The effects of fundamental frequency contour manipulations on speech intelligibility in background noise[a)]

Sharon E. Miller,[b)] Robert S. Schlauch, and Peter J. Watson
Department of Speech-Language-Hearing Sciences, University of Minnesota, Minneapolis, Minnesota 55455

(Received 29 January 2009; revised 21 January 2010; accepted 14 March 2010)

Previous studies have documented that speech with flattened or inverted fundamental frequency (F0) contours is less intelligible than speech with natural variations in F0. The purpose of this present study was to further investigate how F0 manipulations affect speech intelligibility in background noise. Speech recognition in noise was measured for sentences having the following F0 contours: unmodified, flattened at the median, natural but exaggerated, inverted, and sinusoidally frequency modulated at rates of 2.5 and 5.0 Hz, rates shown to make vowels more perceptually salient in background noise. Five talkers produced 180 stimulus sentences, with 30 unique sentences per F0 contour condition. Flattening or exaggerating the F0 contour reduced key word recognition performance by 13% relative to the naturally produced speech. Inverting or sinusoidally frequency modulating the F0 contour reduced performance by 23% relative to typically produced speech. These results support the notion that linguistically incorrect or misleading cues have a greater deleterious effect on speech understanding than linguistically neutral cues.

© *2010 Acoustical Society of America.* [DOI: 10.1121/1.3397384]

PACS number(s): 43.71.Gv, 43.71.Es [KWG] Pages: 435–443

I. INTRODUCTION

Prosody is an inherent feature in spoken language and is realized by the physical and perceptual suprasegmental features present in speech such as pitch, stress, and duration, among others (Lehiste, 1970; Cutler *et al.*, 1997; Pierrehumbert, 1999). In spoken English, the prosodic cue of fundamental frequency (F0), the acoustic correlate of perceived pitch, communicates information regarding speaker intention (Grant and Walden, 1996) and provides crucial linguistic cues for speech perception (Cutler *et al.*, 1997). Past work suggests dynamic changes in F0 cue the location of important content words in an utterance, possibly leading to priority processing and activation of these stored words in the mental lexicon (Cutler *et al.*, 1997). F0 modulation is also thought to help listeners demarcate word and syllable boundaries in connected speech, as words are rarely processed in isolation (Lehiste, 1970; Cutler *et al.*, 1997; Liss *et al.*, 2000; Spitzer *et al.*, 2007).

Prior studies have quantified the importance of a dynamic F0 to speech intelligibility by assessing word recognition in background noise (Wingfield *et al.*, 1984; Laures and Weismer, 1999; Binns and Culling, 2007). Laures and Weismer (1999) found that sentences with a flat F0 contour presented in background noise, resulting in a perceptual monotone pitch, significantly decreased the percentage of correctly identified words compared to speech having a natural, variable F0 contour. Listeners in their study also consistently rated speech with the flat F0 as less intelligible than typically produced speech (Laures and Weismer, 1999). The authors offered multiple explanations of their findings. They theorized that their data support the Cutler and Foss (1977) work suggesting F0 provides crucial cues for speech understanding. A flattened F0 contour could potentially eliminate the linguistic cues associated with a dynamic F0 that serve to direct a listener's attention to important content words in a sentence, decreasing speech understanding. The authors also postulated that a flattened F0 could reduce the contrasts between syllables, making it more difficult for a listener to parse the speech stimuli, degrading speech intelligibility. Finally, Laures and Weismer also hypothesized that a flattened F0 with consistent harmonic spacing could decrease the probability of harmonics falling near vocal tract resonant frequencies, diminishing intelligibility. However, a recent study by Watson and Schlauch (2008) concluded that although harmonic density affects speech intelligibility, it cannot account for the reduced intelligibility observed in studies when the F0 is flattened at the median value.[1]

Binns and Culling (2007) have also assessed the role intonation contours play in speech intelligibility. The authors measured speech recognition thresholds (SRTs) in interfering speech and in speech-shaped noise for low-predictability sentences with manipulated F0 contours. The results indicated that inverting the F0 contour, thereby providing incorrect dynamic F0 cues, significantly decreased intelligibility by the greatest amount relative to an unmodified condition. Significant differences between speech having a flattened F0 and speech with a normal or inverted F0 contour were only observed when the interfering speech also had a variety of F0 contours. The study also documented that having only 50% of the variation of the original F0 contour is sufficient for accurate speech understanding and reducing the contour to 25% of its original variation resulted in performance similar

[a)] Portions of this work were presented in "Examining explanations for the fundamental frequency's contribution to speech intelligibility in noise" J. Acoust. Soc. Am. 118 (1933).

[b)] Author to whom correspondence should be addressed. Electronic mail: mill1979@umn.edu to a monotone speech condition. In addition, low-pass filtering the envelope of the F0 contour proved that envelope frequencies between 2 and 4 Hz in naturally produced speech are critical to speech intelligibility because removing these frequencies from the contour had a significant deleterious effect on SRTs. However, speech intelligibility did not significantly change when the envelope of the F0 contour was unfiltered compared to sentences having an F0 contour low-pass filtered at 4 Hz. The authors concluded that low frequency envelope modulation frequencies in "natural"[2] speech are important for speech intelligibility.

Inverting the F0 contour degrades performance, but it is unknown which envelope modulation frequencies interfere with recognition or whether flattened F0 contours produce identical performance to inverted F0 contours. Earlier studies, employing inverted F0 contours, involved all modulation frequencies (Hillenbrand, 2003; Culling *et al.*, 2003). It is not known if incorrectly presented low frequency envelope modulation frequencies interfere with recognition performance. Indeed, there is evidence from psychoacoustic experiments that frequency modulation (FM) makes vowels more perceptually salient in sustained, speech-shaped noise (McAdams, 1984, 1989). It is possible that a lower-level auditory process related to auditory streaming, not a higher-order linguistic cue such as F0 information identifying content words as described previously, could play a role in the diminished intelligibility of speech having a flat F0 contour.

Psychoacoustic research has documented that the auditory system can take advantage of acoustic regularities in complex stimuli such as harmonicity (Culling and Darwin, 1993), timbre (van Noorden, 1977), level and timing differences (Bregman, 1990; McAdams and Drake, 2002), amplitude modulation, and FM (McAdams, 1989; Bregman, 1990; Culling and Darwin, 1993; Yost *et al.*, 1993; Culling and Summerfield, 1995; Carlyon, 2004) to segregate competing inputs into distinct auditory streams, a process termed auditory scene analysis or auditory streaming (Bregman, 1990). Thus, it can also be said that acoustic properties of sound can facilitate speech understanding when there are multiple, competing messages or background noise by grouping the parts of individual sound sources together (Carlyon, 2004). We define this type of sound processing as a lower-level processing, meaning it occurs prior to or separate from linguistic knowledge processing. Monotone speech is devoid of any F0 FM cues which could hamper a listener's ability to segregate it from competing sources, such as background noise. Past studies investigating the role of FM by McAdams (1984, 1989) and Culling and Summerfield (1995) provided evidence that frequency modulating the F0 of speech and nonspeech sounds against background maskers increased the perceptual salience of the target sounds and improved identification thresholds of the stimuli. Along this same vein, if adding sinusoidal FM to the F0 contour of sentences improves speech understanding, then it can be concluded that nonlinguistic acoustic cues can improve speech intelligibility in difficult listening environments.

Speech devoid of F0 movement is less intelligible than typically intonated speech. However, the previous findings that inverting the F0 contour results in a drop in intelligibility suggests that more than just variation in the F0 contour is necessary to yield optimal performance in background noise; the inverted F0 contours employed by Binns and Culling (2007), which contained all of the modulation frequencies of the original sentences, represent F0 cues that are contrary to those of typical speech. Cues that match those of typical speech would be important if "linguistically correct" cues are required for tasks such as segmenting speech in sentences. This present study expands on this previous work and aims to investigate whether other acoustic F0 modulations can improve speech understanding in noise relative to speech with a flattened F0 contour. In one set of conditions, the natural F0 contour of spoken sentences was replaced by ones with sinusoidal variation at either 2.5 or 5 Hz. These low modulation frequencies correspond to ones in the range of those found effective by McAdams (1984, 1989) for making vowels more salient in speech-shaped noise. Previous amplitude modulated (AM) studies have also documented that the temporal rate of spoken words and syllables is roughly 2.5 and 5 Hz, respectively (Duquesnoy and Plomp, 1980). Frequency and intensity cues reliably covary in both speech perception and production (Neuhoff *et al.*, 1999), but access to AM intensity cues in the temporal envelope that demarcate words and syllables is degraded in background noise. Providing FM cues at the rates of 2.5 and 5.0 Hz could aid speech understanding in noise when intensity cues are diminished. Furthermore, Binns and Culling (2007) found natural modulation frequencies in the temporal envelope for F0 below 4 Hz to be important for speech understanding; the 2.5 Hz modulation frequency is within that range and the 5 Hz one is above it. The influence on speech understanding of these linguistically implausible sinusoidally modulated F0 contours is not known.

By measuring intelligibility when speech has either a linguistically typical or atypical F0 contour or an F0 contour composed of acoustic cues known to make speech targets more salient in background noise, we can document the relative contributions of linguistic and nonlinguistic acoustic cues to the decreased intelligibility of speech having a flattened F0 contour. To achieve this aim, speech intelligibility for low-predictability sentences in noise was measured for speech having the following F0 contours: unmodified/control, flattened at the median, inverted, sinusoidally frequency modulated at 2.5 and 5.0 Hz, and natural with exaggerated F0

The motivations for choosing the listed F0 contour conditions were based on past speech intelligibility in noise studies. To replicate the Laures and Weismer (1999) and Binns and Culling (2007) findings, the F0 contour was flattened at its average value. The inverted F0 condition was included to assess how a dynamic, but linguistically incorrect, F0 contour affects speech intelligibility compared to the Binns and Culling (2007), Culling *et al.* (2003), and Hillenbrand (2003) values. To investigate whether a low-level auditory cue could increase the perceptual salience of the speech in background noise and improve intelligibility relative to the flat F0 condition, the F0 contour was sinusoidally frequency modulated at two rates. If one of the FM conditions improves performance relative to the flattened F0 condition, then it can be concluded that a nonlinguistic acoustic cue can contribute favorably to speech intelligibility in background noise. Past experiments that inverted the F0 contour most likely provided inaccurate sentence- and word-level F0 stress cues. If frequency modulating the F0 improves performance relative to the inverted F0 condition in the present study, then it can be concluded that linguistically neutral, or nonlinguistic acoustic, F0 modulation cues can also advance speech understanding relative to linguistically incorrect cues. Finally, because a 50% reduction in the variation of the F0 contour did not affect speech intelligibility (Binns and Culling, 2007), an exaggerated F0 condition was included to assess how greater than expected F0 variation affects performance. Past work has demonstrated that persons with hearing impairment require greater F0 variation in speech to follow intonation patterns compared to persons with normal hearing (Grant, 1987). We were interested in how persons with normal hearing performed with an exaggerated F0 contour. Given that an exaggerated F0 is used in infant-directed speech and is thought to help infants parse the speech stream (Kuhl *et al.*, 1997), we would expect an exaggerated F0 to improve performance relative to the flat F0 condition.

II. METHODS

A. Participants

Fifteen paid listeners, age 18–30, participated in the task. All subjects were native speakers of American-English, had normal hearing sensitivity, and denied any history of speech-language or neurological disorder.

B. Stimuli

The speech stimuli were recorded from five females who were native speakers of American English and who self-reported normal hearing and no history of a speech-language disorder. Two of the speakers had previously digitally recorded the sentences (22.1 kHz) for another study and three additional female speakers recorded the sentences for the present study. 180 sentences, each sentence containing five key words, were randomly chosen from the IEEE speech corpus (Rothauser *et al.*, 1969) and each speaker was assigned 36 different sentences from the selected list. For this study, the speakers wore a head-mounted, high fidelity microphone (AKG, C420, range 20–20 000 Hz, frequency response −6 to +2 dB) and the sentences were digitally recorded to disk (44.1 kHz) using a CD recorder (Marantz, CDR 300). Recorded sentences were down-sampled to 22.1 kHz to match the previously recorded sentences. Each stimulus sentence was equated for its root-mean-square (rms) level.

The following six conditions were randomly assigned to the 180 sentences: unmodified/control, flattened F0 at the median, exaggerated by a factor of 1.75, F0 frequency modulated at 2.5 Hz, F0 frequency modulated at 5.0 Hz, and inverted F0. Prior to any F0 manipulation, voiced segments were first extracted from the audio files using a 0.0001 s time step and any audible instances of glottal fry were removed. In total, six sentences per speaker in each of the six conditions were used, yielding 30 unique sentences per F0 manipulation condition.

All F0 manipulations of the stimuli were performed using PRAAT software and algorithms (Boersma and Weenink, 1999). For the control condition, no manipulations were performed on the F0 contour.

To flatten the F0 contour, the median F0 value of the sentence was computed and this value was inserted at each point voicing occurred. To exaggerate the F0 contour by a factor of 1.75, the following formula was used:

$$\mathrm{F0}' = 1.75(\mathrm{F0} - \mathrm{F0}_{\mathrm{med}}) + \mathrm{F0}_{\mathrm{med}}. \qquad (1)$$

F0′ represents the new F0 of the frame, F0 is the F0 of the control at a given time sample, and $\mathrm{F0}_{\mathrm{med}}$ the median F0 of the nonmanipulated sentence. The factor of 1.75 was chosen for the exaggerated condition because independent listeners rated it as the highest multiplication factor that still sounded “natural.” Exaggerating the F0 by a factor of 1.75 is also consistent with previous infant-directed speech studies that document adults increase their average F0 range by factors varying anywhere from 1.5 (Grieser and Kuhl, 1988) to 1.68 (Garnica, 1977) when addressing infants compared to adults. In the FM conditions, the F0 contour was flattened and then modulated at 2.5 Hz and 5.0 Hz using the following equation:

$$\mathrm{F0}' = [\mathrm{F0}_{\mathrm{med}} + (\sigma \sin(2\pi X m_f + \theta))], \qquad (2)$$

where σ represents the average standard deviation of the F0 across sentences for each speaker, m_f the modulation frequency (2.5 or 5.0 Hz), and Θ the phase angle. The starting phase of each sentence $(0-2\pi)$ was randomly selected. The motivation for choosing the FM rates of 2.5 and 5.0 Hz is twofold. First, they are consistent with other FM values used in previous speech and nonspeech studies (Culling and Darwin, 1993; Culling and Summerfield, 1995; Carlyon *et al.*, 2000). Second, Binns and Culling (2007) established that F0 contour modulations between 2 and 4 Hz, close to the syllable rate of speech, were the most important cues in speech understanding in background noise. Thus, we chose to frequency modulate the F0 contour at 2.5 Hz. In the same study, Binns and Culling (2007) also established that F0 contours modulated at frequencies greater than 4.0 Hz did not significantly contribute to speech intelligibility. A modulation rate of 5.0 Hz, thus, represents a frequency in the envelope of the F0 contour that is higher than those found to contribute to speech understanding in natural speech.

The F0 contour was inverted using the same formula as Culling *et al.* (2003),

$$\mathrm{F0}' = \frac{\mathrm{F0}_{\mathrm{med}}^2}{\mathrm{F0}}. \qquad (3)$$

The manipulated F0 contours for all sentences were inserted back into each stimulus sentence using PRAAT PSOLA algorithms. To give a sense of how the F0 contours differed, Fig. 1 displays the six F0 contour conditions when applied to the same sentence.

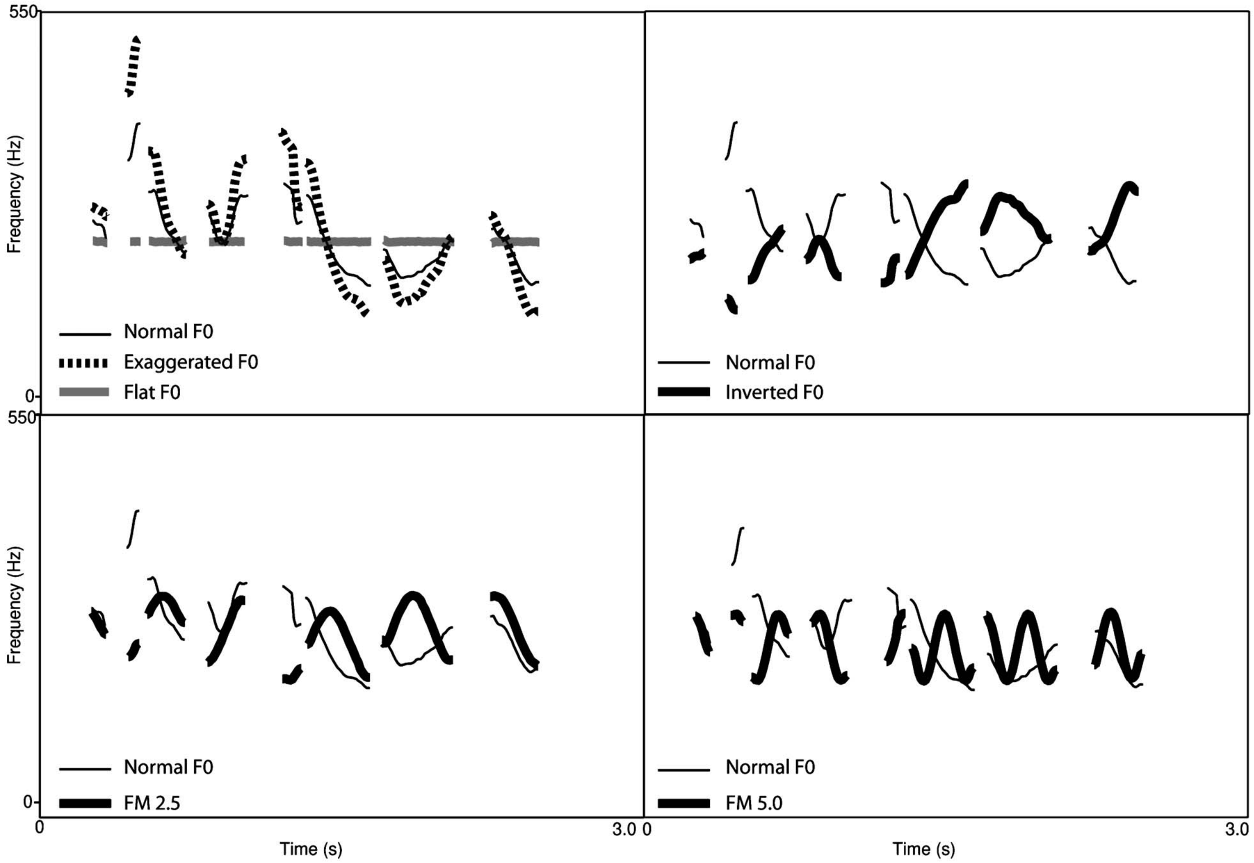

FIG. 1. Illustrates the six F0 contour conditions for the sentence "The cat chased the dog down the street."

The speech-shaped noise used in the study was generated based on the spectrum of the unmodified sentences. The long term spectrum of the concatenated unmodified sentences was extracted in 1/3 octave bands using a spectrum analyzer (Hewlett Packard, model 3561A). Using a computer-based signal processing program, white noise was filtered to take on the form of the speech-shaped spectrum of the sentences. Based on pilot data, the filtered noise was equalized across sentences to produce signal-to-noise ratios (SNRs) of −3 dB sound pressure level (SPL) for three speakers and −2 dB SPL for the other two speakers. The SNR for each talker was selected to eliminate possible ceiling and floor effects in the experiment and to equate keyword intelligibility across talkers.

The arrays of speech and noise were added together and presented to listeners via a single channel. The masker began 150 ms before the onset of the sentence and ended 150 ms after the offset of the sentence. The stimuli were presented through a 16 bit custom designed digital-to-analog converter (22.05 kHz). All stimuli were passed through an antialiasing filter with a cutoff frequency of 10 000 Hz before being routed to an earphone (Sennheiser, HD 250, linear II).

C. Procedure

Listeners were seated in a double-walled sound-attenuated booth and listened to the stimuli monaurally through a left earphone. At the beginning of each session, listeners heard six practice low-predictability sentences not included in the experiment, each sentence representing one of the six experimental conditions. To familiarize the subjects with the stimuli, each practice sentence was presented twice, one time each with and without the speech-shaped noise. After the listener demonstrated familiarity with the task, the experiment began. The 180 sentences were presented one time to each listener in a random order. Upon hearing the sentence, each subject orthographically recorded and also repeated aloud the perceived sentence. An examiner seated in the booth recorded the key words spoken by the listener and these were verified by comparing the experimenter's recordings to the listener's written responses. To begin, the listener pressed a button to hear the first sentence presentation. After recording a response, the listener pressed a button to start the presentation of the next sentence.

To demonstrate that the processed sentences were highly intelligible when presented in isolation (without background noise), three naive listeners were recruited. These listeners were first presented with practice sentences as were the experimental subjects. The results demonstrated ceiling effects consistent with expected performance for naturally produced sentences in quiet. The unmodified/control condition yielded 99% performance, two conditions yielded 98% performance (flattened and FM2.5), one yielded 97% performance (exaggerated), and the remaining conditions yielded 96% performance (inverted and FM 5.0).

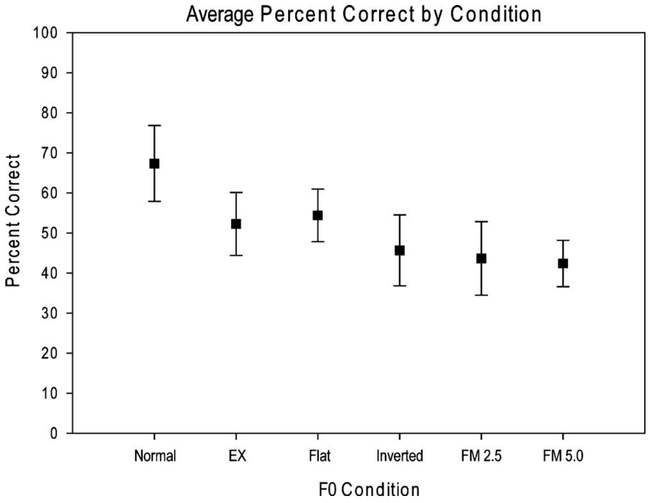

FIG. 2. Mean keyword recognition performance displayed as percent correct of the 15 subjects across the six experimental conditions. The bars represent the standard deviation of the individual data.

III. RESULTS

A. Data analysis

Figure 2 displays the average percent of key words correctly identified in each of the six conditions and demonstrates that any manipulation of the F0 contour degraded speech intelligibility in background noise. A repeated measures analysis of variance (ANOVA) was performed with condition as the within-subject factor and was found to be significant $[F(5,70)=52.97,\ p<0.001]$. The individual subject data for each of the six F0 conditions can be seen in Fig. 3.

Based on visual inspection of the data, an analysis was performed on two contrasts within the significant main effect for F0 condition. First, the unmodified/control condition was compared to the exaggerated and flattened conditions $[F(1,14)=54.52,\ p<0.0001]$. Second, the performance in the exaggerated and flattened F0 conditions was compared to the inverted F0, FM 2.5, and FM 5.0 conditions $[F(1,14)=80.79,\ p<0.001]$. Both contrasts were significant at a Bonferroni-corrected 0.01 level. These results suggest that performance on the speech understanding task can be grouped into three distinct clusters according to condition. As seen in Fig. 2, flattening or exaggerating the F0 similarly degraded performance relative to the normal condition. Likewise, frequency modulating or inverting the F0 contour diminished performance relative to the flattened and exaggerated F0 conditions by a similar amount.

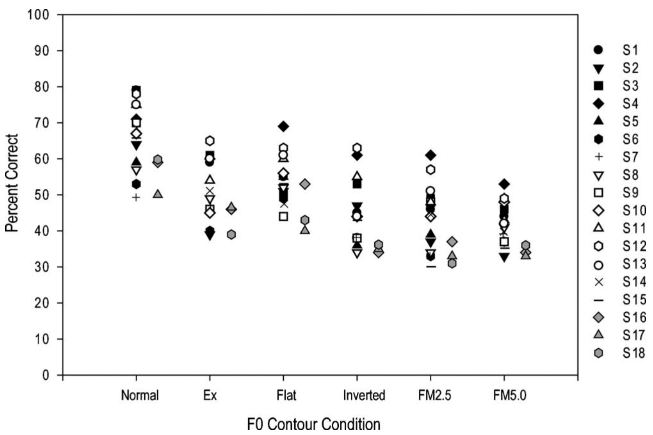

FIG. 3. Individual keyword recognition performance of the 15 listeners displayed as percent correct across the 6 different experimental conditions. The three gray data points shifted to the right in each condition represent the performance for the three subjects that completed the experiment with the processed control stimuli.

TABLE I. Displays the difference, in average percent correct, from the control condition across the six F0 contour conditions for the study using a processed or unprocessed control condition.

F0 contour condition	Unprocessed control stimuli (% difference) ($n=15$)	Processed control stimuli (% difference) ($n=3$)
Control	0	0
Flat	15.1	10.7
Exaggerated	12.9	12.2
Inverted	21.7	21.1
FM2.5	23.7	22.2
FM5.0	24.9	21.8

The stimuli in this experiment were created using synthesized F0 manipulations, meaning the sentences in all conditions, except for the unmodified control condition, were processed. To ensure the validity of the results, it was necessary to prove that the F0 manipulations, and not the processing itself, accounted for the observed intelligibility. To address this issue, three additional participants completed the identical speech understanding experiment with processed control stimuli. Their data are shown in Fig. 3, along with the data from the subjects who listened to the unprocessed, original control condition sentences. Data from the unprocessed and processed control condition sentences show the identical pattern across conditions and the magnitude of the drop in performance to each of the conditions is about the same, as shown in Table I.[3]

B. Keyword analysis

Each stimulus sentence contained five keywords numbered 1–5 (the numbers 1–5 represents the keyword's temporal position in the sentence). rms amplitude was equalized across all sentences in the six F0 conditions; however, Fig. 4 demonstrates that the average percent correct for the five keyword positions was not equal. Because of this trend, a repeated measures ANOVA with F0 manipulation and keyword position as within subjects factors was performed and F0 manipulation $[F(5,70)=27.71,\ p<0.001]$, keyword position $[F(4,56)=4.464,\ p<0.01]$, and the interaction $[F(20,280)=4.505,\ p<0.0001]$ were found to be significant. Figure 4 also documents that keywords at the end of a sentence appear to be less intelligible than earlier keywords in the normal condition, and this appears to be the opposite trend for the inverted F0 condition. Because of this observation, percent correct was then regressed on keyword position for the inverted F0 condition which produced a significant slope coefficient of $1.86(t=2.05,\ p<0.05)$. Figure 4 displays the regression line fit for the inverted F0 data. Percent correct was also regressed on keyword position in the

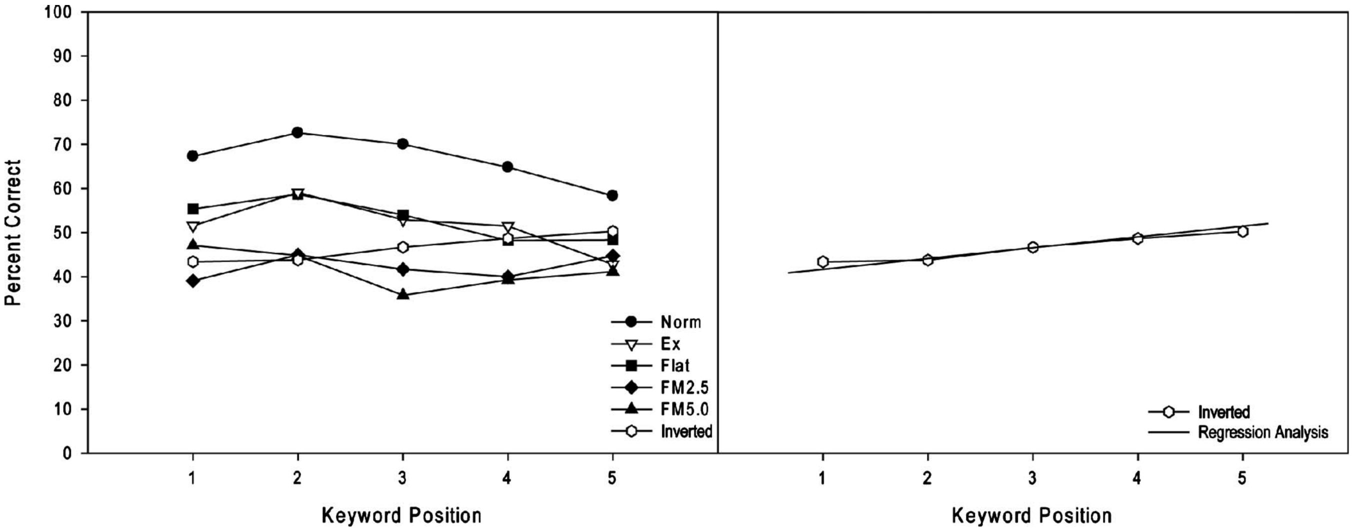

FIG. 4. Average percent correct of keyword recognition performance for keyword positions 1–5 across the six experimental conditions. The regression analysis line is displayed for the inverted F0 data.

unmodified/control condition which also produced a significant slope coefficient of $-2.57(t=-2.758,\ p=.0078)$. These regression analyses reveal that percent correct word understanding improved at the end of the inverted F0 sentences; the opposite trend occurred for the unmodified/control sentences.

IV. DISCUSSION

A. Possible explanations for the results

Consistent with past research, the results of this study clearly document that any deviation from a typically intonated F0 contour pattern has a deleterious effect on speech understanding in background noise (Laures and Weismer, 1999; Binns and Culling, 2007). However, the F0 manipulations in this study did not affect each of the six experimental conditions equally, with the results of the manipulated conditions falling into two distinct groups. Exaggerating or flattening the F0 contour reduced average speech intelligibility relative to typical speech by approximately 13%. The frequency modulated and inverted F0 contour conditions further decreased mean speech intelligibility performance relative to typical speech by 23%. These distinct clusters of results could possibly be explained due the presence of either linguistically neutral or incorrect cues.

A major aim of this study was to investigate whether sinusoidal FM would cause the speech to become more salient than the background noise, thereby improving speech understanding relative to speech with a flat F0 contour. The results demonstrate that, unlike past psychoacoustic experiments, FM of the F0 contour did not cause the speech to become more salient or improve speech understanding. Instead, frequency modulating the F0 contour caused speech understanding to plummet. This finding is surprising given that the FM conditions produced relatively small deviations in F0 from a sentence's median value at a slow rate of change. Table II shows that the standard deviation of the excursions in the FM conditions is smaller than that of the normal condition.

One possible explanation for the poor speech understanding in the FM conditions involves unassociated lexical segmentation. Mattys *et al.* (2005) support a hybrid model of word segmentation. Under optimal listening conditions, they argue that a lexically based segmentation scheme is low cost and efficient. However, under impoverished listening conditions, listeners likely base segmentation decisions on a probabilistic analysis of multiple, sublexical based cues, including prosodic information that coincides with word boundaries. Several studies have demonstrated the important role of F0 for conveying stress as a cue for accurate speech segmentation under adverse listening conditions (Spitzer *et al.*, 2007, 2009). Accordingly, frequency modulating the F0 contour in our experiment most likely destroyed the normal stress pattern of the words, and if listeners rely on changes in F0 to segment speech in noise, linguistically incorrect stress patterns could prevent listeners from establishing accurate word boundaries within the sentence, leading to poor speech understanding. Inaccurate F0 cues were also present in the inverted F0 condition, and it is likely that improper stress patterns interfered with speech understanding in the same manner.

Content words in sentences are often produced with a greater F0 value and at a higher intensity (Lehiste, 1970). In the current study, the presence of background noise most likely prevented listeners from using intensity cues in order

TABLE II. Selected acoustic characteristics averaged for five speakers across the different F0 contour conditions after F0 contour manipulations.

F0 contour condition	Average median F0 value of sentences after manipulation (Hz)	Average σ of F0 of sentences after manipulation (Hz)
Control	201	31
FM 2.5	204	24
FM 5.0	202	22
Inverted	205	31
Flat	202	⋯
Exaggerated	203	54

to attend to content words, making the F0 cues more important. In the flat F0 condition, the words all had the same F0 and lacked this important stress-based segmentation cue; however, the stress cues based on F0 were simply neutralized across the individual words in the sentences as opposed to being inaccurate or misleading. This could possibly explain why frequency modulating or inverting the F0 contour produced significantly poorer speech understanding than the flat F0 condition. Although it is plausible that a person could learn to use F0 cues to segment speech with the inverted contour because it varies in a predictable manner from the original sentences, Neuhoff *et al.* (1999) found that dynamically changing intensity and frequency cues interact to affect loudness and pitch judgments in a manner consistent with a single mechanism. They note that models of pitch and intensity based on static sounds are unable to account for perceptual interactions that appear to take advantage of the finding that their physical correlates, frequency and intensity, covary in many naturally produced sounds. This notion that the F0 contour is important in degraded speech understanding is also supported by the finding that, in the present study, listeners performed equally well across all six conditions when the sentences were presented in quiet. When the stimuli were presented in noise, performance dropped. Likewise, Hillenbrand (2003) also concluded that when speech is degraded, listeners seem to rely more heavily on the F0 contour.

In this study, surprisingly, exaggerating the F0 contour also significantly degraded performance relative to typically intonated speech. Binns and Culling (2007) previously demonstrated that decreasing the excursions in the F0 contour by 50% did not change performance, but a decrease in the variation of F0 to 25% of its original values did. Thus, natural F0 could be reduced considerably without it affecting intelligibility. Increasing F0 movement by exaggerating the F0 contour was predicted to improve intelligibility, or at least sustain it at the same level as for typically intonated speech because the exaggerated F0 contour maintained a typical intonation pattern. However, that was not the case in the present study. This unexpected result may be better understood in the light of recent findings by Watson and Schlauch (2008) who also measured performance for key words in sentences presented in a noise background. Watson and Schlauch (2008) investigated the effect of F0 height on speech intelligibility for speech having F0 contours flattened at low, median, and high F0 values. The authors noted that speech flattened at a higher F0 (near the top of speaker's range) was 7% less intelligible relative to a mean-F0 flattened condition. They hypothesized, based on earlier work (Diehl, *et al.*, 1996; de Cheveigné and Kawahara, 1999), that undersampling of the resonant frequencies of the vocal tract at high F0 values caused this decrease in speech intelligibility. The effects of extreme low F0 values on the intelligibility of sentences are unknown, but it is also possible that these values significantly affected word recognition in this condition.

The present finding that exaggerating the F0 has a deleterious effect on speech understanding is consistent with recent work examining infant-directed speech (IDS). IDS is characterized as having a heightened pitch and broad variations in F0, and these features are thought to help infants parse the speech stream and establish vowel categories (Kuhl *et al.*, 1997). However, Trainor and Desjardins (2002) investigated the independent effects of pitch height and pitch contour on infant vowel discrimination, and the authors found that while broad variations in pitch helped infants discriminate vowels, a steady heightened pitch actually hindered an infant's vowel discrimination.

The exaggerated F0 contour condition degraded performance in the present study with normal hearing listeners. However, this finding does not mean that exaggerated pitch contours are not beneficial to speech perception. Grant (1987) previously documented that when identifying the location of the stressed peak in an intonation contour, persons with hearing loss required far greater frequency transitions (a factor of 1.6–6) than those with normal hearing. Grant concluded that persons with profound hearing loss may have trouble following natural F0 variations in spoken language, and thus, some sort of signal processing in hearing aids that incorporates exaggerated F0 contours could be beneficial. In the present study, it is possible that exaggerating the F0 contour by a factor of 1.75 interfered with the fine structure of the sentences for persons with normal hearing. Persons with sensorineural hearing loss, particularly the ones with profound losses as in Grant, 1987, have poor frequency resolution and hearing loss in the regions of formant frequencies. Exaggerating the F0 contour, for these persons with significant hearing loss, may remain a viable option to improving their speech understanding. Further studies are required to determine the maximum amount of exaggeration in the variability of the F0 contour that does not reduce identification in persons with hearing within normal limits.

The mismatch between unnatural F0 values and formant peaks described above to account for the poorer performance for the exaggerated condition is a possible but unlikely explanation for the drop in performance for the FM conditions. In the FM conditions, the F0 excursions were relatively small, smaller than the F0 variability in typically intonated speech (see Table II and Fig. 1). If a mismatch between the F0 and formant values were responsible for the observed results, it would be expected that the exaggerated F0 condition, with its larger deviations in F0 values, would produce poorer speech understanding performance than the FM conditions. This, however, was not the case.

Several studies of vowel perception and F0 suggest that F0 plays a minor role in the intelligibility of synthetic vowels or vowels in single, isolated words when the values of F0 are selected well within the range of a typical speaker's productions (Katz and Assman, 2001). Whalen *et al.* (1999) argue that the wide range of F0 used by typical talkers does not result in an adjustment of formant frequencies to accommodate these differences and that lexical stress and tonal contrasts have no effect on vowel identity. When differences are found, they tend to be for extremely high values of F0; low and moderate values produce comparable performance (Ryalls and Lieberman, 1982). A dynamically varying F0 tends to *improve* performance over static conditions, but this effect is either small (Hillenbrand and Gayvert, 1993; Bunton, 2006) or is observed mainly for the highest F0s where undersampling of the vocal tract could play a role (Diehl *et al.*,1996). Given what is known about the effect of F0 on vowel identification, the larger drop for the FM conditions compared to the flattened condition is unexpected.

B. Comparisons with past studies

Any manipulation of the fundamental frequency contour in the present study decreased speech intelligibility in background noise relative to typically intonated speech. This finding is consistent with previous studies (Laures and Weismer, 1999; Binns and Culling, 2007); however, the current results vary in some regards to past work. Although different methodologies were employed, making direct comparisons difficult, it is important to discuss the results of this study in the context of previous work. In this study, flattening the F0 contour significantly degraded performance relative to normally intonated speech and inverting the F0 further significantly degraded speech intelligibility relative to monotone speech in the presence of speech-shaped background noise. In contrast, although Binns and Culling (2007) observed that monotone speech degraded speech intelligibility relative to typically intonated speech, this difference did not reach significance when using a speech-shaped background noise masker. However, they did find a significant difference in speech intelligibility between the two conditions when the masker was interfering speech having a variety of F0 contours (e.g., flattened F0, normal F0, and inverse F0). Our results demonstrate that it is possible to observe a dramatic difference in intelligibility between speech having a monotone pitch and speech having an inverted F0 contour in the presence of just speech-shaped noise. This difference between studies could potentially be accounted for by differences in cognitive processing demands. The current study utilized five female talkers, whereas, Binns and Culling (2007) used the same male speaker throughout the majority of their experiment. It is possible that the listeners in their experiment learned characteristics of the male talker, and this familiarity may have provided additional pitch tracking cues that aided speech understanding performance in the speech-shaped noise. Perhaps, then, only in the presence of multitalker babble were cognitive processing demands equivalent across studies. For example, the five female talkers in our study were intermixed across all stimulus sentence conditions, preventing listeners from learning the characteristics of any one speaking voice. This increased uncertainty may be why we observed intelligibility for monotone speech to be significantly poorer in the presence of speech-shaped background noise and Binns and Culling (2007) only observed this difference when multitalker babble was used.

In the current study, adding sinusoidal FM at 5 Hz, close to the syllable rate of speech, caused very poor speech understanding in noise. Binns and Culling (2007) low-pass filtered the F0 contour and found that contour frequencies above 4 Hz did not contribute to intelligibility in naturally intonated sentences. Thus, F0 contours having frequencies greater than 4 Hz would not be predicted to significantly contribute to speech understanding in the presence of other frequencies. However, superimposing an unnatural frequency modulated contour at 5 Hz caused a significant deleterious effect on performance in the current study.

Prior studies have examined the effect of F0 movement on vowel perception, and while this study utilized sentences, it is useful to examine how different types of F0 variability affect vowel identification. Diehl *et al.* (1996) previously examined how F0 movement affected vowel perception and found that the vowels were more intelligible when they had a falling F0 than when the vowels had a static F0. A falling F0 is more typical of natural vowel production, and Diehl *et al.*'s (1996) finding concurs with the results of the current study in that deviations from typical intonation patterns degrades speech understanding. Like Diehl *et al.* (1996), Bunton (2006) also examined the effect of different types of F0 movement on vowel perception in single words, but it is difficult to compare our findings with this study because vowel studies do not provide listeners with an opportunity for segmentation errors.

V. CONCLUSIONS

The present study supports earlier findings that any unnatural F0 contour manipulation decreased speech understanding in background noise (Laures and Weismer, 1999; Binns and Culling, 2007). Relative to typically intonated speech, flattening or exaggerating the F0 contour decreased speech intelligibility by 13% and frequency modulating or inverting the F0 contour decreased intelligibility by 23%. This study makes it clear that incorrect or misleading linguistic cues related to intonation have a more deleterious effect on speech understanding than speech comprised of plausible linguistic cues. It is important to note, however, that while not observed in this study, some artificial F0 manipulations can improve speech intelligibility (Watson and Schlauch, 2009). For example, persons who use an electrolarynx can apply a simple rising or falling intonation contour to speech segments that improve intelligibility relative to typical monotone speech produced with the device.

ACKNOWLEDGMENTS

The authors wish to acknowledge Edward Carney for his assistance with data analysis, Peggy Nelson and Ben Munson for their helpful comments, and Lisa LaMora for her help with data collection.

[1] Undersampling of the spectral envelope can occur for high static and, perhaps, midvalues of F0. The spectral envelope for these conditions is represented more accurately by sweeping the F0 contour (Diehl *et al.*, 1996). However, for low values of F0, where the separation of harmonics of the F0 is smaller than the width of formants and/or the ear's analysis bandwidth, undersampling should not be a concern. This does not rule out that dynamic F0 cues contribute to perception based on some other mechanism.

[2] In order to create an uninterrupted F0 contour in this condition, Binns and Culling (2007) used only continuously voiced speech segments.

[3] This finding is in line with past studies by Watson and Schlauch (2008) and Laures and Weismer (1999) who also found no differences in intelligibility between processed and unprocessed control stimuli.

Binns, C., and Culling, J. F. (**2007**). "The role of fundamental frequency contours in the perception of speech against interfering speech," J. Acoust.

Soc. Am. **122**, 1765–1776.

Boersma, P., and Weenink, D. (**1999**). "Pratt: A system for doing phonetics by computer," Technical Report No. 132, Institute of Phonetic Sciences, University of Amsterdam, Amsterdam, The Netherlands.

Bregman, A. S. (**1990**). *Auditory Scene Analysis: The Perceptual Organization of Sound* (MIT, Cambridge, MA).

Bunton, K. (**2006**). "Fundamental frequency as a perceptual cue for vowel identification in speakers with Parkinson's disease," Folia Phoniatr Logop **58**, 323–339.

Carlyon, R. P. (**2004**). "How the brain separates sounds," Trends Cogn. Sci. **8**, 465–471.

Carlyon, R. P., Moore, B. C., and Micheyl, C. (**2000**). "The effect of modulation rate on the detection of frequency modulation and mistuning of complex tones," J. Acoust. Soc. Am. **108**, 304–315.

Culling, J. F., and Darwin, C. J. (**1993**). "Perceptual separation of simultaneous vowels: Within and across-formant grouping by F0," J. Acoust. Soc. Am. **93**, 3454–3467.

Culling, J. F., Hodder, K. I., and Toh, C. Y. (**2003**). "Effects of reverberation on perceptual segregation of competing voices," J. Acoust. Soc. Am. **114**, 2871–2876.

Culling, J. F., and Summerfield, Q. (**1995**). "The role of frequency modulation in the perceptual segregation of concurrent vowels," J. Acoust. Soc. Am. **98**, 837–846.

Cutler, A., Dahan, D., and van Donselaar, W. (**1997**). "Prosody in the comprehension of spoken language: A literature review," Lang Speech **40**, 141–201.

Cutler, A., and Foss, D. J. (**1977**). "On the role of sentence stress in sentence processing," Lang Speech **20**, 1–10.

de Cheveigné, A., and Kawahara, H. (**1999**). "Missing-data model of vowel identification," J. Acoust. Soc. Am. **105**, 3497–3508.

Diehl, R. L., Lindblom, B., Hoemeke, K. A., and Fahey, R. P. (**1996**). "On explaining certain male-female differences in the phonetic realization of vowel categories," J. Phonetics **24**, 187–208.

Duquesnoy, A. J., and Plomp, R. (**1980**). "Effect of reverberation and noise on the intelligibility of sentences in cases of presbycusis," J. Acoust. Soc. Am. **68**, 537–544.

Garnica, O. (**1977**). "Some prosodic and paralinguistic features of speech to young children," in *Talking to Children: Language Input and Acquisition*, edited by C. E. Snow and C. A. Ferguson, (Cambridge University Press, Cambridge), pp. 63–88.

Grant, K. W. (**1987**). "Identification of intonation contours by normally hearing and profoundly hearing-impaired listeners," J. Acoust. Soc. Am. **82**, 1172–1178.

Grant, K. W., and Walden, B. E. (**1996**). "Spectral distribution of prosodic information," J. Speech Hear. Res. **39**, 228–238.

Grieser, D. L., and Kuhl, P. K. (**1988**). "Maternal speech to infants in a tonal language: Support for universal prosodic features in motherese," Dev. Psychol. **24**, 14–20.

Hillenbrand, J. (**2003**). "Some effects of intonation contour on sentence intelligibility," J. Acoust. Soc. Am. **114**, 2338.

Hillenbrand, J., and Gayvert, R. T. (**1993**). "Vowel classification based on fundamental frequency and formant frequencies," J. Speech Hear. Res. **36**, 694–700.

Katz, W. F., and Assman, P. F. (**2001**). "Identification of children's and adults' vowels: Intrinsic fundamental frequency, fundamental frequency dynamics, and presence of voicing," J. Phonetics **29**, 23–51.

Kuhl, P. K., Andruski, J. E., Chistovich, I. A., Chistovich, L. A., Koshevnikova, E. V., Ryskina, V. L., Stolyarova, E. I., Sundberg, E. I., and Lacerda, F. (**1997**). "Cross-language analysis of phonetic units in language addressed to infants," Science **277**, 684–686.

Laures, J. S., and Weismer, G. (**1999**). "The effects of a flattened fundamental frequency on intelligibility at the sentence level," J. Speech Lang. Hear. Res. **42**, 1148–1156.

Lehiste, I. (**1970**). *Suprasegmentals* (MIT, Cambridge, MA).

Liss, J. M., Spitzer, S. M., Caviness, J. N., Adler, C., and Edwards, B. W. (**2000**). "Lexical boundary error analysis in hypokinetic and ataxic dysarthria," J. Acoust. Soc. Am. **107**, 3415–3424.

Mattys, S. L., White, L., and Melhorn, J. F. (**2005**). "Integration of multiple speech segmentation cues: A hierarchical framework," J. Exp. Psychol. Gen. **134**, 477–500.

McAdams, S. (**1984**). "Spectral fusion, spectral parsing, and the formation of auditory images," Ph.D. thesis, Stanford University, Stanford, CA.

McAdams, S. (**1989**). "Segregation of concurrent sounds. I: Effects of frequency modulation coherence," J. Acoust. Soc. Am. **86**, 2148–2159.

McAdams, S., and Drake, C. (**2002**). "Auditory perception and cognition," in *Stevens' Handbook of Experimental Psychology: Sensation and Perception*, S. Yantis and H. Pashler (Wiley, New York), pp. 397–452.

Neuhoff, J. G., McBeath, M. K., and Wanzie, W. C. (**1999**). "Dynamic frequency change influences loudness perception: A central, analytic process," J. Exp. Psychol. Hum. Percept. Perform. **25**, 1050–1059.

Pierrehumbert, J. (**1999**). *Prosody and Intonation* (MIT, New York).

Rothauser, E. H., Chapman, W. D., Guttman, N., Nordby, K. S., Silbiger, H. R., Urbanek, G. E., and Weinstock, M. (**1969**). "I.E.E.E. recommended practice for speech quality measurements," IEEE Trans. Audio Electroacoust. **17**, 227–246.

Ryalls, J. H., and Lieberman, P. (**1982**). "Fundamental frequency and vowel perception," J. Acoust. Soc. Am. **72**, 1631–1634.

Spitzer, S. M., Liss, J. M., and Mattys, S. L. (**2007**). "Acoustic cues to lexical segmentation: A study of resynthesized speech," J. Acoust. Soc. Am. **122**, 3678–3687.

Spitzer, S. M., Liss, J. M., Spahr, T., Dorman, M., and Lansford, K. (**2009**). "The use of the fundamental frequency for lexical segmentation in listeners with cochlear implants," J. Acoust. Soc. Am. **125**, EL236–EL241.

Trainor, L. J., and Desjardins, R. N. (**2002**). "Pitch characteristics of infant-directed speech affect infants' ability to discriminate vowels," Psychon. Bull. Rev. **9**, 335–340.

van Noorden, L. P. (**1977**). "Minimum differences of level and frequency for perceptual fission of tone sequences ABAB," J. Acoust. Soc. Am. **61**, 1041–1045.

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.

Watson, P. J., and Schlauch, R. S. (**2009**). "Fundamental frequency variation with an electrolarynx improves speech understanding: A case study," Am. J. Speech Lang. Pathol. **18**, 162–167.

Whalen, D. H., Gick, B., Kumada, M., and Honda, K. (**1999**). "Cricothyroid activity in high and low vowels: Exploring the automaticity of intrinsic F0," J. Phonetics **27**, 125–142.

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.

Yost, W. A., Popper, A. N., and Fay, R. R. (**1993**). *Human Psychophysics* (Springer-Verlag, New York).