The role of vowel and consonant fundamental frequency, envelope, and temporal fine structure cues to the intelligibility of words and sentences[a)]

Daniel Fogerty[b)] and Larry E. Humes
Department of Speech and Hearing Sciences Indiana University Bloomington, Indiana 47405

(Received 8 April 2011; revised 16 November 2011; accepted 21 December 2011)

The speech signal contains many acoustic properties that may contribute differently to spoken word recognition. Previous studies have demonstrated that the importance of properties present during consonants or vowels is dependent upon the linguistic context (i.e., words versus sentences). The current study investigated three potentially informative acoustic properties that are present during consonants and vowels for monosyllabic words and sentences. Natural variations in fundamental frequency were either flattened or removed. The speech envelope and temporal fine structure were also investigated by limiting the availability of these cues via noisy signal extraction. Thus, this study investigated the contribution of these acoustic properties, present during either consonants or vowels, to overall word and sentence intelligibility. Results demonstrated that all processing conditions displayed better performance for vowel-only sentences. Greater performance with vowel-only sentences remained, despite removing dynamic cues of the fundamental frequency. Word and sentence comparisons suggest that the speech envelope may be at least partially responsible for additional vowel contributions in sentences. Results suggest that speech information transmitted by the envelope is responsible, in part, for greater vowel contributions in sentences, but is not predictive for isolated words. © *2012 Acoustical Society of America*. [DOI: 10.1121/1.3676696]

PACS number(s): 43.71.Gv, 43.71.Es [MSS] Pages: 1490–1501

I. INTRODUCTION

It is well known that speech is a complex, temporally varying acoustic signal. This signal is composed of many different acoustic and linguistic properties that may, to varying degrees, be informative for understanding the intended message. While some acoustic cues may be correlated and provide similar information for perceiving speech, it is likely that many cues provide different information, and therefore serve different roles for understanding speech. Clearly defining the roles of different auditory speech properties is essential for the design of more advanced signal-processing technology so that processing can focus on preserving and enhancing the particular auditory speech cues that are most informative in a given auditory environment. However, given the highly complex nature of speech, the initial investigation of auditory cue contributions must focus on highly defined classes of speech sounds, with future subdivisions of these classes possible. Previous studies (e.g., Cole *et al.*, 1996; Owren and Cardillo, 2006; Kewley-Port *et al.*, 2007) have focused on using the defined classes of consonants and vowels as one way to divide this acoustic space. Such a division is advantageous for three primary reasons.

First, for auditory speech presentations, consonants and vowels are characterized by fundamentally different acoustic features (Ladefoged, 2001; Stevens, 2002). Whereas coarticulation blurs the boundaries between consonants and vowels, such that even the division of speech sounds into these two general categories may be considered to be merely a convenience (see Ladefoged, 2001), it is generally acknowledged that the acoustic speech stream does contain portions that exhibit predominantly "vowel" or "consonant" characteristics.

Second, a number of studies have now identified differences between the contributions of speech acoustics contained within the putative boundaries of phonetician-defined vowels and consonants (e.g., Cole *et al.*, 1996; Owren and Cardillo, 2006; Kewley-Port *et al.*, 2007). Namely, acoustic cues during the defined vowels have been found to contribute more than during the consonant (Owren and Cardillo, 2006). Furthermore, this is true even when transitional information or durational differences between the consonants and vowels are taken into consideration (Fogerty and Kewley-Port, 2009). Therefore, regardless of whether that acoustic information originated from the production of the concurrent segment (e.g., vowel), or from a neighboring segment (e.g., consonant), the data suggest that the acoustic information present during vowels is essential for speech intelligibility. A general acoustic property responsible for providing that essential information, not limited by ad hoc phonemic conventions, is yet to be defined. The current study represents a beginning of such an investigation. Defining the acoustic locus of this essential vowel information will assist in identifying highly important acoustic properties for speech intelligibility that may be defined on the signal level, rather than the phonemic level of speech analysis.

Third, a number of studies have begun to demonstrate dissociations between the functional roles of consonants and

[a)]Portions of this data were previously presented at the 2011 Cognitive Hearing Science for Communication Conference, Linkoping, Sweden.

[b)]Author to whom correspondence should be addressed. Electronic mail: fogerty@sc.edu. Current address: University of South Carolina, Department of Communication Sciences and Disorders, 1621 Greene St., Columbia, SC 29208.

 0001-4966/2012/131(2)/1490/12/$30.00

vowels (Nespor *et al.*, 2003; New *et al.*, 2008; Toro *et al.*, 2008; Carreiras and Price, 2008; Carreiras *et al.*, 2009). Therefore, the roles of acoustic properties associated with consonants or vowels may also differ.

Granted, using such a pre-defined class of sounds for the partitioning of an acoustic stimulus space is not without its limitations. For example, consonants and vowels each provide information about each other due to overlapping productions (e.g., Liberman *et al.*, 1967; Strange *et al.*, 1983). In sentences, this overlap likely becomes even more pronounced. However, the current investigation does not investigate the identification of individual phonemes. Instead, it is an investigation of how temporal speech segments containing the predominant acoustic properties associated with consonants or vowels convey cues regarding the entire speech utterance. The coarticulatory and suprasegmental consequences of articulatory production likely impose utterance-level cues differentially upon these predominant consonant and vowel units. The current study investigates the importance of several suprasegmental cues to the contribution of consonants and vowels to word recognition.

Furthermore, the investigation of consonant and vowel contributions has led to an important finding: the importance of the vowel. In these studies, word scores for repeating an entire sentence back were better for sentences containing only the vowels (i.e., consonants replaced by noise or silence) than sentences containing only the consonants (i.e., vowels replaced). This finding has been replicated repeatedly (Cole *et al.*, 1996; Kewley-Port *et al.*, 2007; Fogerty and Kewley-Port, 2009), often showing a large two-to-one advantage of sentences preserving the vowels. Even Stilp and Kluender (2010), who argue against segmental contributions, found better performance for vowels compared to consonants when duration was equated. Furthermore, their measure of cochlea-scaled spectral entropy emphasized the importance of vowel acoustics. Vowel segments explained 55% of the variance on this measure ($r = 0.74$), although they state it was a non-significant predictor. Thus, it appears that vowels highlight an important locus of speech information.

The current investigation expanded upon previous segmental studies by exploring three general acoustic properties present in consonant and vowel segments: amplitude envelope (E), temporal fine structure (TFS), and fundamental frequency (F_0). Therefore, this study explored the role of these three general acoustic properties in explaining the consonant/vowel data that have been obtained to date. In such a way, a general acoustic property might be identified as being responsible for the overwhelming importance of information contained within the vowel segment during sentences.

A. Potential acoustic cues

This study explicitly tests how these different types of acoustic information, namely, F_0, E, and TFS, explain relative contributions of consonants and vowels to the recognition of words in sentences and words in isolation. As noted, the relative contribution of vowels and consonants varies in these two speech contexts. In particular, the relative contribution of vowels is greater than that of consonants in a sentence context, but not in isolated words (for a discussion, see Fogerty and Humes, 2010). Of particular interest to this study is whether the enhanced importance of vowels in sentence contexts is explained by the contribution of F_0, E, or TFS information conveyed during the vowel segments of sentences.

1. Fundamental frequency

One property explored by the current study is the contribution of the F_0 contour to sentence and word intelligibility. The natural kinematics of F_0 have been demonstrated to facilitate speech recognition (Laures and Weismer, 1999) as this intonation information facilitates the prediction of syntactic units (Wingfield *et al.*, 1989). Fogerty and Humes (2010) recently proposed that F_0 may be a potential cue that vowels are more apt to carry than consonants, and would likely facilitate sentence recognition more so than isolated words. While studies such as those conducted by Laures and Wesimer (1999) clearly demonstrate the contribution of natural F_0 variation on sentence intelligibility, it has not yet been demonstrated: (1) if those cues are limited to sentence contexts with predictable syntactic constraints; (2) if vowels, the primary periodic element, contain more of these F_0 cues than consonants; or (3) if natural speech periodicity alone provides meaningful cues for word and sentence recognition. The current study was designed to test these questions by using meaningful sentences and isolated words, by examining the contribution of consonants and vowels, and by modifying the natural speech F_0 to be either flat (i.e., removing natural F_0 variations) or removed (i.e., removing speech periodicity, simulating aperiodic, whispered speech).

As noted, one potentially informative acoustic cue for speech recognition, particularly during longer speech samples and during interruption is the time-varying F_0 of the utterance. The prosodic pitch contour conveyed by the F_0 has been demonstrated to aid in sentence recognition tasks, as flattening the pitch contour reduces intelligibility (Laures and Wiesmer, 1999; Watson and Schlauch, 2008). In addition to enhancing syntactic predictability, F_0 cues are believed to aid in stream segregation (Bregman *et al.*, 1990) and facilitate the integration of speech glimpses during interrupted contexts (Darwin *et al.*, 2003); two properties that may be important for understanding sentences under a noise replacement paradigm where alternating segments (i.e., consonants or vowels) are replaced by noise. As the primary periodic intervals of speech, vowel units include the majority of this dynamic F_0 information. This evidence suggests that F_0 contour cues may play a vital role in recognizing words from vowel segments in sentence contexts.

2. Envelope

In addition to F_0, speech prosody also includes the temporal cues conveyed by the speech envelope (E). The speech E is composed of relatively slow modulations of amplitude over time. These temporal amplitude cues facilitate word prediction above what is provided by phonetic information alone (Waibel, 1987). Furthermore, E information has been demonstrated to convey manner and voicing cues (Apoux and Bacon, 2004; Gallun and Souza, 2008; Rosen, 1992;

Shannon *et al.*, 1995). The importance of E information for general speech intelligibility has been demonstrated by a number of studies (e.g., Dorman *et al.*, 1997; Shannon *et al.*, 1995). With E cues available in only three different spectral bands, the number of bands used in the current study, Shannon and colleagues (1995) demonstrated moderately high levels of word recognition in sentences, as well as for consonant and vowel identification in "aCa" and "hVd" phonemic contexts, respectively.

The predominant modulation rate of the speech envelope occurs at around 4 Hz and is correlated across all frequency regions (Crouzet and Ainsworth, 2001; Greenberg *et al.*, 1998). Adjacent frequency bands are also more highly correlated than widely separated bands (Steeneken and Houtgast, 1999). This 4-Hz rate corresponds to the dominant syllabic rate of speech. Vowels are a primary element of the syllable (i.e., nucleus) that may best provide this predominant rate information. (On average, in English consonants occur more frequently than vowels and therefore have a higher segment modulation rate.) Evidence for vowels carrying the predominant rate information of speech is provided by several studies that have demonstrated that the perceptual timing of speech is aligned to the vowel (reviewed by Port, 2003), even across languages with different rate timing (Tajima and Port, 2003). Vowels are also the primary carriers of stress, which is characterized by amplitude and duration changes. Both of these changes directly modulate the speech E and have expressed linguistic meaning, such as distinguishing between a noun and a verb (e.g., as in the word "present"). Therefore, it may be that E cues conveyed during vowels have specific consequences regarding speech recognition performance in sentences. Through comparisons of stimuli that either have or do not have the predominant E preserved, the current study investigates how (or if) this information contributes generally to the contribution of vowels in meaningful sentences.

3. Temporal fine structure

The third acoustic property investigated in this study is temporal fine structure (TFS), which conveys relatively fast frequency modulations over time. TFS information has been demonstrated to be most important for place of articulation (Apoux and Bacon, 2004; Rosen, 1992). However, this dynamic frequency information is believed to be important for "dip listening" (see Lorenzi *et al.*, 2006), or in extracting speech information in brief periods of preserved speech between intervening fluctuating or interrupting maskers. This ability to "glimpse" speech between interruptions may be essential for word recognition in the segmental studies investigating consonant and vowel contributions because of the methodological paradigm. As reviewed, these studies use a noise replacement paradigm. Therefore, sentences (or words) are temporally interrupted. The TFS may provide cues to extract speech information from the remaining speech fragments. What is currently not known, and is under investigation here, is if consonants and vowels carry TFS information necessary for this glimpsing process equally. It may be that vowels provide more robust TFS cues under such conditions, which then results in the observed "vowel advantage." Evidence for the use of TFS cues during interruption comes from studies of E-only speech where the TFS cues are explicitly removed. Listeners with normal hearing and with cochlear implants do not receive a perceptual benefit for interrupted or fluctuating maskers over continuous maskers when they receive E-only speech, although they do for natural speech that preserves TFS information (Nelson and Jin, 2004; Nelson *et al.*, 2003). Fogerty (2011) recently demonstrated that listeners perceptually weight TFS cues most in the mid frequency region conveying predominant cues from the first and second formants. Vowel segments primarily convey the harmonic structure of speech and F1 and F2 dynamics which change substantially over time (Nearey and Assmann, 1986; Hillenbrand and Nearey, 1999). In addition, the TFS may best capture the most varying part of the vowel (i.e., formant transitions), supported by evidence of TFS conveying place cues (Rosen, 1992), which provide significant information regarding neighboring consonants (e.g., Liberman *et al.*, 1967). TFS cues may capture this "consonant" information present within the vowel and lead to probabilistic linguistic cues in sentences. Therefore, the TFS cues present during the vowel segment may be especially important for glimpsing speech during interrupted utterances employed during segment replacement. The current study investigated TFS contributions during glimpsed portions of vowels or consonants.

B. Temporal interruption

As already noted, the relative contributions of consonants and vowels during acoustic presentations have typically been investigated using a noise replacement paradigm. For this method, listeners are tested on consonant-only materials, where all of the vowels are replaced by noise and only the consonants remain, or vowel-only materials, where all of the consonants are replaced and the vowels are preserved. While segmentation procedures follow commonly accepted acoustic landmarks (see Stevens, 2002) for the onset of segmental information, arguably a discrete division between consonants and vowels is not possible. Segments will contain overlapping information. Therefore, previous studies have shifted the location of these segmental boundaries to include either more or less acoustic information within each segmental type (Fogerty and Humes, 2010; Fogerty and Kewley-Port, 2009). These studies have observed that shifting the boundary, in general, modifies performance according to the proportion of the total duration (PTD) for the speech utterance presented. However, such boundary modifications do not change the perceptual difference observed between consonants and vowels when accounting for the PTD. For example, vowel-only sentences result in nearly a 40 percentage-point improvement in word recognition scores over consonant-only sentences across all PTD values while no difference is observed for isolated words (Fogerty and Humes, 2010). Vowels actually account for proportionately less of the sentence than consonants (45% vs 55%, respectively, Fogerty and Kewley-Port, 2009; Ramus *et al.*, 1999) and segmental boundaries need to be shifted so that

consonants account for more than two-thirds of the sentence before perceptual performance of consonant-only sentences exceeds that of vowel-only sentences at the generally accepted segment boundary. These results suggest that estimates of vowel and consonant contributions at the default or typical boundary, defined by common conventions, are appropriate. The current study does not focus on changes in consonant and vowel contributions as a function of boundary location, or the PTD, as the previous studies have accomplished this. Rather, the current study explored the contributions of acoustic properties, specifically F_0, E, and TFS, within each segment.

C. Isolated word versus sentence contexts

The role of the acoustic cue underlying vowel contributions is likely to be dependent upon the acoustic/linguistic context. This is because the relative contribution of vowels is different in isolated words than it is in sentences. Vowels provide large benefits over consonants for speech intelligibility of sentences (Kewley-Port *et al.*, 2007; Fogerty and Kewley-Port, 2009), yet no such difference is apparent in isolated words (Owren and Cardillo, 2006; Fogerty and Humes, 2010). This finding might, in part, be due to fundamental differences between words and sentences.

Production of words within a sentence context results in a number of acoustic consequences. Vowel reduction (Lindblom, 1963), increased co-articulation (Kent and Minifie, 1977), and prosodic changes (i.e., intensity, F_0, and duration) (Shattuck-Hufnagel and Turk, 1996), all result in different acoustic realizations of words when spoken in sentence contexts. These acoustic modifications of the word production result in poor recognition of words excised from sentences (Pickett and Pollack, 1963). Indeed, in a study of naturally produced speech, reduced word forms were recognized only in the acoustic context of sentences, not with other contextual information, such as the visually printed sentence context (Janse and Ernestus, 2011). While reduced words may be recognized with 52% accuracy in isolation, performance improves to 92% in the acoustic context of the sentence (Ernestus *et al.*, 2002). These findings support an already established literature of long-distance effects of the acoustic context (e.g., Coleman, 2003; Local, 2003) which may make the speech signal perceptually coherent (Hawkins, 2003). Indeed, the identification of a vowel in a "bVt" context can be influenced by the preceding acoustic context in sentences (Ladefoged and Broadbent, 1957). Furthermore, production of words spoken in sentences results in speech rate changes that cause nonlinear reductions in duration (discussed by Janse, 2004). Consonants are preserved more than vowels (Gay, 1978; Max and Coleman, 1997) as are stressed, compared to unstressed, syllables (Port, 1981). Such nonlinear reductions in duration alter the speech envelope nonlinearly, which may provide added information contained in the envelope of spoken words in sentences.

In addition to different acoustic realizations of words in sentences, it is well known that sentences provide added linguistic context that facilitates the predictability of words (e.g., Miller *et al.*, 1951). However, there is no reason to suspect that linguistic probability in sentences would favor the contributions of vowels over those of consonants, or vice versa, unless there was also specific acoustic information carried by one or the other segment type that provides linguistic constraints on word probabilities. As noted above, F_0 and E cues may provide such constraints. Thus, there are two primary factors that are different about word recognition in isolation and in sentences. First, the acoustic realization (i.e., different production) of words in sentences results in nonlinear changes at the word level as well as long-distance effects on the sentence. Importantly, it is the acoustic context, not linguistic context alone, which best facilitates the recognition of reduced words in sentences (Janse and Ernestus, 2011). Second, linguistic context facilitates the predictability of words in sentences. However, in order to take advantage of this linguistic (and non-auditory) benefit, the listener must be able to extract sufficient acoustic information for initial lexical access. In addition, these two properties appear to be related, as the intelligibility of excised words is inversely proportional to the predictability of the linguistic context (Leiberman, 1963).

These findings suggest several reasons why the contribution of different acoustic speech properties might be different in isolated words compared to words spoken in a sentence context. Important for this study are direct acoustic differences that impact the acoustic realization of consonants and vowels separately, most related to duration reductions (e.g., Max and Caruso, 1997) that may have the greatest impact on the speech envelope. Furthermore, as reviewed previously, sentences with meaningful syntax may directly influence the natural F_0 variations as well as the E. These findings motivate the use of meaningful sentences, compared to isolated words, and the stimulus manipulations used in this study to directly control these acoustic properties of interest.

D. Purpose

Previous research has demonstrated that vowel segments contribute more than consonant segments during sentence contexts, but not in isolated word contexts (Fogerty and Humes, 2010; Fogerty and Kewley-Port, 2009). The current study was designed to investigate potential acoustic contributions to this vowel advantage that is specific to the linguistic context. Dynamic information conveyed by the fundamental frequency, amplitude envelope, and temporal fine structure was investigated in sentence and word contexts.

II. EXPERIMENT 1: THE ROLE OF F_0 IN VOWEL AND CONSONANT CONTRIBUTIONS

A. Listeners

Fourteen normal-hearing listeners (M = 21 yr, 19–32 yr) were paid to participate in this study. All listeners were native speakers of American English and had pure-tone thresholds no greater than 20 dB HL at octave intervals from 250 to 8000 Hz (ANSI, 2004). Listeners were randomly assigned to one of two experimental groups.

B. Stimuli and design

This study used a 2 (listener group) by 2 (context) by 2 (segmental condition) mixed model design with listener group as the between-subject factor. Listeners were randomly assigned to one of two listening groups corresponding to the flat F_0 (FlatF_0) or removed F_0 (NoF_0) listening conditions. Thus, the processing conditions associated with these listener groups limited the speech information conveyed by F_0. Each listener group heard two types of speech materials: sentences and monosyllabic CVC words. These speech materials were further processed using noise replacement (described below) to preserve only the vowels (V-only) or only the consonants (C-only). In addition, all listeners completed a Full-utterance condition without segment replacement (Full) in 0-dB SNR steady-state noise to determine baseline word recognition abilities under these processing conditions.

1. Stimulus manipulation

The contribution of the prosodic pitch contour was limited in two ways. The FlatF_0 listener group heard speech materials with the natural F_0 variations flattened at the mean value for the utterance. In contrast, the NoF_0 group received sentences for which the natural F_0 information was removed, resulting in speech aperiodicity similar to whispered speech.

Both processing conditions followed the same initial procedure. Figure 1 displays the waveform and spectrogram for a sample of these materials. Stimuli were analyzed by STRAIGHT (Kawahara *et al.*, 1999), a speech analysis and synthesis software program implemented in MATLAB. The F_0 source information was extracted. For the FlatF_0 group, this naturally varying raw F_0 information was replaced by a constant value at the mean F_0. Unvoiced portions were preserved as unvoiced. This flattened F_0 source was then used to resynthesize the speech sample, resulting in high-fidelity speech. Normal speech has normal cycle-to-cycle variations in speech periodicity in addition to intonation (i.e., pitch) changes. This processing method explicitly removed all such variations, resulting in "robotic" sounding monotone speech. Figure 1 (see middle column) displays preservation of the voiceless consonant features and spectral characteristics of the vowels for this FlatF_0 condition. For the NoF_0 group, the F_0 was instead replaced by zeros before resynthesizing the speech sample. This modification resulted in aperiodic speech, as can be observed by the aperiodicity of the waveform and lack of glottal pulsing for NoF_0 in Fig. 1 (see right column). This method also preserved all voiceless consonant information and the spectral characteristics of the speech stimulus (i.e., note the preserved frication, noise bursts, and formant bands for NoF_0 in Fig. 1). This processing resulted in the stimulus sounding similar to whispered speech. However, simulation of whispering would also require modification of the spectral envelope (Fujimura and Lindqvist, 1971), which was not desirable in the current study investigating F_0 contributions.

2. Speech materials

All listeners completed an open-set speech-recognition task with sentences and with monosyllabic consonant-vowel-consonant (CVC) words. Forty-two sentences were selected from the TIMIT database (Garofolo *et al.*, 1990, www.ldc.upenn.edu). Each sentence was spoken by a different talker (21 male, 21 female) from the North Midland dialect region. Sentences averaged eight words per sentence (mean duration: 2480 ms; C: 62 ms; V: 97 ms). As excised words provide poor acoustic realizations of the target for isolated word recognition (Pickett and Pollack, 1963), isolated words were selected from a different database. CVC words (N = 148) were selected from recordings by Takayanagi *et al.* (2002) (mean duration: 437 ms; C: 119 ms; V: 200 ms). These words were recorded by one male talker reported to be of General American dialect and represent two levels of lexical difficulty as determined by the neighborhood activation model (Luce and Pisoni, 1998). Individual words in

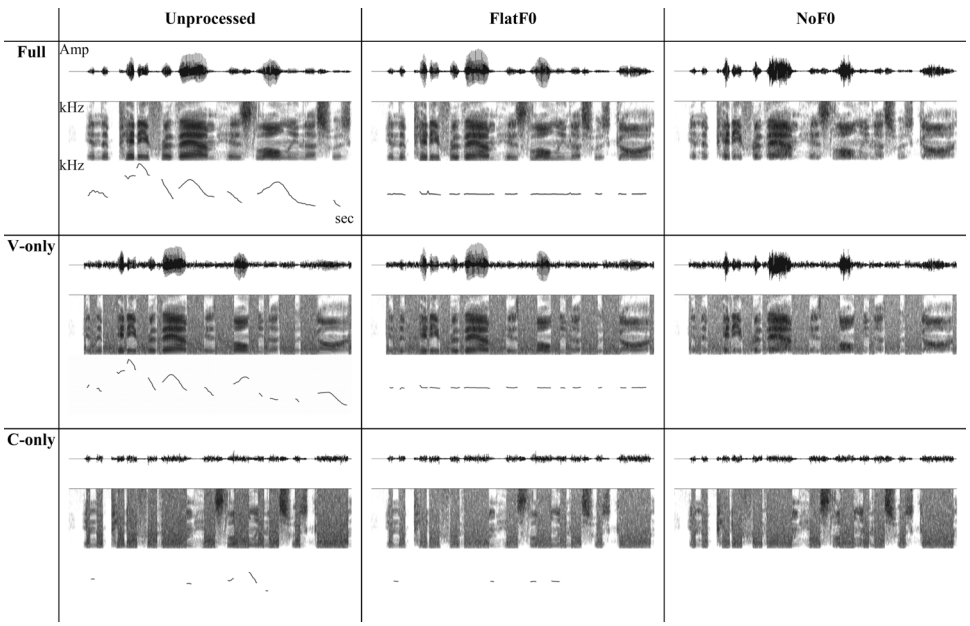

FIG. 1. Stimulus processing examples for experiment 1. The original unprocessed sentence, "In the pity for them his loneliness was gone," is displayed in the left column under full, V-only, and C-only conditions for reference. Stimulus examples for the FlatF_0 and NoF_0 listener groups are displayed in the center and right columns, respectively. Each cell displays the amplitude waveform, spectrogram (0–6 kHz), and F_0 pitch contour (75–225 Hz) for that condition and listener group. No pitch contour is displayed for the NoF_0 group because this regular periodicity was explicitly removed during processing, creating aperiodic speech. Amplitude waveforms were normalized prior to segmentation.

sentences also have varied levels of lexical difficulty, both specific to the word and dependent upon neighboring context. These same sentence and CVC materials have been used in previous noise- replacement studies (Fogerty and Humes, 2010; Fogerty and Kewley-Port, 2009). All words and sentences were normalized in RMS amplitude and presented at a level of 70 dB SPL prior to noise replacement. This ensured that intrinsic level differences between consonants and vowels, present in natural speech, were preserved. For both words and sentences, nominal consonant segments accounted for 55% of the utterance duration, while nominal vowel segments accounted for 45%, consistent with other measurements of English (Ramus *et al.*, 1999).

3. Noise replacement

Noise replacement occurred after the full sentences were processed according to the procedures for each listening group. Segmental boundaries were previously specified in the TIMIT database by experienced phoneticians and the CVC words used here modeled these boundary marking rules (Fogerty and Humes, 2010; Zue and Seneff, 1988). Segmentation used prominent acoustic landmarks, such as abrupt acoustic changes associated with the boundaries for stops and dividing formant transitions in half during slow periods of change for liquids (as these transitions provide information regarding both phonemes). The full stop closure and burst was assigned to the consonant. Vowels followed by /r/ were treated as single rhotocized vowels. Segmental boundaries were adjusted within 1-ms to the nearest local amplitude minima (i.e., zero-crossing) to minimize introduction of transients. A speech-shaped noise matching the long-term average speech spectrum (LTASS) of a concatenation of all experimental test words or sentences was created, scaled to −16 dB relative to the speech level, and used for replacement of the segments. This noise level was significantly below that average vowel and average consonant level and was selected to allow for some perception of stimulus continuity without providing perceptual filling-in of the missing acoustic information. Use of a standard noise level for all replacement intervals also avoided providing gross amplitude information regarding the missing segment. However, the noise level used in replacement does not influence observed consonant and vowel contributions in isolated words (Fogerty and Humes, 2010). Furthermore, while Kewley-Port *et al.* (2007) used different noise levels for vowel and consonant replacement, they obtained very similar findings to Fogerty and Kewley-Port (2009) who used a single standardized replacement level: −16 dB SNR, the same level used in the current study. Indeed, the same relative roles of consonants and vowels have been observed in the absence of noise during replacement (Cole *et al.*, 1996; Owren and Cardillo, 2006).

Two different noise spectra were created, one used for CVC words and one used for the TIMIT sentences. These two different noises were created identically, with the exception that one used the CVC LTASS and one used the TIMIT LTASS. A unique noise waveform was used for all replacement intervals within a given sentence or word. Consonant-only (C-only) words/sentences preserved all consonants while replacing the vowels with noise. Vowel-only (V-only) words/sentences preserved the all vowels and replaced consonant segments with noise. In both cases, the type of preserved consonant or vowel acoustic cues was dependent upon the listening condition of each experimental group.

C. Procedures

All participants were tested alone in a sound-attenuating booth. Stimuli were presented using Tucker-Davis Technologies System III hardware and passed through a headphone buffer (HB-7) to an ER-3 A insert earphone. Presentation levels were calibrated by presenting the speech-shaped noises matching the CVC and sentence RMS at a sound level of 70 dB SPL using an HA-2 2-cc coupler and a Larson Davis model 2800 sound level meter with linear weighting. Speech materials were presented monaurally to the right ear.

For sentence testing, participants were instructed to repeat aloud each sentence as accurately as possible. Digital recordings of responses were made for offline analysis. Sentences in V-only, C-only, and Full-utterance conditions were presented fully randomized to the listeners. No sentence was repeated for a given listener. For CVC testing, participants typed what they thought they heard on a PC running a MATLAB open-set response interface. All words were presented to the participants in a random order. Full-utterance CVCs (i.e., without segmental replacement) were presented in a second block to listeners and were a second presentation of the words tested under segmental replacement. Familiarization trials were provided before sentence and word testing making use of stimuli not used during testing. No feedback was provided during familiarization or testing.

D. Scoring

Sentence responses were scored offline by two trained raters. All words were scored. Words were required to be repeated exactly correct (e.g., no missing or additional suffixes). Inter-rater agreement was previously established at 98%. Typed word responses were automatically corrected for phonetic misspellings, were visually inspected, and were automatically scored using custom-made software. All word percent-correct scores were transformed to rationalized arcsine units to stabilize the error variance prior to analysis (RAU; Studebaker, 1985).

E. Results

Independent-samples t-tests on the Full-utterance condition between the $FlatF_0$ (sentences: $M = 100$ RAU, $SD = 11$; words: $M = 85$ RAU, $SD = 7$) and NoF_0 (sentences: $M = 103$ RAU, $SD = 6$; words: $M = 90$ RAU, $SD = 1$) groups indicated no significant difference in overall performance between groups for either the words or sentences ($p > 0.05$). Therefore, both flattening and removing periodic variations of the speech source appear to have the same degrading effect on speech perception of the target.

Overall, performance across the vowel and consonant conditions was poor. A 2 (segment) by 2 (context) by 2

(listener group) mixed model analysis of variance (ANOVA) was conducted. Results demonstrated significant main effects for segment [$F(1,12) = 80.9$, $p < 0.001$] and context [$F(1,12) = 5.1$, $p < 0.05$]. Interactions with the type of segment occurred with context and listener group ($p < 0.01$). Contrasts were conducted to investigate these effects and are described here.

1. Sentences

Sentence results for the two groups are displayed in Fig. 2(a) along with data for natural sentences from Fogerty and Kewley-Port (2009). Listeners previously performed at 99% (116 RAU) for these same sentences presented in a Full-utterance condition, unprocessed and without segmentation in quiet (Kewley-Port *et al.*, 2007). Paired *t*-tests were performed to examine the differences between segments for the two listener groups. This analysis demonstrated significantly better performance for V-only sentences than for C-only sentences for FlatF_0 [$t(6) = 6.8$, $p < 0.001$] and NoF_0 groups [$t(6) = 17.2$, $p < 0.001$]. No differences were observed between FlatF_0 and NoF_0 groups ($p > 0.05$).

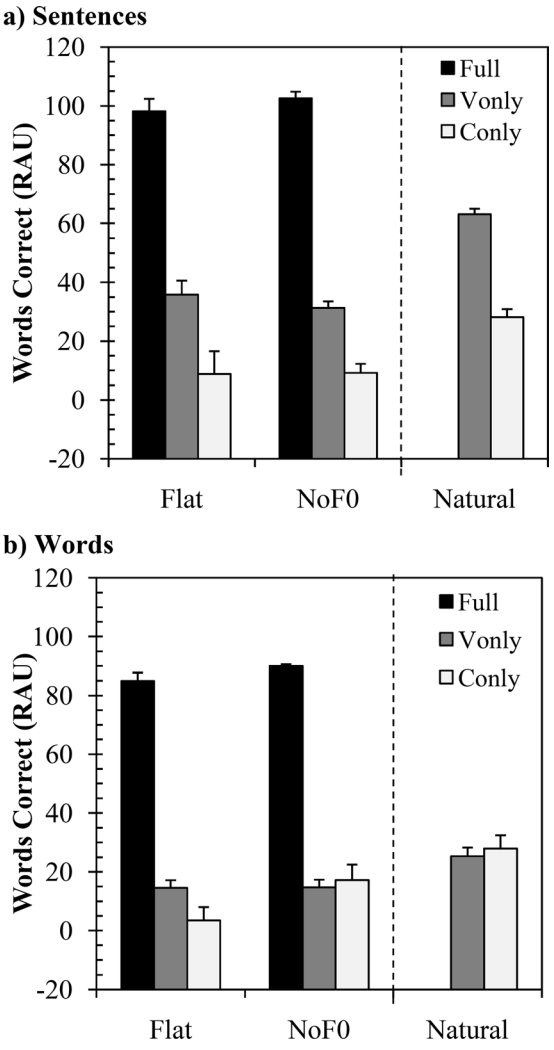

FIG. 2. Word recognition scores in RAU for FlatF_0 and NoF_0 listener groups who received limited acoustic cues in resynthesized speech. Results for (a) sentences and (b) isolated CVC word contexts are displayed. Results of unprocessed, natural speech during segmental replacement for these same sentences (Fogerty and Kewley-Port, 2009) and words (Fogerty and Humes, 2010) are also displayed on the right for reference. Error bars = standard error of the mean.

2. Words

Word results for the two groups are displayed in Fig. 2(b) along with data for natural words from Fogerty and Humes (2010). As confirmed through informal listening, extrapolation of the performance-intensity functions from Dirks *et al.* (2001) indicate that performance for these words in the full-utterance condition, unprocessed and presented in quiet, is near 100% accuracy. FlatF_0 listeners demonstrated significantly better performance for V-only words than for C-only words [$t(6) = 4.8$, $p < 0.01$], while NoF_0 listeners demonstrated no significant difference between these segmental conditions ($p > 0.05$). No differences were observed for V-only and C-only words between FlatF_0 and NoF_0 groups ($p > 0.05$). These performance patterns held for both lexically easy and hard words, with two exceptions. First, no difference was observed between V-only and C-only conditions for easy words during FlatF_0 processing [$t(6) = 2.2$, $p = 0.07$]. Second, for lexically hard C-only words, listeners in the NoF_0 group performed better than FlatF_0 listeners ($p < 0.01$). These two findings combined suggest that the only difference in performance between the FlatF_0 and NoF_0 groups was restricted to the C-only lexically hard words, with worse performance in the FlatF_0 group. Thus, static F_0 cues may actually provide misleading information when the vowel is not available, a problem which is compounded when co-occurring with less phonologically distinct words, leading to reduced performance.

Comparisons across lexical context were also investigated. Across both groups, performance was better for V-only sentences than V-only words [FlatF_0: $t(6) = -3.9$, $p < 0.01$; NoF_0: $t(6) = -7.2$, $p < 0.01$]. However, no significant difference between C-only words and C-only sentences was observed. This pattern was previously observed for natural words and sentences (Fogerty and Humes, 2010). Thus, removing dynamic F_0 cues did not alter the relative vowel and consonant contributions to overall intelligibility.

F. Discussion

The relative contribution of dynamic F_0 cues present during consonant or vowel segments was examined. Flattening or removing the F_0 contour did lower overall intelligibility compared to natural sentences. However, it did so for both C-only and V-only speech. The relative contribution of consonants and vowels to overall speech intelligibility did not change when these dynamic cues of the prosodic contour were removed. That is, V-only sentences still resulted in better intelligibility than C-only sentences [Fig. 2(a)]. For word contexts [see Fig. 2(b)], static F_0 information appears to interfere with consonant contributions to overall word recognition. Removing all F_0 cues, thus creating aperiodic speech, restored consonant contributions to be equal to those of the vowel, as observed for natural speech. This pattern of results suggests that other acoustic information, not conveyed by dynamic F_0 cues, is responsible for the vowel advantage

observed for sentences. Therefore, experiment 2 investigated dynamic amplitude and frequency cues as additional sources of speech information that may differentially contribute to vowel contributions and facilitate global perception of the entire speech sample.

III. EXPERIMENT 2: THE ROLE OF E AND TFS IN VOWEL AND CONSONANT CONTRIBUTIONS

Rosen (1992) divided the modulation rates of speech into three timescales: envelope, periodicity, and temporal fine structure. Experiment 1 examined the contributions of periodicity, conveyed by dynamic F_0 information, to relative segmental contributions. Experiment 2 investigated the remaining two timescales: E and TFS. Of note, E and TFS, as processed here using the Hilbert transform, both include periodicity information (see Faulkner *et al.*, 2000; Moore, 2004; Plack and Oxenham, 2005). The purpose of experiment 2 was to compare the importance of amplitude (E) and frequency (TFS) modulation cues to the relative contribution of vowels and consonants. Previous work has examined the local phonetic features conveyed by E and TFS to consonant and vowel identification (e.g., Xu *et al.*, 2005). The current study extends this investigation to how E and TFS, present during consonants and vowels, contribute globally to word and sentence intelligibility.

A. Listeners

Fourteen normal-hearing listeners (M = 21 yr, 19–23 yr) were paid to participate in this study. All listeners were native speakers of American English and had pure-tone thresholds no greater than 20 dB HL at octave intervals from 250 to 8000 Hz (ANSI, 2004). Listeners were again randomly assigned to one of two experimental groups. No listeners previously participated in experiment 1.

B. Methods

Experiment 2 followed the design and procedures of experiment 1. Two groups of listeners, randomly assigned to either the predominant envelope (E) or predominant temporal fine structure (TFS) listening group, completed testing for sentences and words in full-utterance, V-only, and C-only segmental conditions. Word and sentence materials, segmentation and noise replacement, and test procedures remained identical to experiment 1.

Shown in Fig. 3 are the two temporal conditions for the E and TFS groups. Processing of these materials modeled a new method introduced by Fogerty (2011) for varying the availability of E and TFS without altering the underlying processing of temporal or spectral properties. This was done by independently extracting E and TFS components from noisy speech sources at different SNRs. The noise added to the speech sample matched the power spectrum for that individual sample. For E processing, the E component was extracted using the Hilbert transform over three analysis bands that represent equal cochlear distance (frequency range = 80–6400 Hz) from a speech sample presented at 11 dB SNR, while the TFS component was similarly extracted from the same speech sample presented at −5 dB SNR. For TFS processing, the reverse was true, with TFS extracted at 11 dB SNR and E extracted at −5 dB SNR. Thus, this method preserved both E and TFS cues, but varied

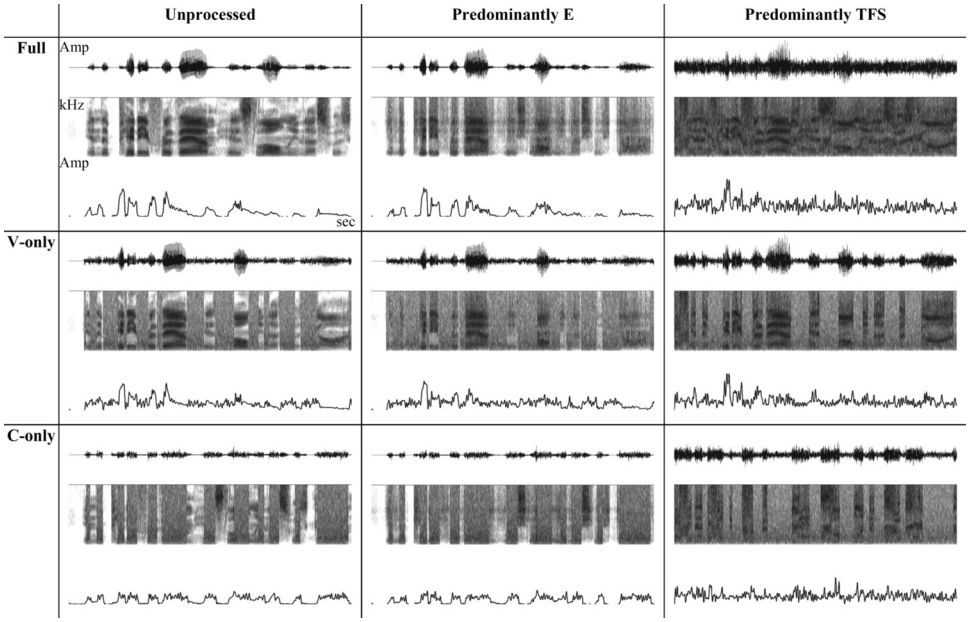

FIG. 3. Stimulus processing examples for experiment 2. The original unprocessed sentence, "In the pity for them his loneliness was gone," is displayed in the left column under full, V-only, and C-only conditions for reference. Stimulus examples for the E and TFS listener groups are displayed in the center and right columns respectively. Note that E and TFS processing still contained both E and TFS components with the non- target component extracted at a less favorable SNR (see text). Each cell displays the amplitude waveform, spectrogram (0–6 kHz), and amplitude envelope using half- wave rectification and low-pass filtering for that condition and listener group. Note that the TFS stimuli did contain an amplitude envelope, although it was highly uncorrelated with the unprocessed stimulus due to the noisy signal extraction. Periodicity cues were provided by both E and TFS processing. Amplitude waveforms were normalized prior to segmentation.

the availability of both types of information by using different noise levels. Fogerty (2011) previously demonstrated that recognition performance for sentences processed according to this method was similar for both E and TFS conditions.

C. Results

An independent-samples t-test between E and TFS groups was first completed on the full-utterance samples. Only data from four listeners in each group were available for this full-utterance condition (due to an error in the stimulus files for three listeners each from the E and TFS groups). Data from all listeners were available for the segmental conditions. Results demonstrated no significant differences between the full-utterance E (sentence: M = 87 RAU, SD = 8; word: M = 50 RAU, SD = 5) and TFS (sentence: M = 82 RAU, SD = 6; word: M = 55 RAU, SD = 6) conditions for either sentences or words ($p > 0.05$), indicating that the processing method for these conditions resulted in similar performance levels for speech containing predominantly E or TFS cues. This is in agreement with Fogerty (2011) using the same processing method for sentence materials.

A 2 (segment) by 2 (context) by 2 (listener group) mixed model ANOVA was conducted. Results demonstrated significant main effects for segment [$F(1,12) = 415.1$, $p < 0.001$] and context [$F(1,12) = 132.4$, $p < 0.001$]. Interactions with the type of segment occurred with context and listener group ($p < 0.001$). Contrasts were conducted to investigate these effects and are described here.

1. Sentences

Sentence results for the E and TFS groups are displayed in Fig. 4(a), again with comparison data for natural sentences from Fogerty and Kewley-Port (2009). Paired t-tests demonstrated significantly better performance for V-only sentences than for C-only sentences for both groups [E: $t(6) = 9.9$, $p < 0.001$; TFS: $t(6) = 19.2$, $p < 0.001$]. Comparison between E and TFS groups demonstrated significantly better performance for TFS listeners in the V-only condition [$t(12) = 5.6$, $p < 0.001$] and for E listeners in the C-only condition [$t(12) = 9.2$, $p < 0.001$].

2. Words

Word results for E and TFS groups are displayed in Fig. 4(b) along with data for natural words from Fogerty and Humes (2010). TFS listeners demonstrated significantly better performance for V-only words than for C-only words [$t(6) = 14.5$, $p < 0.01$), while E listeners demonstrated no significant difference between these segmental conditions ($p > 0.05$). Comparison between E and TFS groups demonstrated significantly better performance for TFS listeners in the V-only condition [$t(12) = 3.8$, $p = 0.003$] and for E listeners in the C-only condition [$t(12) = 8.0$, $p < 0.001$]. These performance patterns held for both lexically easy and hard words.

Comparisons across contexts were also investigated. Across both groups, performance was better for V-only sentences than words [E: $t(6) = -8.4$, $p < 0.001$; TFS: $t(6) = -12.8$, $p < 0.001$]. However, no difference between C-only words and sentences was observed ($p > 0.05$).

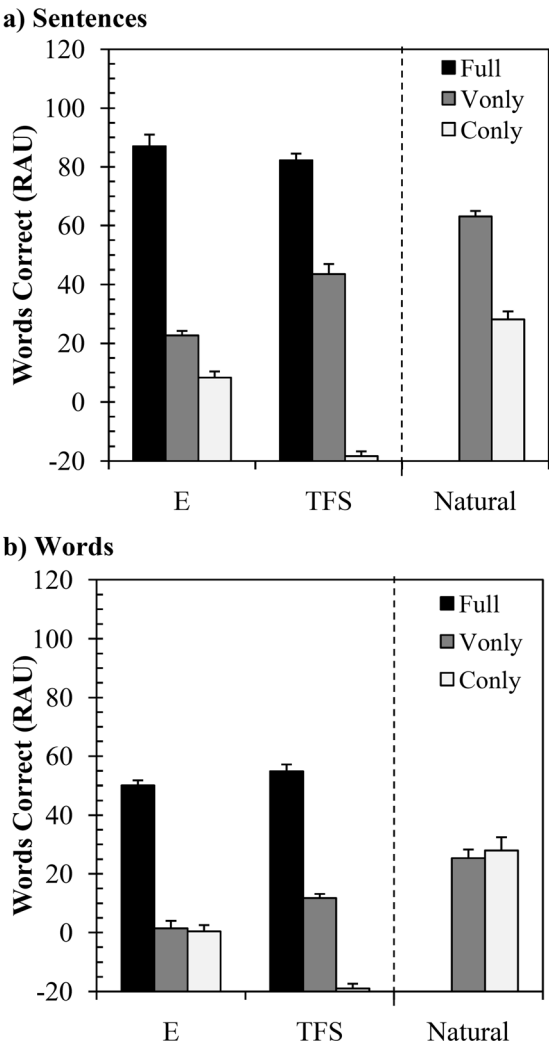

FIG. 4. Word recognition scores in RAU for listener groups who predominantly received E or TFS acoustic cues in resynthesized speech. Results for (a) sentences and (b) isolated CVC word contexts are displayed. Results of unprocessed, natural speech during segmental replacement for these same sentences (Fogerty and Kewley-Port, 2009) and words (Fogerty and Humes, 2010) are also displayed on the right for reference. Error bars = standard error of the mean.

D. Discussion

The relative contribution of amplitude (i.e., E) and frequency (i.e., TFS) cues present during consonant and vowel segments was investigated. Speech was processed to selectively mask E or TFS components, resulting in speech containing predominant TFS or E cues respectively. Results demonstrated that only the vowel provides usable TFS cues. This was true for both sentences and words. In contrast, both consonants and vowels provide E cues. However, the E present during vowels contributes relatively more speech information for the intelligibility of sentences than does the E present during consonants. Equal contribution of the E in consonants and vowels during word context was observed. Thus, the amplitude envelope, at least in part, is responsible

for the context-dependent advantage of vowels. As the TFS of vowels also contributed more during sentences than words, this advantage of the E may also be enhanced by vowel TFS cues.

IV. GENERAL DISCUSSION

The current study was designed to investigate how general acoustic properties might explain differences between the contribution of consonants and vowels that are seen only in sentences, not isolated words. Therefore, this investigation was not about functional differences between consonants and vowels, although some have argued for such dissociations (e.g., Nespor *et al.*, 2003). It was also not about phoneme identification. Instead, the current study was an investigation of the contributions of different acoustic properties to the overall intelligibility of words and sentences, and whether these acoustic properties are conveyed during consonant or vowel segments. Contributions of the dynamic fundamental frequency contour, amplitude envelope, and temporal fine structure were investigated as potential sources of global speech information that may underlie the observed superiority of vowel information in sentences.

Fundamental frequency information was investigated as it is likely to convey global speech information about the entire utterance, rather than being heavily involved in individual phoneme identification. Dynamic fundamental frequency (or pitch) information provides supra-linguistic cues for syntactic units (Lehiste, 1970), facilitates prediction of future syntactic structures (Wingfield *et al.*, 1984), and aids in sentence recognition (Laures and Wiesmer, 1999; Watson and Schlauch, 2008). Furthermore, fundamental frequency information is important for source segregation (e.g., Bregman *et al.*, 1990; Darwin *et al.*, 2003), which may be important as this study essentially employed alternating speech (either V or C segments) and noise stimuli.

However, while the F_0 of consonants and vowels appears to significantly contribute to speech recognition, as observed in this study and in previous work (Laures and Wiesmer, 1999; Watson and Schlauch, 2008), the focus of this study was on whether F_0 cues were differentially important for the contribution of vowel or consonant units. In particular, the focus was on whether F_0 or the F_0 contour explains the context-dependent contributions of vowels. Results demonstrated that even though dynamic F_0 cues were removed for the FlatF_0 and NoF_0 group conditions, listeners still obtained significantly higher word-recognition scores for V-only sentences as compared to C-only sentences. Therefore, F_0 cues alone do not explain this vowel advantage, as removing these cues did not remove this advantage (although it was slightly reduced for the NoF_0 group). Furthermore, no significant differences were obtained between FlatF_0 and NoF_0 groups, indicating that the presence of steady-state mean F_0 information did not facilitate speech intelligibility over aperiodic speech. This was true for both the full-utterance and segmented speech materials.

As for the temporal properties investigated, E and TFS cues also convey important dynamic information that may also contribute to the perception of global, supra-linguistic speech information distributed across the sentence. Amplitude modulations of the E convey local cues about manner and voicing (Rosen, 1992); however, the majority of speech energy occurs at around a 4-Hz modulation which corresponds to the syllabic-rate of English. Therefore, E temporal cues may facilitate syllabification (Rosen, 1992) and also facilitate sentence-level predictions (Waibel, 1987). TFS on the other hand, could be particularly important for sentences presented using the segment replacement method that results in interrupted speech. TFS cues were investigated as these appear to facilitate the perception of speech "glimpses" (see Cooke, 2003) between intervening noise maskers (Lorenzi *et al.*, 2006).

Preserved E and TFS cues do provide information important to the contribution of vowels to word recognition. Consonant contributions to speech intelligibility are not conveyed by TFS cues, as evidenced by the floor performance in that condition for both word and sentence contexts. E cues, however, provide important information that is present in both consonants and vowels. In addition, the E appears to provide additional information during vowels in sentences above what is contained in the consonants, although this additional benefit is not seen in isolated word contexts. Thus, E contributions parallel the context-dependent pattern of vowel contributions that is observed with natural, unmodified speech presentations. This finding highlights the important contributions of the amplitude envelope during sentence presentations, possibly by providing higher-level structural acoustic cues that facilitate top-down contextual processing. Furthermore, as this additional contribution occurred during vowel units, amplitude modulations in the mid-frequencies of speech that have the most speech energy are likely to carry this informative cue. This parallels recent findings that suggest listeners perceptually weight envelope cues in the mid-frequencies higher relative to other frequency regions and relative to TFS contributions (Fogerty, 2011).

The greater importance of E cues in sentences has been highlighted previously by studies of amplitude compression. Compression has the effect of attenuating intensity fluctuations of the modulation transfer function (Plomp, 1988); thereby, reducing E speech cues. Studies have most commonly investigated amplitude compression during phoneme recognition tasks in nonsense syllables. In general, phoneme recognition remains relatively intact during compression (see for example, Dreschler, 1989). In agreement with this finding, for speech containing predominantly E cues, Van Tasell and Trine (1996) found that consonant recognition in nonsense syllables was not impacted by single-band compression until the most severe compression condition. However, sentence recognition was clearly reduced (see also, Jenstad and Souza, 2007; Souza and Kitch, 2001). In addition, in line with the current study, Van Tasell and Trine concluded that listeners rely on the amplitude envelope of sentences, not periodicity cues. These previous findings, in combination with the current study, suggest that compression may not be overly detrimental for phoneme recognition because of the more limited role of amplitude modulations in isolated contexts, such as CVC words or nonsense syllables. In contrast, amplitude modulations within the vowel

segments carry additional information in sentences. Degrading those cues through single-band compression significantly degrades performance, possibly in part due to the reduction of amplitude cues that provide global information about the entire sentence.

The design of more advanced signal-processing technology for speech applications requires clear definitions of the roles played by different auditory speech properties. For maximal speech transmission, signal processing must focus on preserving and enhancing the particular auditory speech cues that are most informative in a given auditory environment. The results of this study demonstrate that at the segmental level, temporal envelope cues distributed across vowel segments are essential for sentence-level speech intelligibility. Further work is required to delineate whether these essential cues are conveyed by intrinsic amplitude modulations within the vowel, or perhaps more likely, result from modulations in amplitude between neighboring segments and/or across the entire sentence. Identification of these properties will inform how best to present temporal envelope cues via hearing aids and cochlear implants to achieve maximal sentence intelligibility.

V. SUMMARY AND CONCLUSIONS

This study investigated the contribution of fundamental frequency, envelope, and temporal fine structure cues to the perceptual contributions of vowels and consonants in words and sentences. Overall, results demonstrated that while dynamic F_0 cues contribute to overall intelligibility, they do not explain perceptual differences between vowel and consonant contributions. Instead, investigation of the temporal fine structure demonstrated that its contribution to intelligibility is conveyed almost exclusively by the vowels in both words and sentences. Envelope cues, on the other hand, were present during both vowels and consonants. Furthermore, results with utterances containing primarily E cues paralleled that of natural speech findings from other studies (Fogerty and Humes, 2010; Fogerty and Kewley-Port, 2009) in that a greater contribution of vowels as compared to consonants was seen for sentences, but not for isolated words. This study provides further evidence for the importance of E cues in the mid- frequency region that contain the vowel formants (see also Fogerty, 2011) and that the informativeness of these cues may be essential in more ecological contexts, such as meaningful sentences, as compared to isolated word contexts.

ACKNOWLEDGMENTS

This work was supported, in part, by NIA R01 AG008293 awarded to L.E.H.

ANSI (**2004**). ANSI S3.6-2004, *American National Standard Specification 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.

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.

Carreiras, M., and Price, C. (**2008**). "Brain activation for consonants and vowels," Cerebral Cortex **18**, 1727–1735.

Carreiras, M., Gillon-Dowens, M., Vergara, M., and Perea, M. (**2009**). "Are vowels and consonants processed differently? Event-related potential evidence with a delayed letter paradigm," J. Cogn. Neurosci. **21**, 275–288.

Cole, R., Yan, Y., Mak, B., Fanty, M., and Bailey, T. (**1996**). "The contribution of consonants versus vowels to word recognition in fluent speech," *Proceedings of the ICASSP'96*, pp. 853–856.

Coleman, J. (**2003**). "Discovering the acoustic correlated or phonological contrasts," J. Phonetics **31**, 351–372.

Cooke, M. (**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.

Dirks, D. D., Takayanagi, S., and Moshfegh, A. (**2001**). "Effects of lexical factors on word recognition among normal-hearing and hearing-impaired listeners," J. Am. Acad. Audiol. **12**, 233–244.

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.

Dreschler, W. A. (**1989**). "Phoneme perception via hearing aids with and without compression and the role of temporal resolution," Audiol. **28**, 49–60.

Ernestus, M., Baayen, H., and Schreuder, R. (**2002**). "The recognition of reduced word forms," Brain Lang. **81**, 162–173.

Faulkner, A., Rosen, S., and Smith, C. (**2000**). "Effects of the salience of pitch and periodicity information on the intelligibility of four-channel vocoded speech: Implications for cochlear implants," J. Acoust. Soc. Am. **108**, 1877–1887.

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 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.

Fujimura, O., and Lindqvist, J. (**1971**). "Sweep-tone measurements of vocal-tract characteristics," J. Acoust. Soc. Am. **49**, 541–558.

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.

Gay, T. (**1978**). "Effect of speaking rate on vowel formant movements," J. Acoust. Soc. Am. **63**, 223–230.

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, 2803–2806.

Hawkins, S. (**2003**). "Roles and representations of systematic fine phonetic detail in speech understanding," J Phonetics **31**, 373–405.

Hillenbrand, J. M., and Nearey, T. M. (**1999**). "Identification of resynthesized/hVd/utterances: Effects of formant contour," J. Acoust. Soc. Am. **105**, 3509–3523.

Janse, E. (**2004**). "Word perception in fast speech: Artificially time-compressed vs. naturally produced speech," Speech Commun. **42**, 155–173.

Janse, E., and Ernestus, M. (**2011**). "The roles of bottom-up and top-down information in the recognition of reduced speech: Evidence from listeners with normal and impaired hearing," J. Phonetics **39**, 330–343.

Jenstad, L. M., and Souza, P. E. (**2007**). "Temporal envelope changes of compression and speech rate: Combined effects on recognition for older adults," J. Speech Lang. Hear. Res. **50**, 1123–1138.

Kawahara, H., Masuda-Kastuse, 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.

Kent, R. D., and Minifie, F. D. (**1977**). "Coarticulation in recent speech production models," J. Phonetics **5**, 115–117.
Kewley-Port, D., Burkle, Z., and Lee, J. (**2007**). "Contribution of consonant versus vowel information to sentence intelligibility for young normal-hearing and elderly hearing-impaired listeners," J. Acoustic. Soc. Am. **122**, 2365–2375.
Ladefoged, P. (**2001**). *Vowels and Consonants: An Introduction to the Sounds of Languages* (Blackwell, Oxford), pp. 1–191.
Ladefoged, P., and Broadbent, D. E. (**1957**). "Information conveyed by vowels," J. Acoust. Soc. Am. **29**, 98–104.
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.
Lehiste, I. (**1970**). *Suprasegmentals* (MIT Press, Cambridge, MA), pp. 1–194.
Leiberman, P. (**1963**). "Some effects of semantic and grammatical context on the production and perception of speech," Lang. Speech **6**, 172–187.
Liberman, A. M., Cooper, F. S., Shankweiler, D. P., and Studdert-Kennedy, M. (**1967**). "Perception of the speech code," Psychol. Rev. **74**, 431–461.
Lindblom, B. (**1963**). "Spectrographic study of vowel reduction," J. Acoust. Soc. Am. **35**, 1773–1781.
Local, J. (**2003**). "Variable domains and variance relevance: Interpreting phonetic exponents," J Phonetics **31**, 321–339.
Lorenzi, C., Gilbert, G., Carn, 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. **103**, 18866–18869.
Luce, P. A., and Pisoni, D. B. (**1998**). "Recognizing spoken words: The Neighborhood Activation Model," Ear Hear. **19**, 1–36.
Max, L., and Caruso, A. J. (**1997**). "Acoustic measures of temporal intervals across speaking rates: Variability of syllable- and phrase-level relative timing," J. Speech Hear. Res. **40**, 1097–1110.
Miller, M. A., Heise, G. A., and Lichten, W. (**1951**). "The intelligibility of speech as a function of the context of the test materials," J. Exp. Psych. **41**, 329–335.
Moore, B. C. J. (**2004**). *An Introduction to the Psychology of Hearing*, 5th Ed. (Elsevier, London), pp. 1–200.
Nearey, T. M., and Assmann, P. (**1986**). "Modeling the role of vowel inherent spectral change in vowel identification," J. Acoust. Soc. Am. **80**, 1297–1308.
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.
Nespor, M., Peña, M., and Mehler, J. (**2003**). "On the different roles of vowels and consonants in speech processing and language acquisition," Lingue Linguaggio **2**, 201–227.
New, B., Araújo, V., and Nazzi, T. (**2008**). "Differential processing of vowels and consonants in lexical access through reading," Psych. Sci. **19**, 1223–1227.
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.
Pickett, J. M., and Pollack, I. (**1963**). "Intelligibility of excerpts from fluent speech: Effects of rate of utterance and duration of excerpt," Lang. Speech **3**, 151–164.
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.
Plomp, R. (**1988**). "The negative effect of amplitude compression in multichannel hearing aids in the light of the modulation-transfer function," J. Acoust. Soc. Am. **83**, 2322–2327.
Port, R. F. (**1981**). "Linguistic timing factors in combination," J. Acoust. Soc. Am. **69**, 262–274.
Port, R. F. (**2003**). "Meter and speech," J. Phonetics **31**, 599–611.
Ramus, F., Nespor, M., and Mehler, J. (**1999**). "Correlates of linguistic rhythm in the speech signal," Cognition **75**, AD3–AD30.
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., Kamath, V., Wygonski, J., and Ekelid, M. (**1995**). "Speech recognition with primarily temporal cues," Science **270**, 303–304.
Shattuck-Hufnagel, S., and Turk, A. E. (**1996**). "A prosody tutorial for investigators of auditory sentence processing," J. Psycholing. Res. **25**, 193–247.
Smith, Z. M., Delgutte, A. J., and Oxenham, A. J. (**2002**). "Chimaeric sounds reveal dichotomies in auditory perception," Nature **416**, 87–90.
Souza, P. E., and Kitch, V. (**2001**). "The contribution of amplitude envelope cues to sentence identification in young and aged listeners," Ear Hear. **22**, 112–119.
Steeneken, H. J. M., and Houtgast, T. (**1999**). "Mutual dependence of the octave-band weights in predicting speech intelligibility," Speech Commun. **28**, 109–123.
Stevens, K. N. (**2002**). "Toward a model for lexical access based on acoustic landmarks and distinctive features," J. Acoust. Soc. Am. **111**, 1872–1891.
Stilp, C. E., and Kluender, K. R. (**2010**). "Cochlea-scaled spectral entropy, not consonants, vowels, or time, best predicts speech intelligibility," Proc. Natl. Acad. Sci. **107**, 12387–12392.
Strange, W., Jenkins, J. J., and Johnson, T. L. (**1983**). "Dynamic specification of coarticulated vowels," J. Acoust. Soc. Am. **74**, 695–705.
Studebaker, G. (**1985**). "A rationalized arcsine transform," J. Speech Hear. Res. **28**, 455–462.
Tajima, K., and Port, R. F. (**2003**). "Speech rhythm in English and Japanese," in *Phonetic Interpretation: Papers in Laboratory Phonology VI*, edited by J. Local, R. Ogden, and R. Temple (Cambridge University Press, Cambridge, UK), pp. 317–334.
Takayanagi, S., Dirks, D., and Moshfegh, A. (**2002**). "Lexical and talker effects on word recognition among native and non-native listeners with normal and impaired hearing," J. Am. Acad. Aud. **16**, 494–504.
Toro, J. M., Nespor, M., Mehler, J., and Bonatti, L. L. (**2008**). "Finding words and rules in a speech stream: Functional differences between vowels and consonants," Psychol. Sci. **19**, 137–144.
Van Tasell, D. J., and Trine, T. D. (**1996**). "Effects of single-band syllabic amplitude compression on temporal speech information in nonsense syllables and in sentences," J. Speech Hear. Res. **39**, 912–922.
Waibel, A. (**1987**). "Prosodic knowledge sources for word hypothesization in a continuous speech recognition system" *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'87*, pp. 856–859.
Watson, P. J., and Schlauch, R. S. (**2008**). "The effect of fundamental frequency on the intelligibility of speech with flattened intonation contours," Am J Speech Lang Pathol. **17**, 348–355.
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.
Wingfield, A., Lombardi, L., and Sokol, S. (**1984**). "Prosodic features and the intelligibility of accelerated speech: Syntactic versus periodic segmentation," J. Speech Hear. Res. **27**, 128–134.
Xu, L., Thompson, C. S., and Pfingst, B. E. (**2005**). "Relative contributions of spectral and temporal cues for phoneme recognition," J. Acoust. Soc. Am. **117**, 3255–3267.
Zue, V. W., and Seneff, S. (**1988**). "Transcription and alignment of the TIMIT database," *Proceedings of the Second Meeting on Advanced Man-Machine Interface through Spoken Language*, 11.1–11.10.