Perceptual weighting of the envelope and fine structure across frequency bands for sentence intelligibility: Effect of interruption at the syllabic-rate and periodic-rate of speech

Daniel Fogerty^{a)}

Department of Speech and Hearing Sciences, Indiana University, Bloomington, Indiana 47405

(Received 23 November 2010; accepted 25 April 2011)

Listeners often only have fragments of speech available to understand the intended message due to competing background noise. In order to maximize successful speech recognition, listeners must allocate their perceptual resources to the most informative acoustic properties. The speech signal contains temporally-varying acoustics in the envelope and fine structure that are present across the frequency spectrum. Understanding how listeners perceptually weigh these acoustic properties in different frequency regions during interrupted speech is essential for the design of assistive listening devices. This study measured the perceptual weighting of young normal-hearing listeners for the envelope and fine structure in each of three frequency bands for interrupted sentence materials. Perceptual weights were obtained during interruption at the syllabic rate (i.e., 4 Hz) and the periodic rate (i.e., 128 Hz) of speech. Potential interruption interactions with fundamental frequency information were investigated by shifting the natural pitch contour higher relative to the interruption rate. The availability of each acoustic property was varied independently by adding noise at different levels. Perceptual weights were determined by correlating a listener's performance with the availability of each acoustic property on a trial-by-trial basis. Results demonstrated similar relative weights across the interruption conditions, with emphasis on the envelope in high-frequencies.

© 2011 Acoustical Society of America. [DOI: 10.1121/1.3592220]

PACS number(s): 43.71.Gv, 43.71.Es, 43.72.Ar, 43.66.Ba [MAH]

Pages: 489–500

I. INTRODUCTION

In everyday listening environments speech is commonly interrupted by some interfering background noise. A number of studies have demonstrated that normal-hearing listeners receive a perceptual benefit for recognizing speech when amplitude dips are introduced into a continuous masker (e.g., Gnasia *et al.*, 2008; Nelson, Jin, Carney, and Nelson, 2003). This phenomenon, termed “masking release,” suggests that listeners are able to take advantage of the brief moments when the speech is at a more favorable signal-to-noise ratio (SNR). These brief spectral-temporal moments have been defined as speech “glimpses” (Cooke, 2003).

Relatively recently, research has begun to focus on the acoustic properties that facilitate glimpsing. Findings indicate that listeners with hearing-impaired, cochlear implants, or normal-hearing listeners with simulated cochlear implants receive less benefit from the presence of speech glimpses (e.g., Bernstein and Grant, 2009; Gnasia *et al.*, 2008; Nelson *et al.*, 2003). Cochlear implant users receive only amplitude modulations of the speech envelope (E), while a complementary temporal cue of frequency modulation, the temporal fine structure (TFS), is absent. These findings suggest that the presence of temporal fine structure information may be essential for the successful processing of speech glimpses (Lorenzi *et al.*, 2006).

The current study was designed to assess how normal-hearing listeners weight different spectral and temporal in-

formation during the presence of interrupted speech. In particular, measurements were made of the relative weighting of the speech envelope and temporal fine structure in each of three frequency bands. Experiment 1 interrupted speech at the syllabic rate of speech (i.e., 4 Hz). Experiment 2 investigated perceptual weighting during interruption near the periodic rate of speech (i.e., 128 Hz). Possible interference of speech periodicity cues was investigated at this rate by also measuring perceptual weighting for speech with a fundamental frequency (F0) shifted away from the interruption rate. Thus, this study measures the importance of different speech acoustic cues during interruption at different rates in order to determine how interruption may interact with speech cues.

The Hilbert transform (Bracewell, 1986; Hilbert, 1912) is a signal processing method that divides the acoustic signal into E and TFS components. The raw Hilbert envelope contains slow varying amplitude fluctuations corresponding to segmental and syllabic modulations of the speech signal, as well as fast fluctuations corresponding to the opening and closing of the vocal folds. Therefore, E contains temporal information on two timescales: low rate modulations corresponding to the traditional speech envelope, and high rate modulations corresponding to speech periodicity. For this study (as well as others, see Gallun and Souza, 2008; Hopkins and Moore, 2010), E is defined according to the entire Hilbert envelope, including the highest modulation rates. However, while E contains speech periodicity information, the TFS contains harmonic relationships important for pitch perception (e.g., Moore, 2004; Plack and Oxenham, 2005). TFS may therefore play an important role in tracking pitch information (i.e., fundamental frequency) during speech

^{a)}Author to whom correspondence should be addressed. Electronic mail: dfogerty@indiana.edu

perception. Therefore, speech periodicity information is carried by both E and TFS cues. It may be that interruption at periodicity rates interferes most with E modulations carrying speech periodicity information, while TFS frequency cues remain relatively preserved. Experiment 2 examines this possibility.

Both E and TFS properties are present across the frequency spectrum, although reduced phase locking at the higher frequencies may reduce the perception of TFS cues (Moore, 2008a). While Shannon *et al.* (1998) noted no difference in the division of frequency bands used for E processing, it may be that E modulations in certain frequency bands contribute more to perception than other frequency bands. Crouzet and Ainsworth (2001) discovered that the temporal envelopes contained in two different and distant frequency regions were only partially correlated with each other. Envelope modulations near the syllabic rate of speech (≤ 4 Hz) correlate across the spectrum, while faster modulations > 8 Hz were only correlated in adjacent bands. This finding suggests that these faster modulations may carry different speech information in different frequency regions, while speech information carried at slower rates is present across the spectrum. In addition, high-frequency bands have more mid-rate modulations (10–25 Hz) than low-frequency bands (Greenberg, Arai, and Silipo, 1998). Gallun and Souza (2008) also found different modulation spectra across frequency bands. They presented listeners with E-only speech that was processed to contain the full envelope modulation spectrum and was modulated by noise. Interestingly, the correlations between the modulation spectra of phonemes predicted consonant identification error rates for E-only speech. That is, phonemes that had highly correlated modulation spectra had higher confusion rates. Findings such as these led Apoux and Bacon (2004) to conclude that the shape of the modulation spectrum may vary across the frequency spectrum.

A few studies have begun to investigate the perceptual weighting of E-only speech across frequency regions (Apoux and Bacon, 2004; Kasturi *et al.*, 2002; Mehr *et al.*, 2001). One predominant method for measuring the perceptual importance of speech information is the correlational method (Berg, 1989; Lutfi, 1995; Richards and Zhu, 1994). The correlational method was developed and applied to speech by Doherty and Turner (1996) to determine how individual listeners perceptually weight individual acoustic components. In the general correlational method used by Doherty and Turner (1996), speech was divided into three frequency bands. On each trial, noise individually and randomly degraded each band. For example, the SNR within each of three bands on a given trial could be -7 , $+1$, and -1 dB. A listener's performance on each trial was then correlated with the SNR in each band. Relative weights for each band were then derived by normalizing the correlations across the three bands to 1.0. Relative weights were interpreted as indexing the proportion to which that band's weight contributed on each trial to the listener's speech understanding performance. Using this method, Doherty and Turner (1996) found the majority of listeners weighted the mid-frequency band (750–2500 Hz) the most for nonsense syllables.

To test whether envelope modulations (including high-rate periodicity < 500 Hz) within certain frequency bands contained more speech information than other frequency regions, Apoux and Bacon (2004) applied this correlational approach (Lutfi, 1994) to obtain frequency-weighting functions of speech-envelope modulated noise. Using four analysis bands and envelope modulations up to 500 Hz, they found that normal-hearing listeners applied equal perceptual weights to envelope modulations in all frequency bands when speech syllables (CV and VC) were presented in quiet. However, listeners placed more weight on the high-frequency band (> 2500 Hz) when speech was presented in noise.

Kasturi, Loizou, Dorman, and Spahr (2002) also estimated importance functions for envelope modulations in different frequency bands. Frequency importance functions were estimated using a least squares approach by removing one or two frequency bands, thereby creating "holes" in the spectrum. Their results demonstrated that vowel identification in "hVd" contexts was most impacted by removing bands containing F1 and F2 information. However, consonant identification in /aCa/ contexts remained relatively constant, regardless of the frequency band removed. The consonant results are similar to Apoux and Bacon (2004) who found flat correlational functions for speech in quiet. Mehr, Turner, and Parkinson (2001) also found flat correlational functions for normal-hearing listeners. However, weighting strategies for listeners with cochlear implants were unequal across frequency and across listeners, demonstrating large individual differences.

These results suggest that while the envelope modulation spectrum may vary across frequency bands, in quiet, listeners are able to use the modulation in each frequency band equally. This may reflect an importance of low-rate syllabic modulations around 4 Hz, that is, present across frequencies (Crouzet and Ainsworth, 2001). However, in the presence of interfering noise, perceptual weighting appears to shift. Greater emphasis appears to be placed on higher frequency bands (Apoux and Bacon, 2004) that contain faster modulation rates (predominantly 10–25 Hz, Greenberg *et al.*, 1998; Silipo, Greenberg, and Arai, 1999) and are not correlated with other spectral regions (Crouzet and Ainsworth, 2001).

While research has begun to investigate the perceptual use of E information across the spectrum, similar data for TFS cues is currently limited. Recently, Hopkins and Moore (2010) have investigated the spectral distributions of TFS cues more comprehensively. They incrementally added TFS to E-only speech starting first with the low-frequency bands and then starting with the high-frequency bands. The most improvement in intelligibility occurred when TFS was added to frequency bands under 1000 Hz, consistent with TFS serving as an important cue for coding F0 (Moore and Moore, 2003; Moore, Glasberg, Flanagan, and Adams, 2006). However, some improvement was also noted when adding TFS above 4000 Hz when no other TFS information was available, suggesting that even though phase-locking is reduced at these frequencies (Moore, 2008a), TFS cues, when presented in quiet, still provide important speech information. This suggests that TFS may contain partially redundant

information in multiple frequency regions. [Hopkins and Moore \(2010\)](#) then explored the importance of TFS acoustic information to the contribution of five independent frequency regions spanning the frequency range of 100–8000 Hz. They found that TFS significantly and equally contributed to speech recognition thresholds in each frequency band above what was provided by E information alone. Thus, TFS information is distributed across the spectrum and provides additional information from what is contained in the envelope. Furthermore, its contributions are not limited to low-frequency coding of the F0.

Recently, [Fogerty \(2010, 2011\)](#) reported the perceptual weighting functions of E and TFS information across frequency bands. During uninterrupted meaningful sentences, normal-hearing listeners weighted mid frequencies the most for E and TFS processing, with E cues weighted more than TFS across all frequency bands. The current study investigates the spectral distribution of E and TFS cues during interruption.

[Hopkins and Moore \(2009\)](#) have compared E and TFS cues and found that TFS is most important for masking release, particularly the TFS in high spectral frequency bands as it contributes to the intelligibility of speech during fluctuating noise, but not steady-state noise. Restoring low-frequency information for cochlear implant processing also aids in the intelligibility of speech in the presence of reversed or competing speech ([Kong and Carlyon, 2007](#); [Li and Loizou, 2008](#)), consistent with the contribution of TFS information. Thus, it appears that TFS cues are important for utilizing speech glimpses between masker fluctuations.

In addition to the increased importance of TFS during interruption is the finding that interruption may also selectively inhibit the transmission of the speech envelope. [Gilbert et al. \(2007\)](#) found that fluctuating noise decreases performance for E-only speech more so than TFS-only speech, specifically at low fluctuating rates under 8 Hz. They discussed these results in terms of the interference fluctuating noise causes for envelope cues. Like the speech envelope, fluctuating noise is an amplitude modulated signal. Modulating the noise at rates close to the natural modulation of speech interferes with the listener's ability to detect the salient envelope modulations of speech. This may be similar to the general psychoacoustic phenomenon of modulation detection interference (e.g., [Bacon and Grantham, 1989](#); [Yost and Sheft, 1989](#)).

Thus, while the importance of TFS cues increases during interrupted speech, the availability of E cues appears to decrease, particularly at rates matching the natural modulation of speech. This suggests that the perceptual weights of listeners will shift to more available or informative cues depending upon the speech context. This study explicitly examines how listeners' perceptual weighting of these cues is modified by listening to speech in interrupted contexts. Changes in relative weights likely reflect more automatic changes in the mode of perceptual processing determined by the availability and informativeness of speech cues.

Slow interruption rates near the syllabic rate of speech have been demonstrated to interfere with the use of speech envelope cues ([Gilbert et al., 2007](#)). Thus, it may be that lis-

teners will have to shift their perceptual weighting to facilitate speech recognition processes for interrupted speech. Specifically, it has been suggested that TFS cues are important and facilitate perception of interrupted speech. However, it is not known how the importance of these cues might compare to the use of preserved E information, or how TFS information might be distributed across the frequency spectrum. Experiment 1 investigates the spectral distribution of perceptual weights to E and TFS cues during slow, syllabic rate interruption of meaningful sentences. Experiment 2 explores the contribution of F0 information during interruption near the natural periodicity of speech.

II. EXPERIMENT 1: PERCEPTUAL WEIGHTING DURING SYLLABIC-RATE INTERRUPTION

Speech is commonly heard in the presence of some interfering sound, such as from competing talkers. Often, that sound may completely obliterate different temporal or spectral portions of the speech signal. In addition, with the rise in digital transmission of speech, often temporal portions may be lost completely during data transmission. Many listeners have increased difficulty understanding speech in such interrupted contexts, particularly those with hearing loss (e.g., [Bernstein and Grant, 2009](#); [Festen and Plomp, 1990](#); [Summers and Molis, 2004](#)). Assistive devices such as hearing aids do little to help in such contexts ([Peters, Moore, and Baer, 1998](#)) and may even compound the problem for these listeners ([Baskent et al., 2009](#); [Moore, 2008b](#); [Wang, 2008](#)). Experiment 1 investigated how listeners weight temporal modulations of the envelope and fine structure in an interrupted speech-in-noise context.

A number of studies have investigated this issue through use of an interrupted noise masker (e.g., [Gnasia et al., 2008](#); [Nelson et al., 2003](#)). By adding a high level fluctuating noise to the speech signal, periodic portions of speech are masked. Results have demonstrated that while normal-hearing listeners perform better with fluctuating noise in contrast to continuous noise, listeners who only receive the speech envelope have little benefit from hearing the glimpses of speech within the fluctuating masker dips ([Nelson and Jin, 2004](#)) and may even perform more poorly than with continuous masking ([Qin and Oxenham, 2003](#)). Therefore, several researchers have argued for the increased importance of fine structure acoustics for such interrupted contexts ([Lorenzi et al., 2006](#); [Hopkins and Moore, 2009](#)) in addition to suggesting how fluctuating maskers interfere with physical properties of the speech envelope, most predominantly at syllabic rates ([Gilbert et al., 2007](#)). While envelope cues alone appear sufficient for continuous speech intelligibility in quiet ([Shannon, et al., 1995](#)), these studies suggest that fine structure information plays a relatively larger role during interruption.

Studies using high level fluctuating noise to mask speech may actually overestimate perceptual contributions of certain speech acoustic cues to speech understanding. This is because interrupting noise that is above the speech level introduces top-down “filling-in” processes from auditory induction ([Sivonen, Maess, Lattner, and Friederici,](#)

2006). Periodically removing speech segments and inserting silence or low-level noise is a common method used in interruption studies that avoids such processes (e.g., Miller and Licklider, 1950). This allows for a more direct examination of how interruption influences the perceptual processes involved. Therefore, experiment 1 removes the speech signal at regular intervals and replaces it with a low level speech-shaped noise.

The purpose of experiment 1 was to measure the perceptual weighting of the envelope and fine structure across frequency bands for interrupted speech. An important goal of this investigation was to determine the relative weight of E and TFS cues when E cues are less available due to syllabic-rate interruption.

A. Listeners

Eight young normal-hearing listeners were paid to participate. All participants were native speakers of American English and had pure-tone thresholds bilaterally no greater than 20 dB hearing level at octave intervals from 250 to 8000 Hz (ANSI, 2004).

B. Stimuli and design

1. Signal processing

Meaningful sentences from the IEEE/Harvard database were selected for use in this study (IEEE, 1969). Sentences were previously recorded by a male talker and are available on CD ROM (Loizou, 2007). For processing, all experimental sentences were first down-sampled to a sampling rate of 16 000 Hz.

The task consisted of varying the SNR for two temporal components (i.e., E and TFS) in each of three frequency bands with cutoff frequencies at 80, 528, 1941, and 6400 Hz, resulting in six acoustic channels. These channels will be noted according to the temporal property and band number, as in E1 (i.e., E modulation in band 1). The three frequency bands selected represented equal distance along the basilar membrane using a cochlear map (Liberman, 1982), and correspond roughly to prosodic, sonorant, and obstruent/fricative linguistic categories respectively. The selected bands for this study have about the same number of ERBs (Moore and Glasberg, 1983) and similar Speech Intelligibility Index values.

A correlational method (Lutfi, 1995; Richards and Zhu, 1994) was used in this experiment which correlates the SNR value of each condition on a given trial to the response accuracy of the listener on that trial. In order to use this analysis method, the SNRs in each of the six acoustic channels were independently and randomly assigned. Therefore, on a given trial the same SNR level could be presented to more than one channel (i.e., a given trial could consist of SNRs presented at -1 , 2 , -1 , 11 , 5 , 2 distributed across the six channels).

The speech stimulus and matching speech-shaped noise were passed through a bank of band-pass filters and the noise level was scaled according to the SNR condition for that trial in each band and added to copies of the filtered speech. The

Hilbert transform then divided each combined speech and noise band into the E and TFS. For example, if the SNR in band 1 for E and TFS was -1 and 5 dB, respectively, two copies of the speech in band 1 were made. These were combined with filtered noise scaled to either -1 or 5 dB SNR and added to each band copy. The Hilbert transform was used to extract the E from -1 dB SNR band copy and the TFS from the 5 dB band copy. The noisy Hilbert components, masked at different SNRs, were then recombined, refiltered at 6400 Hz and up-sampled to 48 828 Hz for presentation through the TDT system III hardware.

Processing into each of these different acoustic channels was performed in MATLAB using custom modification of software designed by Smith *et al.* (2002). The noise used for processing contained the same power spectrum as the original sentence. This ensured that the same SNR was maintained, on average, across all frequency components for each sentence. Thus, the SNR in each channel reflects a true SNR calibrated to the individual stimulus trial. Therefore, each sentence was combined with a unique noise matching its own average long term speech spectrum. SNR values ranged from -1 dB to 11 dB in 3 dB steps, resulting in five different levels. Each SNR was presented in each acoustic channel a total of 120 trials (120×5 SNRs = 600 total trials). Each sentence contained five key words, resulting in 3000 key-words scored. These processed sentences were then subjected to interruption by a noise stimulus.

2. Noise processing

A speech-shaped noise was used for the interruption and was matched to the long-term average speech spectrum for a concatenation of all 600 experimental sentences (within ± 2 dB). A unique noise was used for each replacement within the sentence. The noise level was set to a SNR of 16 dB, a level used in previous noise interruption studies (Fogerty and Kewley-Port, 2009; Kewley-Port *et al.*, 2007) to avoid any possibility of temporally masking consonants or introducing top-down restoration processes. The noise replacement technique used during interruption consisted of removing the speech information within the interruption interval and replacing it with this speech-shaped noise. Thus, during interruption, no speech information was available for perceptual use. To avoid introduction of transients, replacement boundaries were adjusted within a 2 ms window to the nearest amplitude local minimum (i.e., zero-crossing).

A slow rate of interruption, 4 Hz, is near the natural speech amplitude modulation for these sentences (broadband peak modulation was 4.2 Hz for these sentences) and was used to periodically interrupt the speech signal. For each 250 ms of the stimulus, representing one duty cycle for the interruption, the speech was preserved for 75% of the cycle while the noise replaced the speech stimulus during the remaining 25% . This 75% duty cycle was used because these sentences were already highly degraded and the correlational method used required that a listener's performance was above the floor for adequate variability. A larger proportion of noise during each cycle would have greatly reduced overall performance. Stimuli were refiltered at 6400 Hz after this

a) "Dispense with a vest on a day like this."

b) "It's a dense crowd in two distinct ways."

FIG. 1. Waveform and spectrograms of stimulus examples interrupted at a rate of 4 Hz: (a) Preserving mostly TFS acoustics (E: -4 , -4 , -7 dB; TFS: 5, 2, 5 dB), (b) preserving mostly E acoustics (E: 5, 2, 2 dB; TFS: -4 , -4 , -7 dB). Note that the narrow light bars in the spectrogram correspond to the interrupting speech-shaped noise.

interruption processing using a six-order butterworth filter to remove any residual transients. Figure 1 displays examples of these stimuli.

C. Calibration

Stimuli were presented via Tucker–Davis Technologies (TDT) System III hardware using 16-bit resolution at a sampling frequency of 48828 Hz. The output of the TDT D/A converter was passed through a headphone buffer (HB-7) and then to an ER-3A insert earphone. A speech-shaped noise matching a concatenation of 600 experimental sentences was set to 70 dB sound pressure level through the insert earphone in a HA-2 2-cm³ coupler using a Larson Davis model 2800 sound level meter with linear weighting. Therefore, the original unprocessed wideband sentences were calibrated to be presented at 70 dB sound pressure level. However, after filtering and noise masking the overall sound level of the combined stimulus varied according to the individual condition. This ensured that levels representative of typical conversational level (70 dB sound pressure level) were maintained for the individual spectral components of speech across all sentences.

D. Procedure

Listeners were seated individually in a sound attenuating booth and listened to the stimulus sentences presented unilaterally to their right ear. Test stimuli were controlled by TDT system III hardware connected to a PC computer running a MATLAB stimulus presentation interface. Each listener was instructed regarding the task using verbal and written instruction. All listeners completed familiarization trials prior to each experimental test (15 sentences). These sentences were processed according to the stimulus processing procedures for that task. All listeners completed experimental testing regardless of performance on the familiarization

tasks. No feedback was provided during familiarization or testing. All listeners received different randomizations of the sentences. All sentences were only presented once to each listener.

The listeners' task was to repeat aloud as accurately as possible the sentence that was presented. Responses were digitally recorded for offline analysis. Three native English speakers were trained to serve as raters and scored all recorded responses. Inter-rater agreement on a 10% sample of responses was 93%.

E. Results

Listeners were tested on the main correlational task during two different sessions. Average performance on each day was calculated to determine if learning occurred between the two days. No significant difference between performance on each day was observed ($p > 0.05$).

Overall, listeners only improved performance by 1.8 percentage points on the second session. Therefore, data from both sessions were pooled for analysis.

The main experimental task investigated the relative weighting of envelope and fine structure modulations across three frequency bands during interruption at a rate of 4 Hz. Figure 2 displays performance across the five SNRs tested for each acoustic channel. Of particular interest was how each of these six acoustic channels is weighted when speech information is simultaneously available from multiple channels. To determine the relative weighting of each acoustic channel, a point-biserial correlation (Lutfi, 1995) was calculated trial-by-trial between the word score and the SNR in each channel (Table I). Each correlation was calculated over 3000 points (600 sentences $\times$ 5 keywords). Raw correlations (obtained across all 5 keywords) for each listener were tested against the null hypothesis of zero. Of these correlations, 85% were significantly different from zero ($p < 0.05$) and

FIG. 2. Mean keyword performance vs SNR for each acoustic channel during 4 Hz interruption.

are indicated in bold in Table I. On average, TFS3 (i.e., the TFS information in Band 3) did not differ significantly from zero, indicating that this channel was not used by listeners. All other channels received some weight. Correlations were then normalized to 1.0 to reflect the relative perceptual weight assigned to each acoustic channel (see Fig. 3).

A repeated measures analysis of variance was conducted over the normalized weights for keyword scoring (Fig. 3). A main effect for temporal modulation [$F(1,7)=98.9$, $p<0.001$] was found along with an interaction between modulation and frequency band [$F(2,14)=31.4$, $p<0.001$]. Overall, listeners weighted E contributions more than TFS contributions (an average difference of 0.09) and band 3 more than either of the other two bands (by a difference of 0.13). *Post hoc t*-tests demonstrated significant differences between E and TFS for low- and high-frequency bands (Bonferroni-corrected $p<0.006$), but no difference for band 2 ($p>0.05$). E3 was significantly different from E1 and E2, receiving more perceptual weight (Bonferroni-corrected $p<0.006$). Analysis of sentence scores, where all keywords within the sentence were required to be correct, revealed the same analysis results with the exception of no difference obtained between E1 and TFS1 weights ($p>0.05$). Overall, most listeners were consistent with the average configuration of perceptual weights.

TABLE I. Raw point-biserial correlation coefficients for each listener across each of the six channels in experiment 1. The listener's performance was correlated on a trial-by-trial basis with the signal-to-noise ratio in each channel for that trial. Correlations in bold were significantly different from zero ($p<0.05$). Overall percent correct scores across all trials are also listed for each listener.

Listener	Avg. (%)	E1	E2	E3	TFS1	TFS2	TFS3
M4_01	74	0.097	0.053	0.147	0.042	0.063	0.038
M4_02	48	0.104	0.102	0.140	0.021	0.061	-0.005
M4_03	61	0.106	0.059	0.102	0.073	0.079	0.001
M4_04	43	0.082	0.077	0.149	0.019	0.039	0.065
M4_05	57	0.098	0.141	0.170	0.069	0.054	0.030
M4_06	59	0.072	0.069	0.136	0.010	0.065	0.048
M4_07	57	0.068	0.104	0.165	0.042	0.093	0.017
M4_08	58	0.058	0.066	0.177	0.048	0.052	0.063
MEAN	57	0.086	0.084	0.148	0.041	0.063	0.032

FIG. 3. Relative perceptual weights for keyword scoring during 4 Hz interruption obtained by using a point-biserial correlational analysis across the six acoustic channels. Error bars = standard error of the mean. The gray dotted line indicates the reference value if all channels were equally weighted.

The effect of sentence context was also examined to determine if listeners dynamically alter their perceptual weighting across the sentence. Relative weights were calculated over the first two keywords in the sentence and compared to relative weights obtained from the last two keywords, which would have had more sentence context available during processing. Results indicated significant differences only for E1 (Bonferroni-corrected $p<0.008$). Less weight was given to E1 acoustics in the second half of the sentence. The distribution of weights over the second half of the sentence was most reflective of the overall relative weighting configuration. Performance was better on the first half of the sentence compared to the second half for all listeners by an average of 3.4 percentage points ($p<0.01$). Therefore, changing acoustic cues and not contextual predictability appears to be the more likely explanation for this change in perceptual weighting.

F. Discussion

During 4-Hz interruption, listeners overwhelmingly placed most of their perceptual weight on E cues in the high-frequency band, while placing little to no weight on TFS cues in the low- and high-frequency regions. This was not expected. Instead, it was predicted that listeners would place more perceptual weight on band 1, as F0 is assumed to be an important continuity cue during interruption (e.g., Bregman *et al.*, 1990; Darwin *et al.*, 2003). While it was expected that E cues would be adversely impacted by 4-Hz interruption, the overwhelming dominance of relative weight placed on E3 suggests that some E information remains preserved. As the low interruption rate interferes with speech modulation at that same rate (Gilbert *et al.*, 2007), other amplitude modulation rates likely become more important for understanding speech. Higher frequencies (4762–6000 Hz) have additional amplitude modulation information (mid rate 10–25 Hz) that is not present in the lower frequencies (Greenberg *et al.*, 1998). This mid rate modulation may correlate with the segmental rate of speech. Fogerty and Kewley-Port (2009) measured the average duration of vowels and consonants in TIMIT sentences (Garofolo, *et al.*, 1990). These durations correspond to segmental rates of 10 Hz for vowels

and 16 Hz for consonants. The faster rate of consonants helps to explain how band 3, which contains high frequency consonantal noise bursts and frication, may be important for carrying this segmental rate information. This mid-rate modulation would be more resistant to any modulation “masking” caused by the slow 4-Hz interruption. Therefore, envelope modulation in the high frequency band appears to be especially important in understanding speech interrupted at this rate. These results are also consistent with [Apoux and Bacon \(2004\)](#) who found that listeners place more weight, in noise, on E in spectral frequencies greater than 2400 Hz. This appears to contradict [Hopkins and Moore \(2010\)](#) who found little benefit for E during uninterrupted speech in frequency bands above 2000 Hz as compared to when the E was removed by a spectral notch. Both of these studies included low- and high-rate E modulations. Apparently, E cues in this high-frequency region take on new importance when E cues in other frequency regions are no longer available, such as due to modulation masking in this study. In the study by [Hopkins and Moore \(2010\)](#), E cues were still available in the mid-frequency region, and it may be that E cues above 2000 Hz did not provide any additional information beyond what was already available.

No significant weight was placed on TFS3 during interruption. This is likely an effective strategy, as neural encoding of fine structure begins to degrade in these higher frequencies ([Moore, 2008a](#)). It is also consistent with other behavioral work demonstrating reduced perceptual use of TFS above 2000 Hz ([Ardoint and Lorenzi, 2010](#)). However, TFS is believed to be an important cue for glimpsing speech during interruption because TFS cues provide a benefit in intelligibility for modulated maskers more than steady-state maskers and is correlated with masking release (e.g., [Lorenzi et al., 2006](#); [Hopkins and Moore, 2009](#)). This appears to be at odds with the current investigation. If anything, it appears that less weight was placed on TFS acoustics during interruption. Particularly contradictory are results of TFS in the low frequencies, which appears to facilitate speech recognition during reversed speech masking or competing speech ([Kong and Carlyon, 2007](#); [Li and Loizou, 2008](#)), yet this channel received hardly any perceptual weight in the current study.

One explanation may be that the relative perceptual weights were obtained concurrently in this study. Therefore, a low relative weight, such as obtained for TFS1, does not necessarily indicate that a particular channel does not contribute to speech recognition. In addition, a high relative weight does not necessarily indicate that a particular channel, when isolated from other channels, would result in high word recognition scores. It may be that speech information from other channels, such as TFS1, even when highly masked at the lowest SNRs, contributes to the high perceptual weight of a single channel by disambiguating cues. Therefore, these relative weights might not predict individual channel contributions. Rather, they appear to reflect the overall distribution of perceptual weight when multiple channels of speech information are available.

Overall, syllabic-rate interruption resulted in significantly higher relative perceptual weight placed upon the

envelope in the high-frequency band. In contrast, the relative weighting of fine structure cues demonstrated the importance of the mid-frequency band that contains dynamic formant movement.

III. EXPERIMENT 2: PERCEPTUAL WEIGHTING DURING PERIODICITY-RATE INTERRUPTION

The importance of F0 is increased in interrupted contexts as it is believed to facilitate sound segregation ([Bregman, Liao, and Levitan, 1990](#)), particularly during competing talkers ([Brokx and Nooteboom, 1982](#); [Darwin, Brungart, and Simpson, 2003](#)). High-rate periodicity information contained in the Hilbert envelope ([Rosen, 1992](#)) and harmonic resolution provided by the TFS ([Moore et al., 2006](#)) appear to be involved in coding this information. Providing additional TFS to low-frequency bands in vocoded speech increases the perception of low-context sentences ([Hopkins and Moore, 2010](#)) and facilitates glimpsing during competing speech ([Li and Loizou, 2008](#)). Therefore, there appears to be evidence to support the importance of both E and TFS modulation cues in the low-frequency band during interruption.

However, if interruption interferes with E modulations at the interruption rate, then interruption at the periodic rate of speech should limit the contribution of fast-rate E cues that provide information about F0. By shifting the F0 higher relative to the interruption rate, the perceptual use of these fast-rate E cues should be relatively restored. Experiment 2 measures the relative perceptual weight assigned to E and TFS cues in each of these conditions: (1) when interruption occurs at the periodicity of speech and (2) when the interruption rate is maintained at this natural periodicity rate, but the F0 is shifted in frequency to avoid interruption interference.

A. Listeners

Two groups of listeners participated in Experiment 2 for testing perceptual weighting during 128 Hz interruption. One group completed testing at the speaker’s natural fundamental frequency (131 Hz), while the second group completed testing with the fundamental frequency shifted by eight semitones (208 Hz). These groups of listeners will be referred to as M128 and F128 respectively. Eight young normal-hearing listeners participated in each group, for a total of 16 listeners (mean age = 22 yr). All participants were native speakers of American English and had pure-tone thresholds bilaterally no greater than 20 dB hearing level at octave intervals from 250 to 8000 Hz ([ANSI, 2004](#)). No listeners previously participated in experiment 1.

B. Stimuli and procedure

The same IEEE sentences were used as in experiment 1. Processing of the sentences was identical, with two exceptions. First, in order to maintain performance levels, acoustic channels were processed over an SNR range of –4 dB to 8 dB in 3 dB steps. Second, noise interruption was introduced

at a rate of 128 Hz, instead of the 4 Hz interruption rate used previously. The interruption noise and processing of the acoustic channels remained identical to experiment 1.

For F128 (i.e., the F0 shifted group), sentences underwent one additional processing step that was conducted prior to any filtering or noise processing. All sentences were individually processed in STRAIGHT (Kawahara, Masuda-Kastuse, and Cheveigne, 1999) to extract the F0 contour. This natural contour was shifted up in frequency by eight semitones. The sentence was then resynthesized with the new F0 contour. Thus, these shifted sentences remained identical to the original with the exception of the higher F0. No further adjustments, such as for vocal tract size, were made on the stimuli. These resynthesized sentences were then subjected to the processing described above.

Testing procedure was identical to experiment 1. Listeners were seated in a sound attenuating booth and listened to sentences presented monaurally to their right ear via an ER-3A insert earphone. Listeners completed familiarization trials prior to the experimental task. Responses were audio recorded and scored. Two raters scored all responses with an inter-rater agreement of 98.6% on 10% of responses.

C. Results and discussion

1. Group performance

As in experiment 1, overall performance between the two test sessions was examined. Performance was not significantly different between the two sessions for either listener group, with an average difference of 2.4 and 1.4 percentage points, indicating learning did not occur between the two sessions.

TABLE II. Raw point-biserial correlation coefficients for each listener across each of the six channels for the M128 group (top) and F128 group (bottom). The listener’s performance was correlated on a trial-by-trial basis with the signal-to-noise ratio in each channel for that trial. Correlations in bold were significantly different from zero ($p < 0.05$). Overall percent correct scores across all trials are also listed for each listener.

Listener	Avg. (%)	E1	E2	E3	TFS1	TFS2	TFS3
M128_01	74	0.131	0.164	0.199	0.049	0.070	0.076
M128_02	80	0.085	0.120	0.164	0.029	0.056	0.067
M128_04	54	0.098	0.113	0.146	0.056	0.043	0.028
M128_05	68	0.151	0.142	0.192	0.072	0.087	0.065
M128_06	55	0.203	0.160	0.175	0.108	0.014	0.043
M128_07	78	0.142	0.112	0.144	0.057	0.103	0.081
M128_08	71	0.156	0.174	0.185	0.103	0.096	0.119
M128_09	79	0.104	0.091	0.153	0.043	0.088	0.044
MEAN	70	0.134	0.135	0.170	0.065	0.070	0.065
F128_01	62	0.091	0.127	0.143	0.066	0.082	0.065
F128_02	55	0.064	0.112	0.167	0.057	0.091	0.082
F128_03	65	0.089	0.093	0.144	0.052	0.100	0.063
F128_04	62	0.088	0.060	0.145	0.109	0.063	0.067
F128_05	69	0.050	0.082	0.140	0.024	0.095	0.066
F128_06	53	0.144	0.119	0.194	0.052	0.094	0.059
F128_07	58	0.109	0.131	0.181	0.101	0.106	0.082
F128_08	64	0.114	0.134	0.133	0.057	0.118	0.079
MEAN	61	0.094	0.107	0.156	0.065	0.094	0.070

The main experimental task investigated the relative weighting of envelope and fine structure modulations across three frequency bands during interruption at a rate of 128 Hz for two listener groups who heard the spoken sentences with different fundamental frequency: at the natural F0 near 128 Hz (i.e., 131 Hz) and at a F0 shifted higher in frequency by eight semitones (i.e., 208 Hz). Raw correlations (obtained across all five keywords) for each listener were tested against the null hypothesis of zero. Of these correlations, 94 and 98% were significantly different from zero ($p < 0.05$) for the M128 and F128 groups respectively. These are indicated in bold in Table II. All channels received some weight. As done previously, the relative weighting of each acoustic channel was calculated using a point-biserial correlation (Lutfi, 1995) over 3000 points (600 sentences $\times$ 5 keywords) and then normalized to 1.0 to reflect the relative perceptual weight assigned to each acoustic channel (see Fig. 4). Analyses were first conducted separately for each listener group.

A repeated measures analysis of variance was conducted over the normalized weights for keyword scoring separately for each listener group. For M128, a main effect for temporal modulation [$F(1,7) = 105.2$, $p < 0.001$] was found along with an interaction between modulation and frequency band [$F(2,14) = 5.3$, $p < 0.05$]. Similarly, F128 had main effects

FIG. 4. Relative perceptual weights based on keyword scoring during 128 Hz interruption obtained by using a point-biserial correlational analysis across the six acoustic channels. (a) M128 group, (b) F128 group. Error bars = standard error of the mean. The gray dotted line indicates the reference value if all channels were equally weighted.

for temporal modulation [$F(1,7) = 57.7, p < 0.001$] and frequency band [$F(1,7) = 7.7, p < 0.01$], along with an interaction between modulation and frequency band [$F(2,14) = 22.2, p < 0.001$]. While main effects revealed similar results for these two listener groups, *post hoc* comparisons revealed different weight patterns and are discussed separately here.

a. M128: Natural F0. For the natural F0 group, M128, *post hoc t*-tests demonstrated significant differences between E and TFS across all frequency bands (Bonferroni-corrected $p < 0.006$). E3 was significantly different from E2, receiving more perceptual weight [Bonferroni-corrected, $t(7) = -4.9, p = 0.002$]. No significant difference was obtained between E1 and E3 [Bonferroni-corrected, $t(7) = -2.9, p = 0.023$]. This is largely due to the greater between-subject variability for E1, as mean weighting for E1 and E2 was similar (0.208 vs 0.211, respectively). Analysis of sentence scores, where all keywords were required to be correct, revealed the same analysis results with the exception of no difference obtained between E2 and TFS2 weights ($p > 0.05$).

The effect of sentence context was also examined. Relative weights were calculated over the first two keywords in the sentence and compared to relative weights obtained from the last two keywords, which would have had more sentence context available during processing. Results indicated significant differences only for E1 and TFS1 (Bonferroni-corrected $p < 0.008$). As found in experiment 1 with 4 Hz interruption, less weight was given to E1 acoustics in the second half of the sentence. For TFS1, more weight was given to this channel in the second half of the sentence. As found previously, the weighting strategy over the second half of the sentence was most reflective of the overall weighting strategy. Again, performance was better on the first half of the sentence compared to the second half for all listeners by an average of 10.6 percentage points ($p < 0.001$). Therefore, contextual predictability does not explain this change in perceptual weighting. Dynamic tracking of acoustic cues may be more likely.

b. F128: Shifted F0. For the shifted F0 group, F128, *post hoc t*-tests demonstrated significant differences between E and TFS only for band 3 [Bonferroni-corrected $t(7) = 12.6, p < 0.006$]. E3 was significantly different from E1 and E2, receiving more perceptual weight (Bonferroni-corrected, $p < 0.006$). TFS3 was also weighted significantly less than TFS2 [Bonferroni-corrected $t(7) = 4.2, p < 0.006$]. Analysis of sentence scores revealed the same analysis results with the exception of no difference obtained between TFS2 and TFS3 weights (Bonferroni-corrected $p > 0.006$).

Analysis of sentence context revealed significant differences for E2, E3, and TFS2 (Bonferroni-corrected $p < 0.008$). More weight was placed on E2 and TFS2 in the second half of the sentence, while less weight was placed on E3. As found previously, the weighting strategy over the second half of the sentence was most reflective of the overall weighting strategy. Again, performance was best on the first half of the sentence for all listeners by an average of 13.0 percentage points ($p < 0.001$).

2. Correlations

For M128, E1 and TFS1 were significantly correlated, $r = 0.77$, as well as E2 and TFS2, $r = -0.77$. Note that the correlation in band 2 was negative, indicating that listeners who placed more weight on E2, weighted TFS2 less, and vice versa. This may suggest that E2 and TFS2 were redundant and that listeners need to only process speech information in one of these channels. On the other hand, in band 1 the correlation was positively correlated, suggesting that all information in this low-frequency band was weighted equally, possibly indicating that E1 and TFS1 provide complementary cues. Also of note was that average performance across all trials was correlated with TFS2, $r = 0.76$, indicating that those who were more successful at repeating the sentences correctly, placed more weight on TFS2. This is interesting as this group as a whole placed less weight on TFS2 than either of the two other listener groups tested in this study. Therefore, perhaps lower TFS2 weights for the M128 group reflects a group-specific pattern.

For F128, these within band correlations disappeared. Instead, correlations between bands were found for E1 versus TFS3, $r = -0.91$, and TFS1 versus TFS2, $r = -0.77$. Overall performance was not correlated with relative weights, indicating that perhaps there was not one best weighting strategy for this condition.

3. Comparison across interruption conditions

Three groups of listeners were tested across the two experiments of this study. These listeners heard sentences interrupted at the syllabic rate of 4 Hz (M4), the periodic rate of 128 Hz (M128), or 128 Hz interruption with the speech periodicity shifted to 208 Hz (F128). Of interest in this study was how the relative weights of listeners differed between these three conditions. Independent-samples *t*-tests demonstrated no significant differences on relative weights placed on individual channels between any of the three listener groups tested here (Bonferroni-corrected $p > 0.008$). However, trends indicated less weight placed on E3 when moving to a higher interruption rate [M4 vs M128, $t(14) = 3.0, p = 0.010$] and for more weight placed on E1 for the shifted F0 when interrupted at 128 Hz [M128 vs F128, $t(14) = -2.831, p = 0.013$]. These trends were also marginally non-significant for sentence scoring. Therefore, it is possible that E3 is most beneficial during slow interruption rates. In addition, fast interruption rates matching the periodicity of speech may interfere with E1 cues.

The relative channel weights sum to one, which precluded examining main effects of the listener groups. Therefore a mixed repeated-measure analysis of variance with listener group as the between measure was conducted on the raw correlational weights obtained to determine if any group differences existed. Comparisons across the three groups on these raw point biserial correlational weights demonstrated a significant main effect with Geenhouse–Geisser correction for temporal modulation [$F(1, 21) = 210.0, p < 0.001$], frequency band [$F(2,42) = 12.8, p = 0.01$], and interruption group [$F(2, 21) = 9.1, p = 0.001$]. *Post hoc* comparisons using a Bonferroni-corrected $\alpha = 0.008$ demonstrated

significant differences on these raw correlational weights for M4 vs M128 on E1 [$t(14)=-3.2$, $p<0.008$] and E2 [$t(14)=-3.4$, $p<0.008$] and for M4 vs F128 on TFS2 [$t(14)=3.7$, $p<0.008$] and TFS3 [$t(14)=3.9$, $p<0.008$]. No significant differences were obtained between the two conditions with 128 Hz interruption. When compared to M4, higher raw TFS correlations were obtained for F128 and higher raw E correlations were obtained for M128.

IV. GENERAL DISCUSSION

In order to understand how interruption influenced listeners' perceptual weights, a comparison was made with the mean weighting strategy of ten normal-hearing listeners for identical stimulus processing using the same frequency bands during uninterrupted IEEE sentences (Fogerty, 2010, 2011). Figure 5 displays the relative weights for each of the three interruption conditions tested here and this continuous condition. Panel (a) compares E weights while panel (b) compares TFS weights. *Post hoc t*-tests were conducted to compare the current results to the uninterrupted speech case. Table III summarizes comparisons of each of the interruption conditions described here to the perceptual weighting of uninterrupted speech. The overall pattern of results suggests that E3 is essential for interrupted speech contexts. Difficulty

FIG. 5. Comparison of mean relative weighting strategies across all interruption conditions and the uninterrupted condition from Fogerty (2010): (a) comparison of envelope (E) weights across frequency bands, (b) comparison of fine structure (TFS) weights across frequency bands. Error bars = standard error of the mean.

TABLE III. Comparison of relative perceptual weights between the three interruption groups and the uninterrupted condition tested by Fogerty (2010). Channels with significant differences compared to the uninterrupted condition are reported here along with the direction of the significant difference.

Task		Channel weight differences		
M4	E2 < ^a	E3 > ^b		
M128		E3 > ^b	TFS2 < ^a	
F128	E2 < ^b	E3 > ^b		TFS3 > ^a

^a $p < 0.01$.
^b $p < 0.0001$.

hearing high frequency envelope cues may result in the speech understanding problems associated with these types of listening environments. The relative increase in E3 is accompanied by a decrease in the use of E2, except for M128, which also demonstrated a trend in this regard. In addition, TFS information remained relatively preserved during all interruption conditions, although M128 demonstrated reduced importance of TFS2. However, M4 listeners who had higher TFS2 weights had higher accuracy scores. Thus, listeners appear to have different modes of processing speech, determined by the availability of certain informative cues.

Interruption dramatically altered the relative perceptual weights obtained for the Hilbert envelope. A dramatic increase in envelope weighting in the high spectral frequency band was obtained relative to the mid-frequency band. Significantly, 4 Hz interruption only appeared to influence the relative weighting of envelope cues, while the relative weighting of fine structure cues remained constant. These results suggest that listeners do rely on different acoustic cues when speech is interrupted as compared to being continuously available. Therefore, the relative perceptual weighting of listeners is a dynamic process that is dependent upon the listening context. The success of assistive listening devices may therefore be enhanced if processing algorithms are designed to track these dynamic perceptual changes that are dependent upon the listening environment.

Relative TFS weights appear to remain relatively stable between uninterrupted and interrupted listening conditions. Thus, while TFS during interruption was weighted lower relative to E cues, the perceptual weighting obtained in the current study appears to suggest that TFS cues are not adversely influenced by interruption. Rather, the relative perceptual weights may remain preserved. Therefore, TFS may become a reliable cue under certain conditions, especially when the temporal properties of the interfering source are changing or more completely mask the entire speech modulation spectrum, such as in the case of background babble or of a competing talker. In such cases, even E modulation in the high-frequency band might not be available and the listener will need to rely more predominantly on TFS information. Therefore, even though listeners did not alter their relative weights to favor TFS information, their perceptual weights do reflect the continued availability of this information, specifically in the mid-frequency region.

This pattern of results highlights an important consideration for speech processing technologies. TFS cues in the

mid-frequency region appear to contribute most to speech understanding (relative to TFS in other frequency regions) and are modulated by the ability to process acoustic information in other acoustic channels. This was true across interruption conditions, except when the interruption rate matched the speech periodicity rate for M128, yet even in this condition, higher TFS2 weights resulted in better performance among the listeners. Consistent with [Ardoint and Lorenzi \(2010\)](#), TFS modulations in this band appear to convey distinct phonetic information. These modulations reflect formant movement, particularly of F2 in this study, which is important for vowel identification and place cues. Indeed, a number of studies have observed that while E modulations provide cues for voicing and manner, TFS is most important for conveying place information ([Apoux and Bacon, 2004](#); [Gallun and Souza, 2008](#); [Rosen, 1992](#); [Shannon et al., 1995](#)). Perhaps interruption matching the speech periodicity somehow distorts these cues. Further research is required to examine how interruption interacts with E and TFS cues, and what speech information is most adversely impacted by such distortions.

Relatively few differences were observed between 4 and 128 Hz interruption, or 128 Hz interruption with different F0s. One possibility for this overall homogeneity between interruption conditions is that F0 cues were not essential for glimpsing speech when using noise interruption. Studies that have demonstrated the importance of F0 and low frequency information have typically used competing or reversed speech (e.g., [Brokx and Nootboom, 1982](#); [Darwin et al., 2003](#); [Kong and Carlyon, 2007](#); [Li and Loizou, 2008](#)). Such conditions use F0 for talker segregation ([Bregman et al., 1990](#)). A different weighting pattern may indeed have been observed if competing speech instead of noise had been used during interruption, as F0 cues, and perhaps TFS in general, may have had increased importance for successful glimpsing.

One final note of consideration involves the interpretation of the relative perceptual weights obtained in this study. These weights were obtained when multiple acoustic channels were available for perception. Thus, it is likely that an interaction of acoustic speech information occurred. Certain channels may have provided acoustic cues that were insufficient for speech recognition alone, but contributed to the use of other acoustic channels, even when masked at highly unfavorable SNRs. Preliminary work in our lab suggests that E or TFS cues presented at a poorer SNR than those used in this study are not sufficient for word recognition in isolation, but lead to large improvements in recognition scores when presented concurrently with other speech information. Therefore, these relative weights appear to reflect the overall distribution of channel importance when information from multiple channels is available, which is the typical case for everyday listening.

V. CONCLUSIONS

This study investigated how listeners perceptually weight envelope and fine structure acoustics across the frequency spectrum during interruption at the syllabic-rate and

periodic-rate of speech. By varying the signal-to-noise ratio in each channel independently, an estimate of the relative perceptual weight a listener placed on that channel was obtained for each interruption condition. Overall, the results demonstrated similar relative weighting patterns across the interruption conditions, with the majority of weight placed on the high-frequency band envelope. The relative weighting of temporal fine structure, although lower than envelope weights, appears to remain stable during interruption, which may indicate that it is a reliable cue when listening to speech in such contexts.

ACKNOWLEDGMENTS

The author would like to thank Larry Humes, Diane Kewley-Port, David Pisoni, and Jennifer Lentz, who were instrumental in supporting and offering comments regarding this work. Kate Giesen, Eleanor Barlow, and Megan Chaney provided valuable laboratory assistance. A portion of this work was conducted as part of a dissertation completed at Indiana University and was funded, in part, by NIA Grant No. R01 AG008293 awarded to Larry Humes.

- ANSI (2004). "Specifications for audiometers," ANSI S3.6-2004, American National Standards Institute, New York.
- Apoux, F., and Bacon, S. P. (2004). "Relative importance of temporal information in various frequency regions for consonant identification in quiet and in noise," *J. Acoust. Soc. Am.* **116**, 1671–1680.
- Ardoint, M., and Lorenzi, C. (2010). "Effects of lowpass and highpass filtering on the intelligibility of speech based on temporal fine structure or envelope cues," *Hear. Res.* **260**, 89–95.
- Bacon, S. P., and Grantham, D. W. (1989). "Modulation masking: Effects of modulation frequency, depth, and phase," *J. Acoust. Soc. Am.* **85**, 2575–2580.
- Baskent, D., Eiler, C., and Edwards, B. (2009). "Effects of envelope discontinuities on perceptual restoration of amplitude-compressed speech," *J. Acoust. Soc. Am.* **125**, 3995–4005.
- Berg, B. G. (1989). "Analysis of weights in multiple observation tasks," *J. Acoust. Soc. Am.* **86**, 1743–1746.
- Bernstein, J. G., and Grant, K. W. (2009). "Auditory and audio-visual intelligibility of speech in fluctuating maskers for normal-hearing and hearing-impaired listeners," *J. Acoust. Soc. Am.* **125**, 3358–3372.
- Bracewell, R. N. (1986). *The Fourier Transform and its Applications*, 2nd ed. (McGraw Hill, New York), pp. 1–486.
- Bregman, A. S., Liao, C., and Levitan, R. (1990). "Auditory grouping based on fundamental frequency and formant peak frequency," *Can. J. Psych.* **44**, 400–413.
- Brokx, J. P. L., and Nootboom, S. G. (1982). "Intonation and the perception of simultaneous voices," *J. Phonetics* **10**, 23–26.
- Cooke, M. P. (2003). "Glimpsing speech," *J. Phonetics* **31**, 579–584.
- Crouzet, O., and Ainsworth, W. A. (2001). "Envelope information in speech processing: acoustic-phonetic analysis vs auditory figure-ground segregation," *EUROSPEECH-2001*, pp. 477–480.
- Darwin, C. J., Brungart, D. S., and Simpson, B. D. (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.
- Doherty, K. A., and Turner, C. W. (1996). "Use of a correlational method to estimate a listener's weighting function for speech," *J. Acoust. Soc. Am.* **100**, 3769–3773.
- Festen, J. M., and Plomp, R. (1990). "Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing," *J. Acoust. Soc. Am.* **88**, 1725–1736.
- Fogerty, D. (2010). "Perceptual weighting of the envelope and fine structure across frequency bands during continuous and interrupted speech" Dissertation, IN University.
- Fogerty, D. (2011). "Perceptual weighting of individual and concurrent cues for sentence intelligibility: Frequency, envelope, and fine structure," *J. Acoust. Soc. Am.* **129**, 977–988.

- Fogerty, D., and Kewley-Port, D. (2009). "Perceptual contributions of the consonant-vowel boundary to sentence intelligibility," *J. Acoust. Soc. Am.* **126**, 847–857.
- Gallun, F., and Souza, P. (2008). "Exploring the role of the modulation spectrum in phoneme recognition," *Ear Hear.* **29**, 800–813.
- Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallett, D., and Dahlgren, N. (1990). "DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CD-ROM," National Institute of Standards and Technology, NTIS Order No. PB91-505065.
- Gilbert, G., Bergeras, I., Voillery, D., and Lorenzi, C. (2007). "Effects of periodic interruptions on the intelligibility of speech based on temporal fine-structure or envelope cues," *J. Acoust. Soc. Am.* **122**, 1336–1339.
- Gnasia, D., Jourdes, V., and Lorenzi, C. (2008). "Effect of masker modulation depth on speech masking release," *Hear. Res.* **239**, 60–68.
- Greenberg, S., Arai, T., and Silipo, R. (1998). "Speech intelligibility derived from exceedingly sparse spectral information," *Proceedings of the International Conference on Spoken Language Processing*, Sydney, Australia, pp. 2803–2806.
- Hilbert, D. (1912). *Grundzuege einer Allgemeinen Theorie der Linearen Integralgleichungen (Basic Principles of a General Theory of Linear Integral Equations)* (Teubner, Leipzig), pp. 1–282.
- Hopkins, K., and Moore, B. C. J. (2009). "The contribution of temporal fine structure to the intelligibility of speech in steady and modulated noise," *J. Acoust. Soc. Am.* **125**, 442–446.
- Hopkins, K., and Moore, B. C. J. (2010). "The importance of temporal fine structure information in speech at different spectral regions for normal-hearing and hearing-impaired subjects," *J. Acoust. Soc. Am.* **127**, 1595–1608.
- IEEE. (1969). "IEEE recommended practice for speech quality measurements," *IEEE Trans. Audio Electroacoust.* **17**, 227–246.
- Kasturi, K., Loizou, P. C., Dorman, M., and Spahr, T. (2002). "The intelligibility of speech with 'holes' in the spectrum," *J. Acoust. Soc. Am.* **112**, 1102–1111.
- Kawahara, H., Masuda-Katuse, I., and Cheveigne, A. (1999). "Restructuring speech representations using a pitch adaptive time-frequency smoothing and an instantaneous frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Commun.* **27**, 187–207.
- Kewley-Port, D., Burkle, T. Z., and Lee, J. H. (2007). "Contribution of consonant versus vowel information to sentence intelligibility for young normal-hearing and elderly hearing-impaired listeners," *J. Acoust. Soc. Am.* **122**, 2365–2375.
- Kong, Y.-Y., and Carlyon, R. P. (2007). "Improved speech recognition in noise in simulated binaurally combined acoustic and electric stimulation," *J. Acoust. Soc. Am.* **121**, 3717–3727.
- Li, N., and Loizou, P. C. (2008). "A glimpsing account for the benefit of simulated combined acoustic and electric hearing," *J. Acoust. Soc. Am.* **123**, 2287–2294.
- Lieberman, M. C. (1982). "The cochlear frequency map for the cat: Labeling auditory-nerve fibers of known characteristic frequency," *J. Acoust. Soc. Am.* **72**, 1441–1449.
- Loizou, P. C. (2007). *Speech Enhancement: Theory and Practice* (CRC Press, Taylor and Francis, Boca Raton, FL), pp. 1–608.
- Lorenzi, C., Gilbert, G., Cam, H., Garnier, S., and Moore, B. C. J. (2006). "Speech perception problems of the hearing impaired reflect inability to use temporal fine structure," *Proc. Natl. Acad. Sci. U.S.A.* **103**, 18866–18869.
- Lutfi, R. A. (1995). "Correlation coefficients and correlation ratios as estimates of observer weights in multiple-observation tasks," *J. Acoust. Soc. Am.* **97**, 1333–1334.
- Mehr, M. A., Turner, C. W., and Parkinson, A. (2001). "Channel weights for speech recognition in cochlear implant users," *J. Acoust. Soc. Am.* **109**, 359–366.
- Miller, G., and Licklider, J. (1950). "The intelligibility of interrupted speech," *J. Acoust. Soc. Am.* **22**, 167–173.
- Moore, B. C. J. (2004). *An Introduction to the Psychology of Hearing*, 5th ed. (Elsevier, London) pp. 1–413.
- Moore, B. C. J. (2008a). "The role of temporal fine structure processing in pitch perception, masking, and speech perception for normal-hearing and hearing-impaired people," *J. Assoc. Res. Otolaryngol.* **9**, 399–406.
- Moore, B. C. J. (2008b). "The choice of compression speech in hearing aids: Theoretical and practical considerations and the role of individual differences," *Trends Amp.* **12**, 103–112.
- Moore, B. C. J., and Glasberg, B. R. (1983). "Suggested formulae for calculating auditory-filter bandwidths and excitation pattern," *J. Acoust. Soc. Am.* **74**, 750–753.
- Moore, B. C. J., Glasberg, B. R., Flanagan, H. J., and Adams, J. (2006). "Frequency discrimination of complex tones: assessing the role of component resolvability and temporal fine structure," *J. Acoust. Soc. Am.* **119**, 480–490.
- Moore, G. A., and Moore, B. C. J. (2003). "Perception of the low pitch of frequency-shifted complexes," *J. Acoust. Soc. Am.* **113**, 977–985.
- Nelson, P. B., and Jin, S.-H. (2004). "Factors affecting speech understanding in gated interference: Cochlear implant users and normal-hearing listeners," *J. Acoust. Soc. Am.* **115**, 2286–2294.
- Nelson, P. B., Jin, S.-H., Carney, A. E., and Nelson, D. A. (2003). "Understanding speech in modulated interference: Cochlear implant users and normal-hearing listeners," *J. Acoust. Soc. Am.* **113**, 961–968.
- Peters, R. W., Moore, B. C. J., and Baer, T. (1998). "Speech reception thresholds in noise with and without spectral and temporal dips for hearing-impaired and normally hearing people," *J. Acoust. Soc. Am.* **103**, 577–587.
- Plack, C. J., and Oxenham, A. J. (2005). "The psychophysics of pitch," in *Pitch: Neural Coding and Perception*, edited by C. J. Plack, A. J. Oxenham, A. N. Popper, and R. Fay (Springer, New York), pp. 7–55.
- Qin, M. K., and Oxenham, A. J. (2003). "Effects of simulated cochlear-implant processing on speech reception in fluctuating maskers," *J. Acoust. Soc. Am.* **114**, 446–454.
- Richards, V. M., and Zhu, S. (1994). "Relative estimates of combination weights, decision criteria, and internal noise based on correlation coefficients," *J. Acoust. Soc. Am.* **95**, 423–434.
- Rosen, S. (1992). "Temporal information in speech: Acoustic, auditory, and linguistic aspects," *Philos. Trans. R. Soc. B* **336**, 367–373.
- Shannon, R. V., Zeng, F.-G., and Wygonski, J. (1998). "Speech recognition with altered spectral distribution of envelope cues," *J. Acoust. Soc. Am.* **104**, 2467–2476.
- Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., and Ekelid, M. (1995). "Speech recognition with primarily temporal cues," *Science* **270**, 303–304.
- Silipo, R., Greenberg, S., and Arai, T. (1999). "Temporal Constraints on Speech Intelligibility as Deduced from Exceedingly Sparse Spectral Representations," *Proc. Eurospeech 1999*, Budapest, Hungary, pp. 2687–2690.
- Sivonen, P., Maess, B., Lattner, S., and Friederici A. D. (2006). "Phonemic restoration in a sentence context: Evidence from early and late ERP effects," *Brain Res.* **1121**, 177–189.
- Smith, Z. M., Delgutte, A. J., and Oxenham, A. J. (2002). "Chimaeric sounds reveal dichotomies in auditory perception," *Nature* **416**, 87–90.
- Summers, V., and Molis, M. R. (2004). "Speech recognition in fluctuating and continuous maskers: Effects of hearing loss and presentation level," *J. Speech Lang. Hear. Res.* **47**, 245–256.
- Wang, D. (2008). "Time-frequency masking for speech separation and its potential for hearing aid design," *Trends Amp.* **12**, 332–353.
- Yost, W. A., and Sheft, S. (1989). "Across-critical-band processing of amplitude-modulated tones," *J. Acoust. Soc. Am.* **85**, 848–857.