Perceptual weighting of individual and concurrent cues for sentence intelligibility: Frequency, envelope, and fine structure

Daniel Fogerty^{a)}

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

(Received 17 August 2010; revised 19 November 2010; accepted 6 December 2010)

The speech signal may be divided into frequency bands, each containing temporal properties of the envelope and fine structure. For maximal speech understanding, listeners must allocate their perceptual resources to the most informative acoustic properties. Understanding this perceptual weighting is essential for the design of assistive listening devices that need to preserve these important speech cues. This study measured the perceptual weighting of young normal-hearing listeners for the envelope and fine structure in each of three frequency bands for sentence materials. Perceptual weights were obtained under two listening contexts: (1) when each acoustic property was presented individually and (2) when multiple acoustic properties were available concurrently. The processing method was designed to vary the availability of each acoustic property 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 that weights were (1) equal when acoustic properties were presented individually and (2) biased toward envelope and mid-frequency information when multiple properties were available. Results suggest a complex interaction between the available acoustic properties and the listening context in determining how best to allocate perceptual resources when listening to speech in noise. © 2011 Acoustical Society of America. [DOI: 10.1121/1.3531954]

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

Pages: 977–988

I. INTRODUCTION

Examining how speech information is distributed over the spectrum has been of paramount importance for modeling the intelligibility of speech under degraded conditions, such as from reverberation, competing talkers, filtering, or hearing loss (e.g., French and Steinberg, 1947; Fletcher and Galt, 1950; Houtgast and Steeneken, 1985; ANSI, 1997). The measurement of frequency-importance functions has application to understanding the functional impact of hearing loss, as well as designing assistive devices and speech transmission technologies. The primary method of obtaining these importance functions has been through high- and low-pass filtering studies in which the intelligibility of restricted frequency regions has been tested (e.g., French and Steinberg, 1947). Research investigating frequency-band importance functions have found that certain frequency bands contribute more to the intelligibility of uninterrupted speech and this information has been incorporated into the speech intelligibility index (SII; ANSI, 1997). Each frequency band contains acoustic information which can further be divided into two temporal properties: envelope (E) and temporal fine structure (TFS). The spectral distribution of the perceptual use of E and TFS properties has also been examined using high- and low-pass filtering methods (Ardoint and Lorenzi, 2010). Each of these studies has examined the independent contribution of isolated spectral regions where the spectral regions have been defined typically using octave or 1/3-octave bands.

However, in contrast to the filtered speech contexts that form the basis of these frequency-importance functions, in

everyday listening environments, listeners have the entire broadband speech spectrum available. Therefore, these methods index how specific adjacent frequency regions contribute independently to speech intelligibility, but they do not account for how perceptual use of these frequency regions might change when multiple spectral regions are available, particularly if these spectral regions are widely separate (for how listeners may synergistically combine non-adjacent bands see Grant and Braida, 1991; Warren *et al.*, 1995). How the perceptual use of E and TFS temporal properties combines across frequency regions also has yet to be fully explored.

The current study explores the perceptual use of multiple frequency and temporal acoustic channels for sentence intelligibility when these channels are presented separately and concurrently. Experiment 1 explores the perceptual use of three different frequency bands in broadband speech contexts. Furthermore, it extends investigations to the E and TFS. Experiment 2 measures the perceptual weight of the E and TFS information across three frequency bands. In particular, Experiment 2 measures these weights when multiple cues are available to listeners using the correlational method (Berg, 1989; Richards and Zhu, 1994; Lutfi, 1995).

The correlational method was developed and applied to speech (see 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. The speech acoustics in each band were individually and randomly degraded on each trial through the addition of noise. For example, the signal-to-noise ratio (SNR) within each of three bands on a given trial could be -7 , $+1$, and -1 dB. The trial-by-trial data were then analyzed by calculating the point-biserial correlation (Lutfi,

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

1995) of the SNR within each band and the listener's performance (using binary coding: correct/incorrect) on that given trial. These correlations were then normalized to 1.0 across the three bands to obtain relative weights that were interpreted as indexing the proportion to which that band's weight contributed on each trial to the listener's speech recognition performance. This correlational method has the advantage of SII-type filtering experiments in that it can determine the relative contributions of independent frequency bands, but, unlike the SII approach, does so by using the full bandwidth of speech. Another advantage of the correlational method over the filtering experiments is that it is able to obtain weights when the speech information presented is separated into substantially different frequency regions.

The correlational method has been applied to a host of domains, including spectral shape discrimination (Berg and Green, 1990, 1992; Lentz and Leek, 2002) and temporal envelope discrimination (Apoux and Bacon, 2004), as well as, cochlear-implant weighting strategies (Mehr *et al.*, 2001) and electrode function (Turner *et al.*, 2002). In addition, Calandruccio and Doherty (2007) have extended the approach to obtain weighting functions for IEEE sentence materials. This project is designed specifically to address how temporal information within each of several frequency bands contributes to speech understanding.

The contribution of various frequency regions is likely dependent upon the type of linguistic information conveyed within that region. This may help explain why different frequency-weighting functions are obtained for different speech materials (ANSI, 1997). For example, vowels appear to contribute more than consonants in sentence contexts (Kewley-Port *et al.*, 2007; Fogerty and Kewley-Port, 2009), but not in word contexts (Owren and Cardillo, 2006; Fogerty and Humes, 2010). Different linguistic information becomes available and is more informative in different speech contexts, such as the prosodic contour constraining the lexical access of words in sentences (Wingfield *et al.*, 1989; Laures and Weismer, 1999). By bandpass filtering speech into three frequency regions broadly associated with prosodic, sonorant, and obstruent/fricative linguistic categories, it is possible to map linguistic contributions onto more general acoustic properties. In addition to frequency contributions, speech within each frequency band is comprised of complex temporal modulations in amplitude (i.e., E) and frequency (i.e., TFS). Amplitude modulations alone, such as implemented by cochlear implants, convey voicing, and manner cues, while frequency cues convey information regarding the place of articulation (Xu *et al.*, 2005).

Experiment 1a investigates the contribution of different frequency regions of linguistic significance. Grant and Walden (1996) have found that high-frequency speech acoustics convey the linguistic information of syllabic number and stress while low-frequency speech acoustics convey intonation patterns. Thus, the frequency spectrum of speech may be divided into regions of predominant linguistic significance. In contrast to previous filtering studies, the contribution of these frequency regions is examined in a wideband listening context.

Experiment 1b extends these methods to examining the independent contributions of E and TFS temporal dynamics that have been analyzed across these three frequency regions. The

different temporal timescales present in speech, slow E modulations and fast TFS modulations, also convey different types of linguistic information (Rosen, 1992). For example, the E provides information about the syllabic structure and speech rate while the TFS provides acoustic cues related to dynamic formant transitions. Experiment 1b examines the independent contributions of these two types of temporal modulation properties. Previous studies have manipulated the amount of temporal information available in several ways, such as by varying the number of analysis bands (e.g., Shannon *et al.*, 1995; Dorman *et al.*, 1997; Hopkins *et al.*, 2008), filtering the modulation rate (e.g., Drullman *et al.*, 1994a,b; van der Horst *et al.*, 1999), or both (e.g., Xu *et al.*, 2005). By doing so, these studies have altered the temporal information present in the speech stimulus and in some cases also the spectral information. In contrast to these methods, experiment 1b varies the SNR from trial-to-trial and uses the correlational method to derive importance estimates. In doing so, the processing of temporal information remains unchanged, but the availability of these cues is altered systematically, thereby, allowing the measurement of the perceptual contributions of specific temporal modulation properties to speech intelligibility.

Finally, experiment 2 investigates the distributions of perceptual weights for the E and TFS across frequency bands. In contrast to experiment 1 that investigates individual acoustic channels, experiment 2 measures perceptual weights when speech information in multiple acoustic channels is concurrently available. For example, while listeners may be able to use all spectral information equally, they may place more weight on a particular spectral region when information across the frequency spectrum is available.

This study varies the availability of each of these acoustic cues (frequency bands or temporal properties) across a range of SNRs. In doing so, this set of experiments:

- (1) measures the contributions of individual acoustic properties in a masked wideband context;
- (2) examines the contributions of linguistically relevant acoustic divisions;
- (3) demonstrates the feasibility of extending the SNR method to temporal properties; and
- (4) measures perceptual weights for the E and TFS across frequency bands when multiple cues are available.

II. EXPERIMENT 1A: INDIVIDUAL FREQUENCY BAND PERCEPTUAL WEIGHTS

This experiment investigated the individual contributions of each of three frequency bands for wideband speech presentation of meaningful sentences. Performance in each test band was determined as a function of SNR while the contribution of the other two bands was limited by noise masking. The frequency regions used here are related to primary (linguistically relevant) speech information.

A. Listeners

Nine young normal-hearing listeners (18–21 yr) were paid to participate in the study. All participants were native speakers of American English and had pure-tone thresholds

bilaterally not greater than 20 dB HL at octave intervals from 250 to 8000 Hz (ANSI, 2004).

B. Stimuli and design

IEEE/Harvard sentences 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). These stimuli are all meaningful sentences that contain five keywords, such as in the experimental sentence, “The *birch canoe slid* on the *smooth planks*.” All signal processing of the stimuli were completed in MATLAB using custom software and modifications of code provided by Smith *et al.* (2002). All experimental sentences were down-sampled to a sampling rate of 16 000 Hz and passed through a bank of band-pass filters to process speech into three different frequency bands. These frequency bands represent equal distance along the cochlea using a cochlear map (Liberman, 1982) and correspond roughly to prosodic, sonorant, and obstruent/fricative linguistic categories, respectively. Table I displays the frequency range of these bands along with the corresponding number of equivalent rectangular bands (ERBs; Moore and Glasberg, 1983) and the calculated SII values assuming full contribution (i.e., band SNR > 15 dB) for each band. Note that the selected bands for this study have about the same number of ERBs and similar SII values. Fundamental frequency and formant frequency values were calculated for all experimental sentences using STRAIGHT, a speech analysis, modification, and synthesis system (Kawahara *et al.*, 1999). Mean values for each sentence were calculated using a 50-ms sliding window and are provided in Table II.

The contribution of each independent frequency band was investigated by varying the SNR of the target band of interest and maintaining a constant negative SNR in the other two bands. Thus, stimuli contained speech acoustic cues across the entire stimulus spectrum, with cues heavily masked in the two non-target (i.e., masked) bands. The presence of the noise masked bands was included to prevent off-frequency listening and more closely match listening conditions when broadband speech information is available.

Figure 1 displays the processing for the independent band conditions. After passing the speech stimulus through the bank of three bandpass filters, a constant noise was added to the two masked bands to yield a SNR of -5 dB in each of those bands. A unique sample of noise was generated for each sentence to match the same power spectrum as the target sentence. This noise was then bandpass-filtered and summed to the corresponding speech band. For the remaining (target) speech band, the SNR was varied individually over a range of five SNRs (11, 5, 2, -1 , -7 dB). Thus, performance on this task varied as a function of SNR in the target frequency

TABLE I. Frequency range, ERB, and SII values for the three frequency bands used in this study.

Band	Frequency range (Hz)	ERB	SII (ANSI, 1997)
1	80–528	8.3	0.32
2	528–1941	9.9	0.35
3	1941–6400	10.0	0.33

TABLE II. Means and standard deviations (StDev) for the fundamental (F0) and formant (F1, F2, F3) frequency values averaged across 600 IEEE sentences.

Formant	Mean (Hz)	StDev (Hz)
F0	131	7
F1	492	31
F2	1561	71
F3	2910	68

band. All three frequency bands were investigated, creating a total of 15 conditions (3 frequency bands $\times$ 5 SNR levels). Twenty-four sentences (120 keywords) were presented per condition with no sentences being repeated for a given listener. All 360 sentences were presented in a fully randomized order. After processing, all stimuli were up-sampled to 44 828 Hz for presentation through Tucker–Davis Technologies (TDT) System-III hardware.

C. Calibration

A speech-shaped noise was designed in Adobe Audition to match the long-term average speech spectrum (± 2 dB) of a concatenation of 600 unprocessed sentences from the IEEE database. Note that this calibration noise, which was only used during calibration for all experiments described here, was different from the unique noise samples that were used during stimulus processing, described above, which matched the spectrum of each sentence. Stimuli were presented via TDT System-III hardware using 16-bit resolution at a sampling frequency of 48 828 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. This calibration noise was set to 70 dB SPL through the insert earphone in a 2-cc 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 SPL. However, after filtering and noise masking, the overall

FIG. 1. Stimulus processing diagram for experiment 1a. The noise was generated to match the long-term average speech spectrum for the target sentence. The noise level in the target band is indexed by i , corresponding to the five SNR conditions.

sound level of the combined stimulus varied according to the individual condition. This ensured that levels representative of typical conversational level (70 dB SPL) 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 custom designed MATLAB stimulus presentation interface. Each listener was instructed regarding the task using verbal and written instructions. All listeners completed familiarization trials prior to each experimental test that exhibited the same signal processing as the experimental sentences. These sentences were selected from male talkers in the TIMIT database (Garofolo, *et al.*, 1990) and 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 to avoid explicit learning across the many stimulus trials. All listeners received different randomizations of the 360 experimental sentences (120 sentences/target band). Each sentence was presented only once.

During the experimental testing, each sentence was presented individually and the listener was prompted to repeat the sentence aloud as accurately as possible. Listeners were encouraged to guess, without regard to whether their responses made logical sentences. No feedback was provided. All listener responses were digitally recorded for later analysis. Only keywords repeated exactly were scored as correct (e.g., no missing or extra suffixes). In addition, keywords were allowed to be repeated back in any order to be counted as correct repetitions. Each keyword was marked as 0 or 1, corresponding to an incorrect or correct response, respectively. 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 and discussion

Frequency band (three bands) and SNR (five levels) were entered as repeated-measures variables in a general linear model analysis of the percent-correct scores (based on keywords). All percent-correct scores, here and elsewhere, were transformed into rationalized arcsine units to stabilize the error variance (Studebaker, 1985) prior to analysis. A main effect for SNR [$F(4,32) = 173.4$, $p < 0.001$] and target band $\times$ SNR interaction [$F(8,64) = 6.2$, $p = 0.001$] were obtained after Greenhouse–Geisser correction of the degrees of freedom. No main effect of frequency band was obtained. *Post hoc* paired-samples *t* tests demonstrated significant differences between the low-frequency band and both the mid- and high-frequency bands for SNRs at -1 and 5 dB (Bonferroni-corrected $p < 0.003$). Figure 2 plots the performance in each band across SNR. Performance for each of the three bands was significantly correlated across all SNRs (Pearson's $r = 0.69 - 0.90$, $p < 0.05$), indicating that indi-

FIG. 2. Performance function for each frequency band varied independently. Error bars indicate standard error of the mean.

vidual differences in listener performance were consistent across all three bands and SNRs.

The trial-by-trial data were analyzed by calculating the point-biserial correlation (Lutfi, 1995) between the SNR within the target band and the listener's word score (using binary coding: correct/incorrect) on that given trial. These correlations are provided in Table III. Correlations contained 600 points for each frequency band (5 SNRs $\times$ 24 sentences $\times$ 5 keywords). All raw correlations were significantly above the null hypothesis of zero (for a 95% confidence interval, $|r| > 1.96/600^{1/2}$, see Lutfi, 1995), indicating perceptual use of each frequency band. The correlations across the three frequency bands were then normalized to sum to one to obtain relative weights for each band (see Doherty and Turner, 1996; Turner *et al.*, 1998). Figure 3 displays these relative correlational weights. As Turner *et al.* (1998) noted with nonsense syllables, the correlational weights obtained for each target frequency band, each measured in isolation, are approximately equal. No difference between frequency band weights were obtained (Bonferroni-corrected $p > 0.016$), indicating that the perception of each independent band was affected by the noise in a similar fashion. This is consistent with the independent, additive nature of bands in Articulation Theory (French and Steinberg, 1947) and the equivalent SII values estimated for the bands

TABLE III. Point-biserial correlations in each independent frequency band for all listeners. Average performance in percent-correct across the three bands is also listed (second column).

Listener	Average (%)	Band 1	Band 2	Band 3
1a_01	35	0.40	0.39	0.42
1a_02	50	0.39	0.29	0.33
1a_03	37	0.39	0.44	0.40
1a_04	47	0.42	0.31	0.36
1a_05	41	0.38	0.37	0.32
1a_06	41	0.44	0.43	0.33
1a_07	29	0.33	0.23	0.25
1a_08	25	0.24	0.24	0.25
1a_09	40	0.40	0.35	0.35
Mean	38	0.38	0.34	0.33

FIG. 3. Normalized correlations or relative weights for each target band. Dots display individual data for the nine listeners. Fewer than nine dots indicate identical weights for listeners. The solid line represents the mean weight.

used in this study (reported in Table I). Performance for all listeners followed the equal-weighting pattern apparent in the mean data.

Finally, it is important to note that the SII predicts the contribution of bands down to -15 dB SNR. Therefore, the conditions presented here reflect contributions of dominant speech information in the target band in combination with other non-target information. However, evaluation of masked non-target bands alone did not exceed 8% correct word identification during piloting. Furthermore, performance variations occurred across the target SNR, with the availability of masked non-target bands held constant across conditions. Overall, results suggest equal contributions of these three bands when independently tested in the presence of other wideband speech information.

III. EXPERIMENT 1B: INDIVIDUAL ENVELOPE AND FINE STRUCTURE PERCEPTUAL WEIGHTS

Experiment 1b was designed to test the temporal modulation properties of the E and TFS, as defined by the Hilbert transform, using methods analogous to experiment 1a. Previous research has typically measured the perceptual strength of E or TFS information by either varying the number of analysis frequency bands (e.g., Dorman *et al.*, 1997; Shannon *et al.*, 1998; Smith *et al.*, 2002) or systematically adding temporal information to successively more bands (Hopkins *et al.*, 2008). Filtering modulation frequencies of the envelope (Drullman *et al.*, 1994a,b; Shannon *et al.*, 1995; van der Horst *et al.*, 1999; Xu and Pfingst, 2003) or quantifying the strength of speech envelope modulations relative to spurious non-signal modulations (Dubbelboer and Houtgast, 2008) have also been investigated as methods to evaluate the importance of temporal information. However, varying the SNR has proven to be a robust method for examining the contribution of specific speech information (e.g., Miller *et al.*, 1951). Using this same method for the evaluation of temporal information provides a significant advantage for comparing perceptual weights between frequency and temporal dimensions as it is able to place both acoustic properties on the same measurement scale using the same type of speech distortion. This approach has been used increasingly

to investigate the perceptual use of E information (e.g., Apoux and Bacon, 2004), and is extended here to also investigate perceptual weighting of TFS information. This method has the advantage of presenting the entire speech stimulus, while selectively masking E or TFS cues. Therefore, experiment 1b obtained performance functions independently for E and TFS acoustics by varying the SNR of the target speech information. Note that the E was defined by the Hilbert transform and therefore contained even the highest modulation rates, which would include some periodicity cues. These fast-rate E-modulation cues were included so as to provide all speech information, as fast-rate cues may significantly contribute to speech intelligibility. Recent studies have similarly presented the entire Hilbert envelope as the E information (e.g., Gallun and Souza, 2008; Hopkins and Moore, 2010).

A. Methods

Eight new listeners participated in experiment 1b (19–23 yr). All participants were native speakers of American English and had pure-tone thresholds bilaterally no greater than 20 dB HL at octave intervals from 250 to 8000 Hz (ANSI, 2004).

Experiment 1b used the same presentation procedures as in experiment 1a. Sentences were again selected from the IEEE database for testing. The 240 sentences used here were not presented in experiment 1a. The E contains amplitude fluctuations across a range of modulation rates. Modulation rates for E peak around 4 Hz for broadband processing of these IEEE speech materials. Importantly, E and TFS contributions have been shown to be similar when processed using three frequency bands (Smith *et al.*, 2002) and vowels and consonants are identified equally well for three-band envelope vocoders processed over a range of 150–5500 Hz (Xu *et al.*, 2005). The metric for frequency band analysis (linear vs logarithmic) also does not influence speech perception using only envelope cues (Shannon *et al.*, 1998). Therefore, the three frequency bands used in experiment 1a were used as the analysis bands for the current E and TFS processing.

A diagram for the stimulus processing is displayed in Fig. 4. A speech-shaped noise matching the long-term average of the target sentence was created. This noise was scaled to -5 dB SNR and added to the target sentence. Then, the masked sentence was passed through a bank of three analysis

FIG. 4. Stimulus processing diagram for envelope (E) and fine structure (TFS) stimuli (experiment 1b). The same SNR was applied across all three frequency bands and is indexed by i , corresponding to the five SNR values available.

filters. Within each frequency band, the E and TFS were extracted using the Hilbert transform, resulting in isolated E and TFS properties in each of the three bands. These masked modulation properties will be referred to as -5_E and -5_TFS , respectively. The speech-shaped noise was also scaled over the range of SNR values used in experiment 1a (11, 5, 2, -1 , -7 dB) and added to the original sentence, from which the target E (E_target) and target TFS (TFS_target) were extracted over the three analysis bands. For envelope testing, in each analysis band the E_target and masked -5_TFS were combined and summed across frequency bands. The same was performed for TFS testing, where the E in each band was replaced by the masked -5_E for that band and combined with the TFS_target . Thus, for both E and TFS testing, the non-test modulation property was masked at a constant -5 dB SNR while the target portion was varied over the range of test SNRs. Finally, the entire stimulus was re-filtered at 6400 Hz and upsampled to 48 828 Hz to produce the final stimulus. There were a total of ten conditions (2 temporal properties $\times$ 5 SNRs). Twenty-four sentences (120 keywords) were presented per condition.

As in experiment 1a, listeners were seated alone in a sound-attenuating booth. Each participant listened to a unique randomization of the sentences in each of the ten conditions monaurally via an ER-3A insert earphone and responded aloud by repeating as accurately as possible the sentence that was presented. Responses were digitally recorded for offline analysis and scoring. All listeners first received familiarization trials, as in experiment 1a. No feedback was provided.

B. Results and discussion

A two (temporal property: E and TFS) by five (SNR level) repeated-measures analysis of variance (ANOVA) resulted in significant main effects for the temporal property [$F(1,7) = 123.7$, $p < 0.001$] and SNR [$F(4,28) = 346.0$, $p < 0.001$], with a significant interaction between these two variables as well [$F(4,28) = 18.1$, $p < 0.001$]. *Post hoc t* tests between E and TFS conditions at each SNR demonstrated significant differences at -1 , 2, and 5 dB SNR (Bonferroni-corrected $p < 0.01$). As seen in Fig. 5, in general, the percep-

FIG. 5. Performance function for each temporal property varied independently. Error bars indicate standard error of the mean. E = envelope, TFS = temporal fine structure.

TABLE IV. Point-biserial correlations in each independent modulation band across all SNRs for all listeners. Average percent-correct performance across the two types of modulation is also listed (second column).

Listener	Average (%)	E	TFS
1b_01	57	0.50	0.44
1b_02	63	0.41	0.43
1b_03	61	0.41	0.43
1b_04	57	0.43	0.44
1b_05	51	0.46	0.49
1b_06	68	0.39	0.34
1b_07	51	0.45	0.49
1b_08	47	0.50	0.52
Mean	57	0.44	0.45

tual use of E and TFS acoustic information varies similarly across SNRs with listeners tending toward better performance, by about 9 percentage points on average, for the E conditions. As noted for the frequency band stimuli of experiment 1, correlations between E and TFS conditions were significant across all SNRs (Pearson's $r = 0.84-0.87$, $p < 0.01$) with the exception of at 11 dB SNR. Also of note is that performance appears to level off at 81%. This is in agreement with Shannon *et al.* (1995) who found that E-only speech reached a maximum of about 80% for sentences in quiet when processed through three frequency bands.

As in experiment 1a, point-biserial correlations were calculated between the word score and the SNR in the target temporal property for that trial and these appear in Table IV. Correlations contained 600 points for both temporal modulation property conditions (5 SNRs $\times$ 24 sentences $\times$ 5 keywords). Again both E and TFS raw correlations were significantly above zero ($p < 0.05$). The correlations across the two temporal conditions were then normalized to sum to one to obtain relative weights. These weights are plotted in Fig. 6. No significant difference was obtained between E and TFS perceptual weights [$t(7) = -2.84$, $p = 0.79$]. All listeners followed this pattern, indicating that they placed equal relative perceptual weight on these two types of temporal acoustic information.

IV. DISCUSSION OF EXPERIMENT 1

Results of experiment 1a demonstrated that performance varies as a function of the SNR similarly within each of the

FIG. 6. Normalized correlations for the envelope (E) and fine structure (TFS) target stimuli. Dots display individual data for the eight listeners who completed both E and TFS testing. Individual data points overlap. The solid line represents the mean weight.

three frequency bands. Furthermore, all listeners placed equal weight on each of these individual spectral regions. These frequency bands convey very different types of speech information, yet these results indicate that listeners are able to use these different acoustic cues to obtain similar speech-recognition performance for these speech materials.

Each of these bands was presented in a wideband speech context that best models the real world acoustic environment for normal-hearing listeners. As the SNR in this target band was the only independent variable, changes in performance may be attributed to the use of speech information in that target band. However, it is important to note that each target band was presented along with other wideband speech information, albeit noise masked. Therefore, it cannot be determined whether each target band provides direct information for speech recognition, or indirectly facilitates sentence intelligibility via the disambiguation of other off-band speech information.

Results of experiment 1b demonstrated that performance varies as a function of the SNR similarly for E and TFS information. The results obtained here for E and TFS masking are in good agreement with other methods that have indicated approximately equal contributions of E and TFS when processed in three frequency bands. Using a chimaeric paradigm where the E of one sentence and the TFS of another sentence were combined, [Smith et al. \(2002\)](#) found that listeners had the same accuracy for both the E and TFS sentences when processed through three frequency bands. The equal perceptual weights obtained in the current study support this finding. [Ardoit and Lorenzi \(2010\)](#) also examined the spectral distribution of E and TFS cues using high- and low-pass filtering. They found a cross-over frequency of these filters at 1500 Hz for both E and TFS information, which is in the middle of the mid-frequency band in this study. This suggests that E and TFS information were equally distributed across each of the three frequency bands used here and is consistent with the equal perceptual weights obtained.

This study introduced a method of varying the availability of temporal properties without altering the underlying analysis and extraction of E and TFS, thereby, preserving the modulation and frequency spectra of the underlying speech stimulus. This method allows equivalent signal processing to be applied to temporal modulation information as applied to the frequency band information. Results demonstrated that listeners are equally able to use acoustic information in each of these acoustic channels when presented in wideband contexts. However, it may be that E or TFS cues are more informative in certain spectral regions than others. The processing method used here allows for the examination of perceptual weights placed on E and TFS cues in different spectral regions, which experiment 2 explores.

V. EXPERIMENT 2: PERCEPTUAL WEIGHTING OF ENVELOPE AND FINE STRUCTURE ACROSS FREQUENCY FOR SPEECH IN NOISE

Experiment 1 demonstrated the feasibility of using noise masking and the correlational method to measure the perceptual weights assigned to E and TFS speech information when each temporal cue was independently available. Experiment

2 was designed to investigate the perceptual weight that is assigned to these two temporal modulation properties in different frequency regions when multiple acoustic cues are concurrently available.

Temporal modulation properties may be weighted differently in each of the frequency bands. It may be that the TFS is perceptually more available in the low-frequency band, as neural fine structure coding decreases at higher frequencies (reviewed by [Moore, 2008](#)). However, TFS added to individual bands across the spectrum facilitates speech understanding above what is provided by E cues alone ([Hopkins and Moore, 2010](#)). Envelope acoustics might be weighted more heavily in high-frequency regions ([Apoux and Bacon, 2004](#)) where the onsets and offsets of consonants are most present, as abrupt E changes are coded by slow synchronous neural responses ([Moore et al., 2001](#)). Experiment 2 investigated the perceptual weight of E and TFS properties across frequency bands when multiple sources of speech information were available. A correlational method ([Richards and Zhu, 1994](#); [Lutfi, 1995](#)) was again used to relate the availability of speech information in a specific channel, indexed by the SNR, with performance on that given trial. Thus, perceptual weights were obtained using a trial-by-trial analysis of performance and signal information.

A number of researchers have used this correlational method to examine the perceptual weights that individuals place on speech information in different frequency regions for nonsense speech ([Doherty and Turner, 1996](#); [Turner et al., 1998](#)), sentences ([Calandruccio and Doherty, 2007](#)), and with the hearing impaired ([Calandruccio and Doherty, 2008](#)). [Apoux and Bacon \(2004\)](#) have also examined E contributions in isolated band-limited stimuli. However, the current study extends these investigations by independently and concurrently varying E and TFS information in three frequency bands. Thus, the distribution of perceptual weights applied to temporal information across the spectrum in wideband speech was obtained in this experiment. That is, in experiments 1a and 1b there was only one “channel” of speech information predominantly available at a time—either one of the three spectral channels (exp 1a) or one of two possible temporal channels (exp 1b). Here, six channels of information (3 spectral bands $\times$ 2 temporal modulation properties) are available.

A. Listeners

Ten young normal-hearing listeners (18–26 yr) were paid to participate in the study. All participants were native speakers of American English and had pure-tone thresholds bilaterally at no greater than 20 dB hearing level (HL) at octave intervals from 250 to 8000 Hz ([ANSI, 2004](#)). These listeners did not participate in experiment 1.

B. Stimuli and design

The experimental task presented multiple channels using 600 IEEE sentences originally recorded by [Loizou \(2007\)](#). The task consisted of varying the SNR along two different dimensions. Figure 7 displays a processing diagram used for these stimuli. The first dimension along which SNR

FIG. 7. Stimulus processing diagram for experiment 2. An IEEE sentence and a speech-shaped noise matching the target sentence were filtered into three bands. The noise was added at different SNRs to copies of each band prior to the Hilbert transform, which extracted either E or TFS. All channels were then recombined to form the final noisy stimulus. The SNR independently and randomly assigned to each channel is indexed by i , corresponding to the five SNR values available.

varied was spectral frequency or frequency band. The same three frequency bands described in experiment 1 were used here. The second dimension corresponded to the temporal properties that are present within each frequency band: E and TFS. As noted, this resulted in six acoustic information “channels” (two temporal properties in each of three frequency bands) and will be noted according to the temporal property and band number, as in E1 (i.e., E modulation in band 1). 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 the 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 for E and TFS was -1 and -7 dB for the E and TFS in band 1, respectively, two copies of the speech in band 1 were made. These were combined with filtered noise scaled to either -1 or -7 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 -7 dB band copy. The noisy Hilbert components, masked at different SNRs, were then recombined, re-filtered at 6400 Hz and up-sampled to 48 828 Hz for presentation through the TDT System-III hardware.

The SNR was varied independently in each of these six different acoustic channels. 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 the average long term speech spectrum of that trial sentence. SNR values ranged from -7 to 5 dB in 3 dB steps, resulting in five different levels. A correlational method (Richards and Zhu, 1994; Lutfi, 1995) 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 pre-

sented at $-1, 2, -1, -7, 5$, and 2 dB distributed across the six channels). Each SNR was presented in each acoustic channel a total of 120 trials (120×5 SNRs = 600 total trials). Each sentence contained five keywords, resulting in 3000 keywords scored.

C. Procedure

Test procedures were the same as in experiment 1, including familiarization trials processed according to the stimulus design used here. All listeners received different randomizations of the experimental sentences. All sentences were only presented once to each listener. As in experiment 1, stimuli were presented unilaterally to the right ear. The listeners’ task was to repeat aloud as accurately as possible the sentence that was presented. Responses were digitally recorded for offline analysis. The same raters as in experiment 1 scored the responses for experiment 2.

D. Results

1. Mean perceptual weights

Listeners were tested on the correlational task during two different sessions. Average keyword-recognition performance on each day was calculated to determine if learning occurred between the 2 days even though feedback was not provided. No significant difference between the performances on each day was observed ($p > 0.05$). Overall, listeners only improved performance by 0.1 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. 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 was calculated trial-by-trial between the word score and the SNR in each channel and these are presented in Table V. Each correlation was

TABLE V. Point-biserial correlation coefficients for each listener across each of the six channels. The listener’s performance was correlated on a trial-by-trial basis with the signal-to-noise ratio in each channel for that trial. All correlations were significantly different from zero ($p < 0.05$). Overall percent-correct scores across all trials are also listed for each listener (second column).

Listener	Average (%)	E1	E2	E3	TFS1	TFS2	TFS3
2_01	62	0.20	0.20	0.17	0.11	0.12	0.09
2_02	59	0.19	0.27	0.17	0.11	0.17	0.07
2_03	69	0.13	0.15	0.0	0.07	0.13	0.07
2_04	63	0.16	0.20	0.20	0.08	0.13	0.06
2_05	60	0.14	0.20	0.19	0.09	0.16	0.09
2_06	44	0.15	0.19	0.16	0.06	0.14	0.10
2_07	79	0.12	0.13	0.12	0.06	0.15	0.05
2_08	66	0.11	0.18	0.15	0.05	0.17	0.08
2_09	60	0.15	0.19	0.15	0.06	0.13	0.08
2_10	65	0.18	0.21	0.15	0.08	0.16	0.05
Mean	63	0.15	0.19	0.17	0.08	0.15	0.07

FIG. 8. Relative perceptual weights obtained by using a point-biserial correlational analysis across the six acoustic channels for keyword scoring. Error bars = standard error of the mean. Gray dotted line indicates value for equal channel weighting.

calculated over 3000 points (600 sentences $\times$ 5 keywords). Correlations were then normalized to 1.0 to reflect the relative perceptual weight assigned to each acoustic channel and are plotted in Fig. 8.

A repeated-measures ANOVA was conducted on the normalized weights for keyword scoring (Fig. 8). Significant main effects for temporal modulation [$F(1,9)=194.8$, $p < 0.001$] and frequency band [$f(2,18)=28.4$, $p < 0.001$] were found, as well as a significant modulation $\times$ band interaction [$F(2,18)=8.0$, $p < 0.01$]. Overall, listeners weighted E contributions more than TFS contributions (an average difference of 0.09) and band 2 more than either of the other two bands (by a difference of 0.07). *Post hoc t* tests demonstrated significant differences between E and TFS in all frequency bands (Bonferroni-corrected $p < 0.006$). Bands 1 and 2 were significantly different for E and TFS channels, while bands 2 and 3 were significantly different for the TFS channel only (Bonferroni-corrected $p < 0.006$). Note that no difference was obtained between bands 1 and 3 for either E or TFS channels, indicating that these are weighted similarly and that differences in frequency band weighting are driven by the mid-frequency band (528–1941 Hz). Analysis of sentence scores (where all keywords within the sentence were required to be correct) revealed the same results with the exception of no difference obtained between E and TFS weights in the mid band ($p > 0.05$).

All raw correlations (obtained across all five keywords) for each listener were significantly different from the null hypothesis of zero ($p < 0.05$), indicating that each channel received some perceptual weighting. However, in experiment 1 each frequency band and each temporal modulation property received the same perceptual weight. If that were the case in this study when these channels were presented concurrently, then we would expect them all not to differ from a relative weight of 0.17. Figure 8 displays a dotted horizontal line at this value as a reference. Points above this line indicate that acoustic channel received more perceptual weight than would be expected from an equal distribution of weighting across channels, consistent with the predictions from experiment 1 with isolated bands. Points below this line indicate less perceptual weight. Note that while each band in isolation is able to be processed equally, some chan-

nels receive more weight when acoustic cues are available from multiple channels. Namely, E2 and E3 channels received more perceptual weight ($p < 0.05$). However, this occurred at a perceptual cost of placing relatively less perceptual weight on TFS1 and TFS3 bands ($p < 0.05$).

2. Individual listeners

All listeners followed this same configuration of relative weights with few exceptions. Of the listeners who varied from the mean weighting pattern, two of them varied in three channels, subjects 2_01 and 2_07. The highest performing listener, 2_07, weighted all E channels equally as would be expected from experiment 1, but placed the most weight on TFS2 acoustics. On the other hand, 2_01 placed significantly less weight on the TFS2 channel, instead favoring more weight on E1. This individual performed close to the group mean in overall accuracy.

Correlations across each of the raw point-biserial correlation coefficients in each channel demonstrated that E1 was significantly correlated with E2 ($r=0.76$) and TFS1 ($r=0.86$) channels. Raw coefficients for E2 and TFS1 were also significantly correlated ($r=0.64$). Thus, the availability of E and TFS modulations in band 1 were influenced similarly by the noise level and were predictive of E2 weighting. Perceptual weights for E1 and TFS1 were also correlated ($r=0.65$), indicating that listeners who used E1 more also used TFS1 more.

3. Sentence context

As the stimuli used were natural sentences that may involve dynamic processing, the stability of perceptual weights across the sentence was explored. To investigate this, relative weights were calculated over the first two keywords in the sentence and compared to relative weights obtained from the last two keywords. Results indicated significant differences between the first and second halves of the sentence only for E1 and TFS3 channels ($p < 0.008$). In the second half of the sentence, less weight was given to E1 acoustics while more weight was placed on TFS3 cues. The configuration of perceptual weights over the entire sentence mostly reflects the pattern obtained for weights on the second half of the sentence.

Sentence context, which can facilitate the prediction of future words (Grosjean, 1980), was not likely to have contributed to this perceptual shift as performance was actually better on the first half of the sentence compared to the second half for all listeners by 16 percentage-points ($p < 0.001$). Instead, other sentence properties, such as intonation contours or other acoustic properties may be involved in shifting relative weights dynamically across the sentence. Alternatively, the changing perceptual weights could also reflect a dynamic perceptual processing shift, where listeners equally process the envelope in all frequency bands initially, but then focus processing on E2 as well as TFS3 cues later.

E. Discussion

1. Frequency band contributions

Individual frequency bands yielded equal performance weights in isolation (see experiment 1). However, when

multiple bands were available, differential preference was given to the mid-frequency band in this study. [Turner et al. \(1998\)](#) suggested that channels that are weighted higher in broadband conditions, as compared to isolated narrow-band conditions, indicate that such channels yield a sum greater than their parts. This suggests that important speech information is not equally distributed across the spectrum. However, the low relative weights of certain channels do not necessarily imply that those channels make no significant contributions to speech perception, or could not contribute under different listening conditions. Other convergent methods are necessary to define what actually underlies the relative perceptual weights that are obtained.

Work from [Turner et al. \(1998\)](#) and [Calandruccio and Doherty \(2007\)](#) suggests that listeners weight frequencies below 1120 Hz most when multiple frequency regions are available for IEEE sentences and nonsense syllables. This frequency region includes the low band (80–528 Hz) and part of the mid-frequency band (528–1941 Hz) used in this study. Given the high weighting of the mid-frequency region for both E and TFS acoustics, it appears that the 528–1120 Hz range in this mid band supplies crucial speech information for IEEE sentences. For the current stimuli, this band contains an overlap of F1 and F2 formant frequencies. Contributions of the mid-frequency band in this study are consistent with frequency-importance studies ([DePaolis et al., 1996](#); [Sherbecoe and Studebaker, 2002](#)) and the contribution of narrow spectral slits ([Greenberg et al., 1998](#)). Indeed, [Greenberg et al. \(1998\)](#) claimed the “supreme importance” of the mid-frequency region (750–2381 Hz) for speech intelligibility.

2. Temporal property contributions

Listeners in this study placed the most perceptual weight on E acoustics, particularly in the mid band. This is in contrast to [Apoux and Bacon \(2004\)](#) who used the correlational method with E-only nonsense syllables that contained high-rate E modulations up to 500 Hz. They found equal weight of E across frequencies in quiet and more weight on frequencies > 2400 Hz in noise. This higher frequency band has additional E modulation information (mid rate 10–25 Hz) that is not present in the lower frequencies ([Greenberg et al., 1998](#)). The difference in findings between this study and that of [Apoux and Bacon \(2004\)](#) may also reflect the difference between closed-set syllable identification and the open-set sentence tasks used here or the different frequency bands investigated.

However, the mid-frequency region in this study may be most important for E and TFS cues because of the linguistic information it conveys. [Kasturi et al. \(2002\)](#) found that E bands 300–487 Hz and 791–2085 Hz are weighted most for vowel identification and contain most of the F2 information (frequency-band contributions were flat for consonant identification). Given that vowels are essential for sentence intelligibility ([Kewley-Port et al., 2007](#); [Fogerty and Kewley-Port, 2009](#)), it is likely that this mid-frequency region contains more important amplitude modulations for the intelligibility of sentences that were used in the current study. [Phatak and Allen \(2007\)](#) also found that vowels are easier to recognize than consonants in continuous speech weighted noise, per-

haps because of the continued perceptual availability of this mid band (for both E and TFS). Both TFS and E convey distinct and important phonetic cues in this mid-frequency band between 1000 and 2000 Hz ([Ardoint and Lorenzi, 2010](#)), further supporting the higher perceptual weights for both E2 and TFS2 found in this study.

Also consistent with this study, [Hopkins and Moore \(2010\)](#) found greater perceptual benefit for adding E to the 400–2000 Hz range than to frequency bands above or below. TFS contributions above what was provided by the E alone across all frequency bands was also noted, consistent with the significant weight placed on TFS in each of the bands here (albeit significantly limited in bands 1 and 3).

The reduced weighting of TFS in the high-frequency region is consistent with reduced phase locking ([Moore, 2008](#)). Perceptually, the speech envelope contributes up to 6000 Hz, while TFS contributions begin to break down above 2000 Hz ([Ardoint and Lorenzi, 2010](#)). In contrast to the high-band TFS acoustics, it is clear that low-band TFS provides contributions to speech intelligibility ([Kong and Carlyon, 2007](#); [Li and Loizou, 2008](#); [Hopkins and Moore, 2009](#)). However, the mid-band TFS was more important during the continuous noise presented here. Therefore, it may be more important to provide TFS2, rather than TFS1, to listeners using combined acoustic and electric hearing methods in listening contexts with continuous noise.

Support for this is provided by listener 2_07 who had the best performance and placed the most weight on TFS2 with equal weights across frequency bands for E. Individual patterns such as this may highlight important training parameters to facilitate speech-recognition processes. TFS in the mid-frequency range may provide a greater perceptual benefit than low-frequency information, particularly because perceptual use of E1 and TFS1 was significantly correlated. Therefore, these channels may contain overlapping or correlated perceptual information. E1 and TFS1 do not appear to be completely independent perceptual channels. This may be because both E1 and TFS1 appear to provide cues to F0, coded in the repetition rate of the waveform or the resolution of harmonics (discussed by [Hopkins and Moore, 2010](#)).

VI. GENERAL DISCUSSION

The purpose of this set of experiments was to investigate the relative importance of envelope and fine structure acoustic cues across linguistically significant frequency bands during the perception of uninterrupted speech in noise. Experiment 1 demonstrated that equal relative perceptual weight was placed on each acoustic channel when that channel's availability was individually varied across a range of SNRs. All acoustic channels were equally susceptible to degradation by varying the SNR. Perhaps due to possible synergistic interactions, experiment 2 demonstrated that these perceptual weights do not remain equal when multiple acoustic information channels are concurrently available. Instead, relative to other cues, listeners placed most of their perceptual weight on E and band-2 acoustic information.

Of particular importance may be the perceptual weighting observed in the mid-frequency band (500–2000 Hz).

During meaningful, uninterrupted sentences, listeners weighted both E and TFS cues the highest in this region. This suggests that preserving the acoustic cues presented in this band may be mandatory for maximal speech understanding.

Particularly interesting is the perceptual weighting of TFS acoustics. Traditionally, TFS in the low-frequency region has been viewed as being most important for speech perception (Hopkins *et al.*, 2008; Hopkins and Moore, 2009), partly because of its involvement conveying pitch (see Moore, 2008), and therefore has been an important consideration for facilitating E-only speech (Kong and Carlyon, 2007; Li and Loizou, 2008). However, it now appears that TFS contributes to speech perception across the entire frequency spectrum of speech (Hopkins and Moore, 2010) and the current study argues that TFS between 500 and 2000 Hz may be most important for speech in continuous noise, perhaps because it codes the formant frequency dynamics important for place cues (Rosen, 1992). This may help explain why vowels—predominantly conveyed by F1 and F2 formants—are essential for sentence intelligibility (Kewley-Port *et al.*, 2007; Fogerty and Kewley-Port, 2009).

It was surprising that prosodic band-1 information did not receive more weight. The prosodic contour of speech conveyed by the F0 is important for speech perception during quiet and competing contexts (e.g., Laures and Weismer, 1999; Darwin *et al.*, 2003; Winfield *et al.*, 1989). Therefore, it may make contributions, particularly at low SNRs or in the context of competing signals. These lower frequencies provide supralinguistic cues for syntactic units and source segregation (e.g., Leiste, 1970; Darwin *et al.*, 2003). Further research is needed to explore the possibility of contributions of low-weighted acoustics, such as from this band.

VII. SUMMARY AND CONCLUSIONS

This study investigated how listeners perceptually weight envelope and fine structure acoustic information across the frequency spectrum. These cues were presented under two listening conditions: (1) when each acoustic property was individually presented and (2) when multiple acoustic properties were concurrently presented. By varying the SNR in each channel independently, an estimate of the relative perceptual weight a listener placed on that cue was obtained for each listening condition. The results of this study demonstrate the following:

- (1) Equal contributions to sentence intelligibility were obtained for three frequency bands conveying predominant prosodic, sonorant, and obstruent/fricative linguistic cues when each band was tested individually in a masked wideband context.
- (2) A method that manipulated the SNR in order to vary the availability of temporal-modulation information was validated and demonstrated equal perceptual weighting of envelope and fine structure acoustics processed in three frequency bands over an 80–6400 Hz range.
- (3) When multiple cues were simultaneously available for perception (Experiment 2), the mid-frequency band (528–1941 Hz) was weighted the most for both the enve-

lope and fine structure. Overall, across all spectral regions, the envelope was the predominant cue that listeners used.

- (4) Little perceptual weight was placed on low-frequency envelope or fine structure cues in uninterrupted speech contexts. It may be that acoustic channels that received little perceptual weight, and therefore had performance that was not strongly associated with the variations in SNR in that channel, still provide significant contributions even when highly degraded.

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. Larry Humes also provided helpful suggestions regarding an earlier version of this manuscript. Kate Giesen and Eleanor Barlow assisted in scoring participant responses. 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 (1997). ANSI S3.5–1997, Methods for the Calculation of the Speech Intelligibility Index (American National Standards Institute, New York).
- ANSI (2004). ANSI S3.6–2004, Specifications for Audiometers (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.
- Berg, B. G. (1989). “Analysis of weights in multiple observation tasks,” *J. Acoust. Soc. Am.* **86**, 1743–1746.
- Berg, B. G., and Green, D. M. (1990). “Spectral weights in profile listening,” *J. Acoust. Soc. Am.* **88**, 758–766.
- Berg, B. G., and Green, D. M. (1992). “Discrimination of complex spectra: Spectral weights and performance efficiency,” in *Auditory Physiology and Perception*, edited by Y. Cazals, L. Demany, and K. Homer (Pergamon, London), pp. 373–379.
- Calandruccio, L., and Doherty, K. A. (2007). “Spectral weighting strategies for sentences measured by a correlational method,” *J. Acoust. Soc. Am.* **121**, 3827–3836.
- Calandruccio, L., and Doherty, K. A. (2008). “Spectral weighting strategies for hearing-impaired listeners measured using a correlational method,” *J. Acoust. Soc. Am.* **123**, 2367–2378.
- 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.
- DePaolis, R. A., Janota, C. P., and Frank, T. (1996). “Frequency importance functions for words, sentences, and continuous discourse,” *J. Speech Hear. Res.* **39**, 714–723.
- 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.
- Dorman, M. F., Loizou, P. C., and Rainey, D. (1997). “Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs,” *J. Acoust. Soc. Am.* **102**, 2403–2411.
- Drullman, R., Festen, J. M., and Plomp, R. (1994a). “Effect of temporal envelope smearing on speech reception,” *J. Acoust. Soc. Am.* **95**, 1053–1064.
- Drullman, R., Festen, J. M., and Plomp, R. (1994b). “Effect of reducing slow temporal modulations on speech reception,” *J. Acoust. Soc. Am.* **95**, 2670–2680.
- Dubbelboer, F., and Houtgast, T. (2008). “The concept of signal-to-noise ratio in the modulation domain and speech intelligibility,” *J. Acoust. Soc. Am.* **124**, 3937–3946.

- Fletcher, H., and Galt, R. (1950). "Perception of speech and its relation to telephony," *J. Acoust. Soc. Am.* **22**, 89–151.
- Fogerty, D., and Humes, L. E. (2010). "Perceptual contributions to monosyllabic word intelligibility: Segmental, lexical, and noise replacement factors," *J. Acoust. Soc. Am.* **128**, 3114–3125.
- Fogerty, D., and Kewley-Port, D. (2009). "Perceptual contributions of the consonant-vowel boundary to sentence intelligibility," *J. Acoust. Soc. Am.* **126**, 847–857.
- French, N. E., and Steinberg, J. C. (1947). "Factors governing the intelligibility of speech sounds," *J. Acoust. Soc. Am.* **19**, 90–119.
- 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.
- Grant, K. W., and Braid, L. D. (1991). "Evaluating the articulation index for auditory-visual input," *J. Acoust. Soc. Am.* **89**, 2952–2960.
- Grant, K. W., and Walden, B. E. (1996). "Spectral distribution of prosodic information," *J. Speech Hear. Res.* **39**, 228–238.
- Greenberg, S., Arai, T., and Silipo, R. (1998). "Speech intelligibility derived from exceedingly sparse spectral information," in *Proceedings of the International Conference on Spoken Language Processing*, pp. 2803–2806.
- Grosjean, F. (1980). "Spoken word recognition processes and the gating paradigm," *Percept. Psychophys.* **28**, 267–283.
- 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.
- Hopkins, K., Moore, B. C. J., and Stone, M. A. (2008). "Effects of moderate cochlear hearing loss on the ability to benefit from temporal fine structure information in speech," *J. Acoust. Soc. Am.* **123**, 1140–1153.
- Houtgast, T., and Steeneken, H. J. M. (1985). "A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria," *J. Acoust. Soc. Am.* **77**, 1069–1077.
- 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**(Pt.1), 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.
- Laures, J., and Weismer, G. (1999). "The effect of flattened F0 on intelligibility at the sentence-level," *J. Speech Lang. Hear. Res.* **42**, 1148–1156.
- Leiste, I. (1970). *Suprasegmentals* (MIT Press, Cambridge, MA), pp. 1–194.
- Lentz, J. J., and Leek, M. R. (2002). "Decision strategies of hearing-impaired listeners in spectral shape discrimination," *J. Acoust. Soc. Am.* **111**, 1389–1398.
- 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.
- 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. A., Heise, G. A., and Lichten, W. (1951). "The intelligibility of speech as a function of the context of test material," *J. Exp. Psychol.* **41**, 329–335.
- Moore, B. C. J. (2008). "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., and Glasberg, B. R. (1983). "Suggested formulae for calculating auditory-filter bandwidths and excitation pattern," *J. Acoust. Soc. Am.* **74**, 750–753.
- Moore, D. R., Schnupp, J. W. H., and King, A. J. (2001). "Coding the temporal structure of sounds in auditory cortex," *Nat. Neurosci.* **4**, 1055–1056.
- Owren, M. J., and Cardillo, G. C. (2006). "The relative roles of vowels and consonants in discriminating talker identity versus word meaning," *J. Acoust. Soc. Am.* **119**, 1727–1739.
- Phatak, S. A., and Allen, J. B. (2007). "Consonant and vowel confusions in speech-weighted noise," *J. Acoust. Soc. Am.* **121**, 2312–2326.
- 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. London, Ser. B*, **336**, 367–373.
- Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., and Ekelid, M. (1995). "Speech recognition with primarily temporal cues," *Science* **270**, 303–304.
- 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.
- Sherbecoe, R. L., and Studebaker, G. A. (2002). "Audibility-index functions for the connected speech test," *Ear Hear.* **23**, 385–398.
- Smith, Z. M., Delgutte, A. J., and Oxenham, A. J. (2002). "Chimaeric sounds reveal dichotomies in auditory perception," *Nature* **416**, 87–90.
- Turner, C. W., Kwon, B. J., Tanaka, C., Knapp, J., Hubbart, J. L., and Doherty, K. A. (1998). "Frequency-weighting functions for broadband speech as estimated by a correlational method," *J. Acoust. Soc. Am.* **104**, 1580–1585.
- Turner, C., Mehr, M., Hughes, M., Brown, C., and Abbas, P. (2002). "Within-subject predictors of speech recognition in cochlear implants: A null result," *ARLO* **3**, 95–100.
- van der Horst, R., Leeuw, A. R., and Dreschler, W. A. (1999). "Importance of temporal-envelope cues in consonant recognition," *J. Acoust. Soc. Am.* **105**, 1801–1809.
- Warren, R. M., Reiner, K. R., Bashford, J. A., Jr., and Brubaker, B. S. (1995). "Spectral redundancy: Intelligibility of sentences heard through narrow spectral slits," *Percept. Psychophys.* **57**, 175–182.
- Wingfield, A., Lahar, C. J., and Stine, E. A. L. (1989). "Age and decision strategies in running memory for speech: Effects of prosody and linguistic structure," *J. Gerontol. Psychol. Sci.* **44**, 106–113.
- Xu, L., and Pongst, B. E. (2003). "Relative importance of temporal envelope and fine structure in lexical-tone perception," *J. Acoust. Soc. Am.* **114**(Pt.1), 3024–3027.
- Xu, L., Thompson, C. S., and Pongst, B. E. (2005). "Relative contributions of spectral and temporal cues for phoneme recognition," *J. Acoust. Soc. Am.* **117**, 3255–3267.