally in their stress pattern. One of the words always had primary stress on the first syllable and the other word had secondary stress on the first syllable. The non-critical word pairs overlapped both segmentally and suprasegmentally in the first two syllables. The non-critical word pairs (e.g., *con'verter*–*con'vergence*) had primary lexical stress on the first, second, or third syllable. The critical and non-critical word pairs were matched in their spoken word frequency ($t(47) = 0.31$, $p > 0.05$; Corpus of Contemporary American English; Davies, 2008). In addition, critical words were matched in frequency across stress type ($t(23) = 0.01$, $p > 0.05$). The critical words were three- or four-syllable words (with the exception of one five-syllabic word) and the non-critical words were three- to five-syllable words (all critical words are available in the Appendix of Jesse *et al.*, 2017).

All materials were recorded by a female native speaker of American English at a 44.1 kHz sampling rate in a sound-proof booth. The words were produced in the carrier sentence "Click on the word ____." The recorded sentences were normalized to have the same root-mean-squared (RMS) amplitude. Acoustic analyses on the first syllable of the critical word pairs in the unprocessed materials² showed a significantly larger pitch excursion when the syllables carried primary ($M = 18$ semitones per second) than secondary ($M = 12$ semitones per second) lexical stress ($t(22) = 3.86$, $p < 0.001$). Previous reports have shown that F_0 is a salient cue for lexical stress perception in English (e.g., Beckman, 1986; Fry, 1958; Lehiste, 1970; Mattys, 2000). Primary-stressed syllables also had a lower mean RMS amplitude than secondary-stressed syllables by 1.4 dB ($t(22) = -3.47$, $p < 0.01$), but did not differ from each other in duration or spectral tilt. It is important to note that the first and second formant frequencies were not significantly different on the first and second syllables of the critical word pairs ($p < 0.05$), further ruling out that any segmental cues could potentially support distinguishing stress pairs during the first two syllables.

C. Signal processing

Recorded stimuli were subjected to two types of signal processing: noise-channel vocoding and LP filtering. Noise-channel vocoding was used to simulate (1) the preservation

of temporal envelope cues and (2) the spectral degradation and lack of temporal fine-structure cues in CI processing. Similar to the processing described by [Shannon et al. \(1995\)](#), broadband speech (141–7000 Hz) was first processed through a pre-emphasis filter and then bandpass filtered into 8 logarithmically-spaced frequency bands ([Greenwood, 1990](#)). The amplitude envelope of the signal from each band was extracted using the Hilbert transform, followed by LP filtering with a 300-Hz cutoff frequency. The envelope within each band was modulated with white noise, which was then filtered by the same bandpass filter that had been used to generate the frequency band. All bands were then summed to produce the final vocoded signal.

To create the LP-filtered supplementary speech for the vocoder + LP condition, LP filtering was performed on the original stimuli using a tenth-order Butterworth filter with a roll-off slope of 60 dB/octave and a cutoff frequency of 500 Hz. These LP parameters approximated a common sloping hearing loss above 500 Hz in the non-implanted ear of CI users. Using these LP filtering parameters with a group of young normal-hearing listeners, [Kong and Braida \(2011\)](#) showed that the amount of information transmission for consonant and vowel recognition in quiet was similar to that obtained from a group of real bimodal users (CI in one ear and residual hearing in the other).

D. Procedures

The study consisted of two 2-h sessions, taking place on two different days, about 1–2 weeks apart. Each participant was tested in two listening conditions—vocoder (simulation of electric hearing) and vocoder + LP (simulation of bimodal hearing). Participants were tested in a sound-proof booth, seated 60 cm away from a 1024 × 768 (17 in., diagonal) computer screen (60 Hz refresh rate).

All stimuli were presented at a comfortable listening level. Vocoded speech was presented to one ear over Sennheiser (Wedemark, Germany) HD 600 headphones at an RMS level of 68 dB A. In the vocoder + LP listening condition, LP speech was presented additionally to the opposite ear of the vocoded speech. Half of the participants received the vocoded stimuli in the left ear and LP stimuli in the right ear; the other half received the stimuli in the opposite configuration. The presentation level for the LP speech was determined using a loudness balancing procedure, in which the level of LP speech in one ear was adjusted to have the same perceived loudness as the vocoded speech in the other ear. Loudness was balanced across ears to maximize the potential bimodal benefit ([Dorman et al., 2014](#)) and to avoid listeners' bias toward the louder signal (e.g., they may pay greater attention to the perceived louder vocoded speech on one side in the vocoder + LP condition). A single word was used for this loudness-balancing task. First, the vocoded version of that word was presented alone five times at a fixed level of 68 dB A. The LP version of that word was then presented alone in the other ear five times, at the same level as the vocoded stimulus. Listeners were asked to indicate how the perceived loudness of the LP speech compared to that of the vocoded speech. We then used a bracketing technique to

increase the level of the LP speech in steps of 1 dB, and repeated the comparison between vocoded and LP speech at each new LP level. Once the LP speech was perceived as being louder than the vocoded speech, the level of LP speech was decreased in 1 dB steps until it was perceived as being softer. This procedure was completed three times to determine the equal-loudness level of the vocoded and LP stimulus. Last, both the vocoded and LP speech were presented simultaneously for a direct comparison, in order to further confirm the equal loudness of vocoded and LP speech across ears. For the majority of the participants, the final presentation level of the LP speech was 70 dB A, i.e., 2 dB higher than the level of the vocoded speech.

Session one—After having completed the loudness-balancing procedure, each participant underwent a training session to adapt to the spectrally-degraded vocoded speech. First, participants practiced listening to three lists (a total of 30 sentences) of 8-channel vocoded IEEE sentences ([IEEE, 1969](#)). During training, participants could listen to each sentence as many times in a row as they wished before typing in the words they heard. Written feedback was then displayed on the computer screen. After training, participants were tested on three additional lists of IEEE sentences. During this test phase, there was no replay of the stimuli or written feedback. Participants who attained at least 80% correct word recognition of the IEEE sentence task were also tested on three other lists of IEEE sentences in a vocoder + LP condition to confirm that performance remained at high levels, that is, above 80% correct. All participants fulfilled these two performance criteria.

Next, each participant was familiarized with all words of the main experiment. Following the procedure used by [Jesse et al. \(2017\)](#), during this familiarization phase, all words were randomly presented visually on a computer screen one at a time for participants to read out loud. This procedure was self-paced without feedback.

The main experiment employed the visual world paradigm with printed words, following [Jesse et al. \(2017\)](#). Participants' fixations of their right eye on the words were recorded with a desktop-mounted Eyelink 1000+ system (SR Research, Ltd., Canada) at a sampling rate of 1 kHz. During the experiment, each trial started with a fixation cross shown in the center of the computer screen for 500 ms, followed by a black screen of 200 ms. Next, four printed words (Lucida Sans Typewriter, size 20) were displayed, with each word centered in one of the four quadrants of the screen. Participants heard the sentence "Click on the word (*target*)."

The timing of this presentation was such that the acoustic onset of the target word occurred 1800 ms after display onset. On each trial, displays contained the target word (e.g., *admiral*), a competitor word (*admiration*), and two unrelated distractor words (*converter*, *convergence*). Participants responded by clicking on the corresponding word displayed on the screen. Each participant had 2 s to respond and the display went away once a response was given or after the 2-s time limit had passed. The inter-trial interval was 440 ms.

The experiment began with one practice block, which contained two presentations of six practice sets. Participants then completed four main test blocks per listening condition.

They were tested only in one listening condition during the first session, and completed testing in the other listening condition in the second session. The order of the listening condition was counterbalanced across participants. In each of the main test blocks, 24 critical sets and the 8 filler sets were presented once in random order. Auditory targets within a block had equally often primary stress or secondary stress on the first syllable. The assignment of which word within a pair was used as a target was counterbalanced across participants within each listening condition. On the critical trials of the first block, a word from the critical stress pairs was always presented as auditory target (e.g., *admiral* or *admiration*). Repetitions of the word sets in subsequent blocks were set up such that each word from a set was equally likely to be the target. The order of blocks 2 to 4 within each listening condition was counterbalanced across participants, and the presentation order within a block was random for each participant. Printed words were pseudo-randomly assigned to the quadrants of the screen, such that targets and competitors occurred equally often in each position within each block. Drift corrections were conducted before the main experiment and after every eighth trial.

Session Two—Participants were tested in the eye-tracking experiment on the second listening condition on a different day. As in the first session, a familiarization phase was provided again before the main experiment. The procedures for the main experiment were the same as those used in the first test session.

After the eye-tracking experiment, each participant was tested on their ability to recognize the words (words from all critical and non-critical pairs) that had been presented in the main experiment in each listening condition. The words were presented via the headphones randomly one at a time (with no visual display), and participants were asked to type in the word they heard. No feedback was given. The order of listening condition tested was counterbalanced across participants.

III. RESULTS

Participants' performance in recognizing individual words after the eye-tracking experiment was equally high in the vocoder condition [$M = 99\%$, standard deviation (SD) = 1.33] and in the vocoder + LP condition ($M = 100\%$, SD = 0.73). The degree of spectral degradation implemented in this study therefore does not significantly affect listeners' ability to recognize the test words from segmental cues.

Eye-fixation data for the first auditory presentation of each critical word in each listening condition were analyzed for trials with a correct response and no fixations off screen. These criteria led to 95% of the critical trials in the vocoder condition and 94% in vocoder + LP condition being included in the analyses. Correct responses were defined as any response given within 192 pixels around the printed version of the target word. Accuracy on critical trials was 99% for both listening conditions (vocoder: mean reaction time of correct responses = 1168 ms, SD = 242 ms; vocoder + LP: mean reaction time of correct responses = 1151 ms, SD = 237 ms).

Following the data analysis methods used in our previous study (Jesse *et al.*, 2017), fixations were categorized as belonging to a printed word if they fell within 192 pixels around the word. For each target stress type (i.e., primary vs secondary), we calculated the proportions of fixations on critical targets, competitors, and averaged-over unrelated distractors for time windows corresponding to the first and second syllable of the target word. To account for the estimated time for human listeners to program an eye movement (e.g., Matin *et al.*, 1993), the time windows were shifted by 200 ms. Figure 1 shows the proportions of fixations on targets (black solid line), competitors (gray dashed line), and distractors (black dotted line) over time by stress type of the first syllable of the critical target words (left: primary; right: secondary) for the vocoder condition (top panel) and the vocoder + LP condition (bottom panel). Vertical lines indicate (from left to right; all shifted by 200 ms) the onset of the target word, the average offset of the first and second syllables, and the average offset of the segmental overlap between target and competitor. The data suggest that listeners fixated on targets more than on competitors before the end of the second syllable in the vocoder + LP condition only. This pattern of results suggests that listeners can distinguish targets and competitors from suprasegmental cues to lexical stress (i.e., fixations to target and competitor diverged after the first syllable when the items of critical pairs only differed in pitch and amplitude), only in the condition where they were also provided with low-frequency fine-structure information (vocoder + LP condition). Listeners use cues to primary and secondary lexical stress to distinguish targets from their segmentally-matching but suprasegmentally-mismatching competitors.

For the statistical analyses, target preference was calculated as the difference in logit-transformed fixation proportions of targets and competitors, and competitor preference was calculated as the difference in logit-transformed fixation proportions of competitors and distractors. We used mixed-effect models, as implemented in the lmer function (lme4 package, Bates and Sarkar, 2009) in the R statistical program (Version 3.1.2; R Core Team, 2016), to analyze the effect of stress type of the target word and listening condition (vocoder vs vocoder + LP) on these two dependent variables. Stress type and listening condition were contrast-coded fixed factors (primary stress = -0.5 , secondary stress = 0.5 ; vocoder = -0.5 , vocoder + LP = 0.5). Subjects and items were random factors. By-subject slope adjustments were included for stress type, and by-subject and by-item slope adjustments were included for listening condition (Barr *et al.*, 2013). Satterthwaite approximations were used to estimate p -values.

Analyses of target preference during the first syllable showed no overall target preference [$\beta = 0.07$, standard error (SE) = 0.14, $p = 0.6$], but target preference was modulated by listening condition ($\beta = 0.7$, SE = 0.27, $p < 0.01$), such that targets were fixated significantly more than competitors during the first syllable when presented with vocoded + LP speech ($\beta = 0.42$, SE = 0.19, $p < 0.05$) but not when presented with vocoded speech ($\beta = -0.27$, SE = 0.21, $p = 0.19$). Stress type had no effect on target preference and

FIG. 1. Fixation proportions over time for targets, competitors, and distractors for the vocoder condition (upper panel) and the vocoder + LP condition (lower panel). Data are plotted separately by stress type of the first syllable of the target (left: primary stress; right: secondary stress). Vertical lines show the onset of the target word (200 ms), the average offset of the first and the second syllables, and the average offset of the segmental overlap between pairs. These time markers were all shifted by 200 ms.

did not interact with listening condition (all $p > 0.05$). During the second syllable, targets were fixated significantly more than competitors ($\beta = 0.35$, $SE = 0.16$, $p = 0.03$). Again, this target preference was modulated by listening condition ($\beta = 0.89$, $SE = 0.28$, $p < 0.01$), in that target preference emerged only in the processing of vocoded + LP speech ($\beta = 0.79$, $SE = 0.20$, $p < 0.0001$) but not in the processing of vocoded speech ($\beta = -0.09$, $SE = 0.21$, $p = 0.67$). The stress type of the initial syllable had no influence on target preference ($\beta = 0.04$, $SE = 0.34$, $p = 0.91$) and the interaction between stress and listening condition was not significant ($p > 0.05$). In summary, these results suggest that, when listening to spectrally-degraded vocoded speech, listeners were unable to use suprasegmental cues to lexical stress to recognize words. Regardless of the stress type of target words, proportions of fixations on target and competitor words only diverged after the two competing words were disambiguated by the segmental information in the third syllable. However, when low-frequency fine-structure information was added to the vocoded speech to simulate bimodal listening, listeners efficiently utilized suprasegmental cues to lexical stress to distinguish words from their phonological competitors before they differed segmentally. Unlike for unprocessed speech (Jesse *et al.*, 2017), in the present study listeners were able to exploit suprasegmental cues to both

primary and secondary stress on the first syllable of the target word.

Analyses of competitor effects (i.e., the difference in logit-transformed fixation proportions between competitors and distractors) showed, similar to our previous findings for English with unprocessed speech, only a general preference for competitors over distractors for both time windows (first syllable: $\beta = 0.82$, $SE = 0.14$, $p < 0.00001$; second syllable: $\beta = 1.91$, $SE = 0.18$, $p < 0.00001$), indicating general phonological competition. Words that were phonologically overlapping with the targets competed for recognition, regardless of the degree of stress on their first syllable and the listening condition (all $p > 0.05$). This result indicates that the degree of phonological competition due to segmental similarity to the target word is approximately the same in both listening conditions.

IV. DISCUSSION

The current study investigated the effect of speech degradation with and without additional low-frequency fine-structure information on the perception of prosodic information for word recognition in English. Using a visual world paradigm, we examined how spectral degradation affects English listeners' ability to efficiently use suprasegmental

cues to lexical stress to change the time course of online spoken-word recognition. Normal-hearing listeners were tested with vocoded speech that simulated electric hearing and with vocoded + LP speech that simulated bimodal hearing. We found that spectral degradation negatively affected the perception of lexical stress and that only when additional low-frequency fine-structure information was added to the vocoded speech were listeners able to use suprasegmental cues to lexical stress for the facilitation of online spoken-word recognition.

A. Effect of spectral degradation on using lexical stress cues in online spoken-word recognition

Similar to findings for Dutch listeners (Reinisch *et al.*, 2010), Jesse *et al.* (2017) provided evidence for English listeners' early use of suprasegmental cues to lexical stress for word recognition; that is, listeners utilized these cues as soon as they became available during online word recognition. However, the early use of primary stress information for word recognition was not found in the present study when English listeners were presented with spectrally degraded speech. For the vocoder condition, a preference for fixating on the target over the competitor did not emerge for either primary or secondary stress-initial targets until segmental information disambiguated the two words in the third syllable. As described in our acoustic analyses above, while the first and second formant frequencies in our unprocessed stimuli were similar in the first two syllables between primary- and secondary-stressed target words (i.e., no difference in segmental cues), there was a significant difference in pitch excursion and a small, significant difference in amplitude (by 1.4 dB) between the two stress types. Noise-channel vocoding discards, however, the fine-structure cues necessary for pitch encoding. Consequently, this lack of access to pitch cues in the vocoded speech compared to the unprocessed speech slowed down listeners' word recognition process, as the listeners then had to wait for disambiguating segmental information in the third syllable of these words. Although channel vocoding preserves the amplitude difference between primary- and secondary-stressed target words, listeners did not use this cue to lexical stress in the vocoder condition. There are several possible explanations for this result: (1) the small difference in amplitude may not be sufficient to help in distinguishing primary and secondary stress, (2) amplitude is not a reliable cue to the degree of lexical stress (e.g., Beckman, 1986; Fear *et al.*, 1995; Mattys, 2000; Sluijter and Van Heuven, 1996) and seems to be a less important cue for lexical stress in English (e.g., Chrabaszcz *et al.*, 2014; Fry, 1955, 1958; Lehiste, 1970; Mattys, 2000), and/or (3) listeners' ability to perceive slight temporal and intensity changes in the spectrally degraded signal is reduced, as suggested by previous vocoder and CI studies (Oxenham and Kreft, 2014; Rogers *et al.*, 2006; Sagi *et al.*, 2009; Won *et al.*, 2011).

Aside from the delayed word recognition for critical primary-stressed target words, indicating an inability to use suprasegmental cues to lexical stress, vocoder processing with eight channels did not affect the time course of recognizing words from segmental information. First, fixations on target and competitors diverged once segmental information

disambiguated the words during the third syllable. Second, there was a significant effect of general phonological competition (i.e., participants fixated more on the competitors than on the unrelated *distractors*), already apparent during the first syllable when segmental information distinguished the critical words from the distractor words. These patterns of results suggest that the delayed recognition of critical target words in the vocoder condition cannot be attributed to an overall longer processing time that may be necessary for spectrally degraded speech signals.

B. Importance of low-frequency fine-structure cues for the use of lexical stress in online spoken-word recognition

Lexical stress was not exploited in online word recognition when listening to vocoded speech alone, but it did modulate the time course of word recognition in the simulated bimodal hearing condition. Remarkably, listeners used these prosodic cues immediately to recognize target words, capitalizing on them during a time when segmental information did not yet distinguish target and competitor. Given that pitch excursion was the only viable cue to lexical stress in our stimulus set that was present in the vocoder + LP condition but not in the vocoder condition, we suggest that it was the better pitch encoding provided by the low-frequency fine-structure cues that restored listeners' ability to use prosodic cues for faster word recognition, despite the high degree of degradation (i.e., vocoded speech) in the higher frequency regions. The present eye-tracking results showed that differentiating primary and secondary stress-initial words was significantly faster with simulated bimodal listening than with simulated CI listening. This finding is consistent with previous studies using pitch-related tasks, such as melody identification (e.g., Kong *et al.*, 2005), lexical tone identification (e.g., Li *et al.*, 2014), and question-statement discrimination (Most *et al.*, 2011; Straatman *et al.*, 2010), that showed improved performance with EAS listening compared to CI alone.

The difference in the eye-fixation results for the use of lexical stress with and without low-frequency fine-structure cues cannot be attributed to the reduced ability to recognize the words from segmental cues in the vocoder-only condition. First, speech intelligibility (assessed during the eye-tracking experiment with forced-choice responses using the computer mouse and tested in a word recognition task after the eye-tracking experiment) was at ceiling level performance of 99%–100% in both vocoder and vocoder + LP conditions. Second, the degree of lexical competition (measured as the preference for fixating competitors over segmentally-unrelated distractors) was similar in both listening conditions ($p > 0.05$). Third, the first two syllables of primary- and secondary-stressed words were identical in terms of their segmental cues and only differed suprasegmentally in pitch excursion and amplitude cues. Listeners therefore cannot disambiguate the primary- and secondary-stressed words after the first syllable based on segmental cues alone. These results further argue against an increase in

speech intelligibility in the vocoder + LP condition driving the earlier recognition of critical target words.

The online use of prosodic cues for simulated bimodal hearing was observed for *both* primary and secondary stress-initial target words in the current study. An additional analysis was carried out to compare target preference during the second syllable on trials with secondary stress-initial target words in our vocoder + LP condition and in the unprocessed speech condition from Jesse *et al.* (2017). We found that listening condition, evaluated as a contrast-coded factor ($-0.5 = \text{unprocessed}$, $0.5 = \text{vocoder} + \text{LP}$), was not significantly different ($\beta = 0.75$, $SE = 0.43$, $p = 0.095$). This result means that although we observed preference for target words over competitor words during the second syllable in the vocoder + LP condition, this target preference is not statistically different from the lack of such an earlier target preference observed in the unprocessed speech condition.

C. Clinical implications

Many recently implanted CI users have low-frequency residual hearing in the implanted or the non-implanted ear, or both. Clinically, the presence of an EAS benefit is usually determined by showing a difference in the percentage of correct word recognition between EAS listening and CI alone listening. Potential EAS users may have been discouraged to continue to use their hearing aid in the implanted and/or non-implanted ear based on this conventional practice. The findings of the present study reinforce the need for new tools or evaluation metrics in capturing benefits or treatment effectiveness beyond the traditional measure of percentage correct for speech intelligibility. Recent studies have shown a reduction in listening effort (measured by the reaction time to the secondary task or by the change of pupil size) as the quality of the auditory signal improved beyond the asymptotic performance in a recognition task (e.g., Sarampalis *et al.*, 2009; Pals *et al.*, 2013; Winn *et al.*, 2015).

Our study offers a new approach to evaluate the EAS benefit in terms of the time course of processing during spoken-word recognition. Our results showed that although percentage word recognition was at ceiling for both vocoder and vocoder + LP conditions, the simulated bimodal hearing provided an additional benefit to listeners as it allowed them to efficiently use prosodic information online to speed up word recognition. The current findings thus contribute to our understanding of the temporal dynamics of spoken-word recognition by demonstrating that low-frequency fine-structure information can facilitate the resolution of lexical competition for word recognition by providing suprasegmental cues to lexical stress. Reduction of processing time at the word level could have a significant cascading benefit on downstream processing of sentences and discourse. In addition, faster and more efficient encoding and updating of speech information online could potentially free up cognitive resources for other processing demands and improve memory for language (Zellin *et al.*, 2011).

ACKNOWLEDGMENTS

We thank Hanna Millstine and Suzette Xie for their assistance in data collection, and Ala Somarowthu and Katja

Poellmann for technical support and stimulus generation. This work was supported by NIH R01-DC012300 to Y.-Y.K.

¹Stress is indicated with the standard diacritic marks of the International Phonetic Association. For example, 'ad is the primary-stressed beginning of admiral and ,ad is the secondary-stressed first syllable of admiration.

²The amplitude difference in the first syllable between the primary and secondary stress in vocoded speech and in LP-filtered speech was almost identical (1.5 dB) to that in unprocessed speech; also the pitch excursion difference in LP-filtered speech was the same as in unprocessed speech, as pitch information is preserved in the low-frequency portion of the speech signal. Importantly, this finding means that vocoded speech and LP-filtered speech only differed in their availability of pitch excursion as a cue.

- Alloppenna, P. D., Magnuson, J. S., and Tanenhaus, M. K. (1998). "Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models," *J. Mem. Lang.* **38**, 419–439.
- Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (2013). "Random effects structure for confirmatory hypothesis testing: Keep it maximal," *J. Mem. Lang.* **68**, 255–278.
- Bates, D. M., and Sarkar, D. (2009). lme4: Linear mixed-effects models using eigen and variational algorithms (Version R package version 0.999375-27).
- Beckman, M. E. (1986). *Stress and Non-Stress Accent* (Netherlands Phonetics Archives, Foris, Dordrecht), Vol. VII, 239 pp.
- Bernstein, J. G., and Oxenham, A. J. (2003). "Pitch discrimination of diotic and dichotic tone complexes: Harmonic resolvability or harmonic number?," *J. Acoust. Soc. Am.* **113**, 3323–3334.
- Bierer, J. A., Faulkner, K. F., and Tremblay, K. L. (2011). "Identifying cochlear implant channels with poor electrode-neuron interfaces: Electrically evoked auditory brain stem responses measured with the partial tripolar configuration," *Ear Hear.* **32**, 436–444.
- Carlyon, R. P., Deeks, J. M., and McKay, C. M. (2010). "The upper limit of temporal pitch for cochlear-implant listeners: Stimulus duration, conditioner pulses, and the number of electrode stimulated," *J. Acoust. Soc. Am.* **127**, 1469–1478.
- Chatterjee, M., and Peng, S.-C. (2008). "Processing F0 with cochlear implants: Modulation frequency discrimination and speech intonation recognition," *Hear. Res.* **235**, 143–156.
- Chrabaszcz, A., Winn, M., Lin, C. Y., and Idsardi, W. J. (2014). "Acoustic cues to perception of word stress by English, Mandarin, and Russian speakers," *J. Speech Lang. Hear. Res.* **57**, 1468–1479.
- Ciocca, V., Francis, A. L., Aisha, R., and Wong, L. (2002). "The perception of Cantonese lexical tones by early-deafened cochlear implantees," *J. Acoust. Soc. Am.* **111**, 2250–2256.
- Cooper, N., Cutler, A., and Wales, R. (2002). "Constraints of lexical stress on lexical access in English: Evidence from native and non-native listeners," *Lang. Speech.* **45**, 207–228.
- Cullington, H. E., and Zeng, F.-G. (2008). "Speech recognition with varying numbers and types of competing talkers by normal-hearing, cochlear-implant, and implant simulation subjects," *J. Acoust. Soc. Am.* **123**, 450–461.
- Cutler, A., and Pasveer, D. (2006). "Explaining cross linguistic differences in effects of lexical stress on spoken-word recognition," in *Proceedings of the Third International Conference on Speech Prosody*, edited by R. Hoffman and H. Mixdorff (TUD Press, Dresden), pp. 250–254.
- Dahan, D., Magnuson, J. S., Tanenhaus, M. K., and Hogan, E. M. (2001). "Subcategorical mismatches and the time course of lexical access: Evidence for lexical competition," *Lang. Cog. Process.* **16**, 507–534.
- Dai, H. (2000). "On the relative influence of individual harmonics on pitch judgment," *J. Acoust. Soc. Am.* **107**, 953–959.
- Davies, M. (2008). The Corpus of Contemporary American English: 450 million words, 1990–present. Available online at <http://corpus.byu.edu/coca/> (Last viewed March 14, 2014).
- Dorman, M. F., Gifford, R. H., Spahr, A. J., and McKarns, S. A. (2008). "The benefits of combining acoustic and electric stimulation for the recognition of speech, voice and melodies," *Audiol. Neurotol.* **13**, 105–112.
- Dorman, M. F., Loizou, P., Wang, S., Zhang, T., Spahr, A., Loisel, L., and Cook, S. (2014). "Bimodal cochlear implants: The role of acoustic signal level in determining speech perception benefit," *Audiol. Neurotol.* **19**, 234–238.
- Fear, B. D., Cutler, A., and Butterfield, S. (1995). "The strong/weak syllable distinction in English," *J. Acoust. Soc. Am.* **97**, 1893–1904.

- Fishman, K. E., Shannon, R. V., and Slattery, W. H. (1997). "Speech recognition as a function of the number of electrodes used in the SPEAK cochlear implant speech processor," *J. Speech Lang. Hear. Res.* **40**, 1201–1215.
- Friedrich, C. K., Kotz, S. A., Friederici, A. D., and Gunter, T. C. (2004). "ERPs reflect lexical identification in word fragment priming," *J. Cog. Neurosci.* **16**, 541–552.
- Friesen, L., Shannon, R., Başkent, D., and Wang, X. (2001). "Speech recognition in noise as a function of the number of spectral channels: Comparison of acoustic hearing and cochlear implants," *J. Acoust. Soc. Am.* **110**, 1150–1163.
- Fry, D. B. (1955). "Duration and intensity as physical correlates of linguistic stress," *J. Acoust. Soc. Am.* **27**, 765–768.
- Fry, D. B. (1958). "Experiments in the perception of stress," *Lang. Speech* **1**, 126–152.
- Fu, Q.-J., and Nogaki, G. (2005). "Noise susceptibility of cochlear implant users: The role of spectral resolution and smearing," *J. Assoc. Res. Otolaryngol.* **6**, 19–27.
- Galvin, J. J., III, Fu, Q. J., and Shannon, R. V. (2009). "Melodic contour identification and music perception by cochlear implant users," *Ann. N.Y. Acad. Sci.* **1169**, 518–533.
- Greenwood, D. (1990). "A cochlear frequency-position function for several species—29 years later," *J. Acoust. Soc. Am.* **87**, 2592–2605.
- Hilbert, D. (1912). *Principles of a General Theory of the Linear Integral Equations* (Teubner, Leipzig), 282 pp.
- Holt, C. M., Demuth, K., and Yuen, I. (2016). "The use of prosodic cues in sentence processing by prelingually deaf users of cochlear implants," *Ear. Hear.* **37**, e256–e262.
- Holt, C. M., and McDermott, H. J. (2013). "Discrimination of intonation contours by adolescents with cochlear implants," *Int. J. Audiol.* **52**, 808–815.
- Houtsma, A. J. M., and Smurzynski, J. (1990). "Pitch identification and discrimination for complex tones with many harmonics," *J. Acoust. Soc. Am.* **87**, 304–310.
- Hughes, M. L. (2008). "A re-evaluation of the relation between physiological channel interaction and electrode pitch ranking in cochlear implant," *J. Acoust. Soc. Am.* **124**, 2711–2714.
- IEEE (1969). "IEEE recommended practice for speech quality measurements," *IEEE Trans. Audio Electroacoust.* **17**, 225–246.
- Ito, K., and Speer, S. R. (2008). "Anticipatory effects of intonation: Eye movements during instructed visual speech," *J. Mem. Lang.* **58**, 541–573.
- Jesse, A., and McQueen, J. M. (2014). "Suprasegmental lexical stress cues in visual speech can guide spoken-word recognition," *Q. J. Exp. Psychol.* **67**, 793–808.
- Jesse, A., Poellmann, K., and Kong, Y.-Y. (2017). "English listeners use suprasegmental cues to lexical stress early during spoken-word recognition," *J. Speech Lang. Hear. Res.* **60**, 190–198.
- Jones, G. L., Won, J. H., Drennan, W. R., and Rubinstein, J. T. (2013). "Relationship between channel interaction and spectral-ripple discrimination in cochlear implant users," *J. Acoust. Soc. Am.* **133**, 425–433.
- Kong, Y. Y., and Braid, L. D. (2011). "Cross-frequency integration for consonant and vowel identification in bimodal hearing," *J. Speech Lang. Hear. Res.* **54**, 959–980.
- Kong, Y. Y., Cruz, R., Jones, J. A., and Zeng, F. G. (2004). "Music perception with temporal cues in acoustic and electric hearing," *Ear. Hear.* **25**, 173–185.
- Kong, Y. Y., Deeks, J. M., Axon, P. R., and Carlyon, R. P. (2009). "Limits of temporal pitch in cochlear implants," *J. Acoust. Soc. Am.* **125**, 1649–1657.
- Kong, Y. Y., Stickney, G. S., and Zeng, F. G. (2005). "Speech and melody recognition in binaurally combined acoustic and electric hearing," *J. Acoust. Soc. Am.* **117**, 1351–1361.
- Lee, K. Y., van Hasselt, C. A., Chiu, S. N., and Cheung, D. M. (2002). "Cantonese tone perception ability of cochlear implant children in comparison with normal-hearing children," *Int. J. Pediatr. Otorhinolaryngol.* **63**, 137–147.
- Lehiste, I. (1970). *Suprasegmentals* (MIT, Cambridge, MA), 194 pp.
- Li, Y., Zhang, G., Galvin, J. J., III, and Fu, Q. J. (2014). "Mandarin speech perception in combined electric and acoustic stimulation," *PLoS One* **9**(11), e112471.
- Luce, P. A., and Pisoni, D. B. (1998). "Recognizing spoken words: The neighborhood activation model," *Ear. Hear.* **19**, 1–36.
- Marx, M., James, C., Foxton, J., Capber, A., Frayssé, B., Barone, P., and Deguine, O. (2015). "Speech prosody perception in cochlear implant users with and without residual hearing," *Ear. Hear.* **36**, 239–248.
- Matin, E., Shao, K. C., and Boff, K. R. (1993). "Saccadic overhead: Information-processing time with and without saccades," *Percept. Psychophys.* **53**, 372–380.
- Mattys, S. L. (2000). "The perception of primary and secondary stress in English," *Percept. Psychophys.* **62**, 253–265.
- McDermott, H. J. (2004). "Music perception with cochlear implants: A review," *Trends Amplif.* **8**, 49–82.
- Meister, H., Landwehr, M., Pyschny, V., Wagner, P., and Walger, M. (2011). "The perception of sentence stress in cochlear implant recipients," *Ear. Hear.* **32**, 459–467.
- Meister, H., Landwehr, M., Pyschny, V., Walger, M., and von Wedel, H. (2009). "The perception of prosody and speaker gender in normal-hearing listeners and cochlear implant recipients," *Int. J. Audiol.* **48**, 38–48.
- Morris, D., Magnusson, L., Faulkner, A., Jönsson, R., and Juul, H. (2013). "Identification of vowel length, word stress, compound words and phrases by postlingually deafened cochlear implant listeners," *J. Am. Acad. Audiol.* **24**, 879–890.
- Most, T., Harel, T., Shpak, T., and Luntz, M. (2011). "Perception of supra-segmental speech features via bimodal stimulation: Cochlear implant on one ear and hearing aid on the other," *J. Speech Lang. Hear. Res.* **54**, 668–678.
- Norris, D., and McQueen, J. M. (2008). "Shortlist B: A Bayesian model of continuous speech recognition," *Psychol. Rev.* **115**, 357–395.
- Oxenham, A. J., and Kreft, H. A. (2014). "Speech perception in tones and noise via cochlear implants reveals influence of spectral resolution on temporal processing," *Trends Hear.* **18**, 1–14.
- Pals, C., Sarampalis, A., and Başkent, D. (2013). "Listening effort with cochlear implant simulations," *J. Speech Lang. Hear. Res.* **56**, 1075–1084.
- Peng, S. C., Lu, N., and Chatterjee, M. (2009). "Effects of cooperating and conflicting cues on speech intonation recognition by cochlear implant users and normal hearing listeners," *Audiol. Neurotol.* **14**, 327–337.
- Peng, S. C., Tomblin, J. B., Cheung, H., Lin, Y. S., and Wang, L. S. (2004). "Perception and production of Mandarin tones in prelingually deaf children with cochlear implants," *Ear. Hear.* **25**, 251–264.
- Peng, S. C., Tomblin, J. B., and Turner, C. W. (2008). "Production and perception of speech intonation in pediatric cochlear implant recipients and individuals with normal hearing," *Ear. Hear.* **29**, 336–351.
- Pfingst, B. E., Colesa, D. J., Hembrador, S., Kang, S. Y., Middlebrooks, J. C., Raphael, Y., and Su, G. L. (2011). "Detection of pulse trains in the electrically stimulated cochlea: Effects of cochlear health," *J. Acoust. Soc. Am.* **130**, 3954–3968.
- Plomp, R. (1967). "Pitch of complex tones," *J. Acoust. Soc. Am.* **41**, 1526–1533.
- R Core Team (2016). R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/> (Last viewed June 12, 2016).
- Reinisch, E., Jesse, A., and McQueen, J. M. (2010). "Early use of phonetic information in spoken word recognition: Lexical stress drives eye movements immediately," *Q. J. Exp. Psych.* **63**, 772–783.
- Ritsma, R. J. (1967). "Frequencies dominant in the perception of the pitch of complex sounds," *J. Acoust. Soc. Am.* **42**, 191–198.
- Rogers, C. F., Healy, E. W., and Montgomery, A. A. (2006). "Sensitivity to isolated and concurrent intensity and fundamental frequency increments by cochlear implant users under natural listening conditions," *J. Acoust. Soc. Am.* **119**, 2276–2287.
- Sagi, E., Kaiser, A. R., Meyer, T. A., and Svirsky, M. A. (2009). "The effect of temporal gap identification on speech perception by users of cochlear implants," *J. Speech Lang. Hear. Res.* **52**, 385–395.
- Salverda, A. P., Dahan, D., and McQueen, J. M. (2003). "The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension," *Cognition* **90**, 51–89.
- Salverda, A. P., Dahan, D., Tanenhaus, M. K., Crosswhite, K., Masharov, M., and McDonough, J. (2007). "Effects of prosodically modulated sub-phonetic variation on lexical competition," *Cognition* **105**, 466–476.
- Sarampalis, A., Kalluri, S., Edwards, B., and Hafter, E. (2009). "Objective measures of listening effort: Effects of background noise and noise reduction," *J. Speech Lang. Hear. Res.* **52**, 1230–1240.
- See, R. L., Driscoll, V. D., Gfeller, K., Kliethermes, S., and Oleson, J. (2013). "Speech intonation and melodic contour recognition in children

- with cochlear implants and with normal hearing," *Otol. Neurotol.* **34**, 490–498.
- Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., and Ekelid, M. (1995). "Speech recognition with primarily temporal cues," *Science* **270**, 303–304.
- Singh, S., Kong, Y. Y., and Zeng, F. G. (2009). "Cochlear implant melody recognition as a function of melody frequency range, harmonicity, and number of electrodes," *Ear Hear.* **30**, 160–168.
- Sluijter, A. M., and van Heuven, V. J. (1996). "Spectral balance as an acoustic correlate of linguistic stress," *J. Acoust. Soc. Am.* **100**, 2471–2485.
- Smith, Z. M., Delgutte, B., and Oxenham, A. J. (2002). "Chimaeric sounds reveal dichotomies in auditory perception," *Nature* **416**, 87–90.
- Soto-Faraco, S., Sebastian-Galles, N., and Cutler, A. (2001). "Segmental and suprasegmental mismatch in lexical access," *J. Mem. Lang.* **45**, 412–432.
- Spitzer, S., Liss, J., Spahr, T., Dorman, M., and Lansford, K. (2009). "The use of fundamental frequency for lexical segmentation in listeners with cochlear implants," *J. Acoust. Soc. Am.* **125**, EL236–EL241.
- Stickney, G. S., Zeng, F.-G., Litovsky, R., and Assmann, P. (2004). "Cochlear implant speech recognition with speech maskers," *J. Acoust. Soc. Am.* **116**, 1081–1091.
- Straatman, L. V., Rietveld, A. C. M., Beijen, J., Mylanus, E. A. M., and Mens, L. H. M. (2010). "Advantage of bimodal fitting in prosody perception for children using a cochlear implant and a hearing aid," *J. Acoust. Soc. Am.* **128**, 1884–1895.
- Sulpizio, S., and McQueen, J. M. (2012). "Italians use abstract knowledge about lexical stress during spoken-word recognition," *J. Mem. Lang.* **66**, 177–193.
- Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., and Sedivy, J. C. (1995). "Integration of visual and linguistic information in spoken language comprehension," *Science* **268**, 1632–1634.
- van Donselaar, W., Koster, M., and Culter, A. (2005). "Exploring the role of lexical stress in lexical recognition," *Q. J. Exp. Psychol.* **58**, 251–273.
- van Heuven, V. J., and Hagman, P. J. (1988). "Lexical statistics and spoken word recognition in Dutch," in *Linguistics in the Netherlands*, edited by P. Coopmans and A. Hulk (Foris, Dordrecht), pp. 59–69.
- Wei, C. G., Cao, K., and Zeng, F.-G. (2004). "Mandarin tone recognition in cochlear-implant subjects," *Hear. Res.* **197**, 87–95.
- Winn, M. B., Edwards, J. R., and Litovsky, R. Y. (2015). "The impact of auditory spectral resolution on listening effort revealed by pupil dilation," *Ear Hear.* **36**, e153–e165.
- Won, J. H., Drennan, W. R., Nie, K., Jameyson, E. M., and Rubinstein, J. T. (2011). "Acoustic temporal modulation detection and speech perception in cochlear implant listeners," *J. Acoust. Soc. Am.* **130**, 376–388.
- Zellin, M., Pannekamp, A., Toepel, U., and van der Meer, E. (2011). "In the eye of the listener: Pupil dilation elucidates discourse processing," *Int. J. Psychophysiol.* **81**, 133–141.
- Zeng, F. G. (2002). "Temporal pitch in electric hearing," *Hear. Res.* **174**, 101–106.
- Zeng, F.-G., Nie, K., Stickney, G. S., Kong, Y.-Y., Vongphoe, M., Bhargave, A., Wei, C., and Cao, K. (2005). "Speech recognition with amplitude and frequency modulations," *Proc. Natl. Acad. Sci. U.S.A.* **102**, 2293–2298.
- Zeng, F.-G., Tang, Q., and Lu, T. (2014). "Abnormal pitch perception produced by cochlear implant stimulation," *PLoS One* **9**(2), e88662.
- Zwitserslood, P. (1989). "The locus of the effects of sentential-semantic context in spoken-word processing," *Cognition* **32**, 25–64.