

Title:

**Short Time-Reversal Windows Used to Investigate the
Processing of Intonation**

Running title:

Short Time Reversal Envelopes

Primary author contact information:

Thomas C. Purnell
Department of Linguistics, University of Wisconsin-Madison
1168 Van Hise Hall, 1220 Linden Drive
Madison, WI 53706
tcpurnell@wisc.edu
(608) 265-3080 (phone)
(608) 265-3193 (fax)

Second author contact information:

Eric Raimy
Department of English, University of Wisconsin-Madison
7123 Helen C. White Hall, 600 N. Park Street
Madison, WI 53706
raimy@wisc.edu
(608) 263-6870 (phone)
(608) 263-3709 (fax)

Word count: 4,474 (excluding title page, abstract, figures and tables)

Acknowledgements:

Thanks to Ryan Hanke, Bill Idsardi, Monica Macaulay, Blake Rogers,
Jenny Saffran, Joe Salmons and Christian Stilp for discussion and
comments on earlier drafts. All errors are ours.

Short Time-Reversal Windows Used to Investigate the Processing of Intonation

ABSTRACT

Acoustic time-reversal (Saberri & Perrott 1999) is a signal processing technique where an audio signal is divided over time into equal sequences that are reversed *in situ*. Previous studies have used this technique to investigate how large reversal windows must be to disrupt auditory processing. We investigate whether small time-reversed windows disrupt the perception of intonation. Small time-reversal windows are found to affect the processed signal in a way that acoustically disrupts the calculation of pitch differentially based on whether pitch is calculated temporally or spectrally. Subjects reconstituted the original pitch for a subset of the manipulated stimuli. This recovery suggests that intonation and aspects of the stimuli (i.e. real vs. nonce word, monosyllabic vs. polysyllabic, stress pattern) interact in complicated ways in auditory processing. Utilizing small time-reversal windows with other psychoacoustic perceptual tasks extends the methodology to investigate multi-channelled aspects of information in human speech.

Keywords: *discourse processing, language comprehension, lexical processing, speech perception*

Short Time-Reversal Windows Used to Investigate the Processing of Intonation

Acoustic time-reversal is a signal processing technique whereby a portion of a digitized audio signal is divided up over time into chunks that are then reversed and pieced back together. Figure 1 depicts the reversal process of a speech signal (a sound wave of acoustic energy over time). Each window (ω) in the reversed speech signal is flipped front to back, so that the last edge (ϵ_2) comes before the first edge (ϵ_1) of the original signal. That these clumps of signal or window are arranged in their original sequence ($\omega_1, \omega_2, \dots \omega_n$ throughout the reversal) prevents the audio from resembling something like backmasking where the entire signal sounds backwards (e.g., the Beatle's 1966 album *Revolver*). The cognitive operation active in interpreting a time-reversed signal falls partially under the purview of digital signal processing because, unknown to the auditory system; the temporal reorganization of the signal effectively makes portions of the speech signal discontinuous. Remarkably, within limits, the brain reconstitutes the signal and perceives the signal in its original unreversed order even if the chunk separates portions of a phoneme, syllable or word.

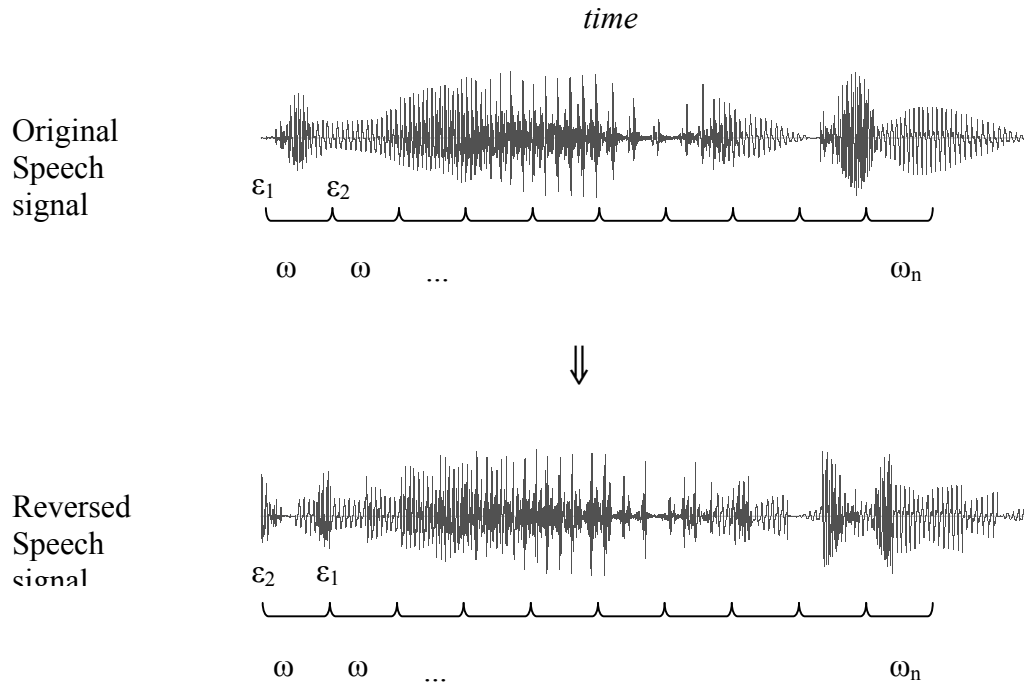


Fig. 1. Resultant signal when applying a 40 ms reversal window (ω) with two edges (ϵ_1 , ϵ_2) to a speech signal.

If intelligibility—that is, whether or not the lexical content of a word or sentence can be determined by a listener—is used as the measure of experimental success, then the human brain reliably processes time-reversal envelopes efficiently even when the reversal window exceeds 100 ms (Saberri & Perrott, 1999).¹ Since the upper limit of the relationship between intelligibility and time-reversed speech has been determined, time-reversed speech has been used extensively in the testing of perceptual abilities by obfuscating portions of the acoustic signal (Hoen et al., 2007; Lakshminarayanan et al., 2003; Shannon, 2005; Smith, Delgutte, & Oxenham, 2002). A common theme to this line of research is the general intelligibility of speech. Speech perception makes use of a very rich informational stream (Lieberman, 1996; Scott & Johnsrude, 2003; Wang, 2007 among others). Intelligibility is thus an extremely

important aspect of speech perception since it likely provides a fairly direct view of lexical access and sentence processing, and demonstrates in particular the presence of abstract relations in lexical processing (McQueen, Cutler, & Norris, 2006; Mirman, McClelland, & Holt, 2006). However, the task involved in determining intelligibility does not accurately represent the extent to which our brains manage all of the information contained in the speech stream. The speech stream contains many different types of information beyond simple lexical content; speaker identity, emotive content, sociophonetic information and intonation are well established additional channels of information in the speech stream. One of the goals of this project is to investigate whether the time-reversed speech paradigm can be used to explore channels of information other than intelligibility in the speech stream, thereby expanding our understanding of both the processing of language as well as the organizational structure of lexical processing. Specifically, we focus on intonation here and question whether intonation interacts with the prelexical phonetic processing component. An additional goal of this paper is to exemplify a novel use of the reversal technique as a methodological tool in research on cognition generally and speech perception and lexical access specifically.

For present purposes, we use the term intonation to refer to the pattern of fundamental frequency (F0) change that indicates whether a sentence is a question or a statement in English. Questions in English seeking a yes or no answer have a distinct phrase-ending intonation pattern (rising F0, no fall) as compared to either questions seeking an answer of content or a declarative statement (rise-fall or fall) (Bollinger, 1958; Ladd, 1996; Pierrehumbert, 1985, 1987) with the final F0 movement being a

reliable cue for distinguishing questions from statements (Majewski & Blasdell, 1969; Studdert-Kennedy & Hadding, 1973). Our knowledge of the general robustness and the up-and-down patterning of F0 is complemented by the relational aspect of F0 and concomitant stress features (Fourakis, 1991). For English, the height of the overall F0 contour interacts with the pattern of final F0 movement so that a high mean F0 attenuates the need for a rise in F0 to signal a question (Ladd, 1996). Thus, a high mean F0 with a slight drop is more likely to be understood as a question than a low mean F0 with a sharp drop. This trade-off between mean F0 and its trajectory is largely an effect of the F0 declination in speech relative to the starting F0 in a word.

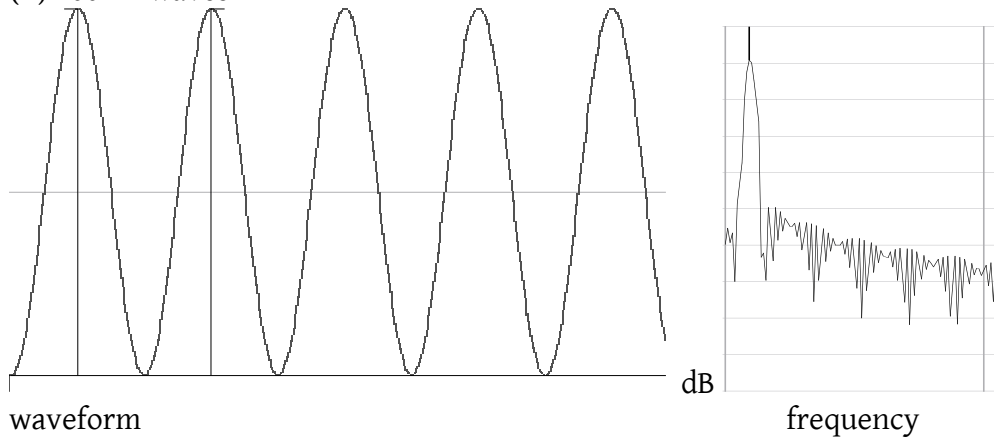
With the grammatical relationship between the change of F0 and statements and questions reasonably well understood, we can turn to asking questions about the psychophysics of this grammatical relationship. The null hypothesis for processing small time-reversals should be that perception follows straightforwardly from well-known facts about pitch, which Plack and Oxenham (2005a: 2) define as “that attribute of sensation whose variation is associated with musical melodies.” The robustness of pitch perception arises from the calculation of pitch from both the waveform or spectrum using autocorrelation or cepstrum methods, respectively (de Cheveigne, 2005; Hess, 1983). To be specific, modeling the temporal calculation of pitch involves a period-to-period calculation, often by finding a repeating pitch peak or zero crossing on the waveform whereas modeling the spectral calculation relies on the harmonic or spectral structure of the signal.

Figures 2 through 5 illustrate the way that small envelope reversals can affect both the temporal and spectral calculations of pitch. From three waveforms and the

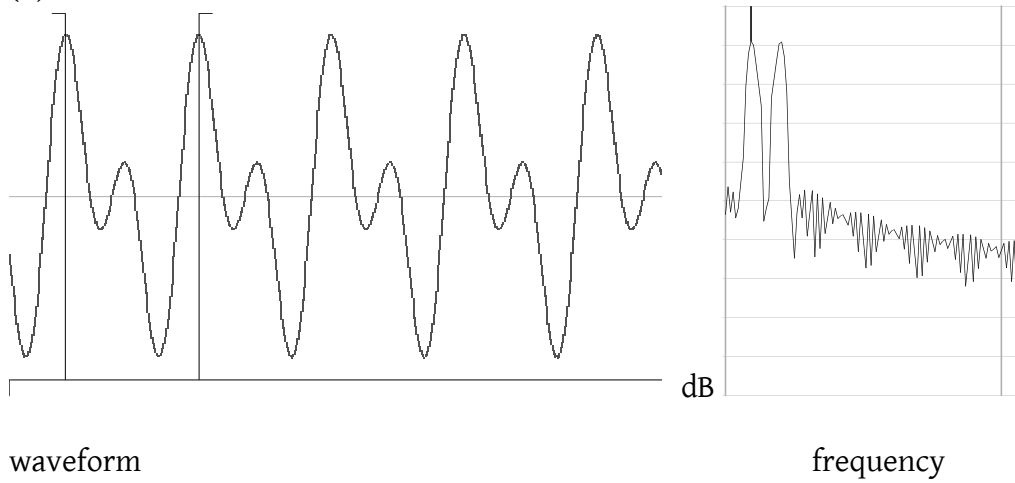
corresponding power spectrum of each waveform in Figure 2 we can see how harmonic structure is reflected at different levels of analysis. In panel (A) a simple 100 Hz sine wave can be measured temporally by measuring the time between peaks (window ω). The period of ω is 10 ms and marked with cursors on the waveform on the left. We know that for a wave vibrating at a fundamental frequency (F_0) there are sympathetic resonances vibrating at whole number multiples of F_0 (the harmonics) that are observable as normal fluctuations of the waveform and in a power spectrum. In panels (B) and (C), harmonics (simulated here by the combination of 200 and 300 Hz waves to the 100 Hz wave) are seen as smaller peaks between the fundamental peak on the waveform (left panel). Notice too the leftmost high peaks in the spectra to the right of each waveform. Because the spectral peaks are spaced apart by the value of the lowest peak—in the case at hand, 100 Hz apart—a listener should be able to calculate pitch from both the waveform and the power spectrum. However, the waveform appears more sensitive to corruption of F_0 under reversal, whereas the higher peaks in a power spectrum could retain the original pitch. Thus we suspect that the more corrupt the temporal aspects of the waveform become, the more a listener accesses power spectrum information. Figure 3 demonstrates how a small-time reversal with a envelope duration that is slightly below and slightly above a peak-to-peak cycle can reduce the effectiveness of a temporal algorithm. A simple sine wave with an F_0 of 100 Hz in Figure 3 (A, henceforth referred to as “ T_0 ”) is not reversed while the waves in panels (B) and (C) are reversed. From the cycle to cycle variation in the wave slightly above the original in panel (C, henceforth “ $>T_0$ ”) one can see a repeated fundamental. However, as seen in the wave reversed just below the mean pitch in panel (B,

henceforth “<TO”) it is possible that the temporal pitch calculation will be affected while the spectral calculation will not due to the rearranged temporal proximity of energy peaks.

(A) 100 Hz waves



(B) 100 + 200 Hz waves



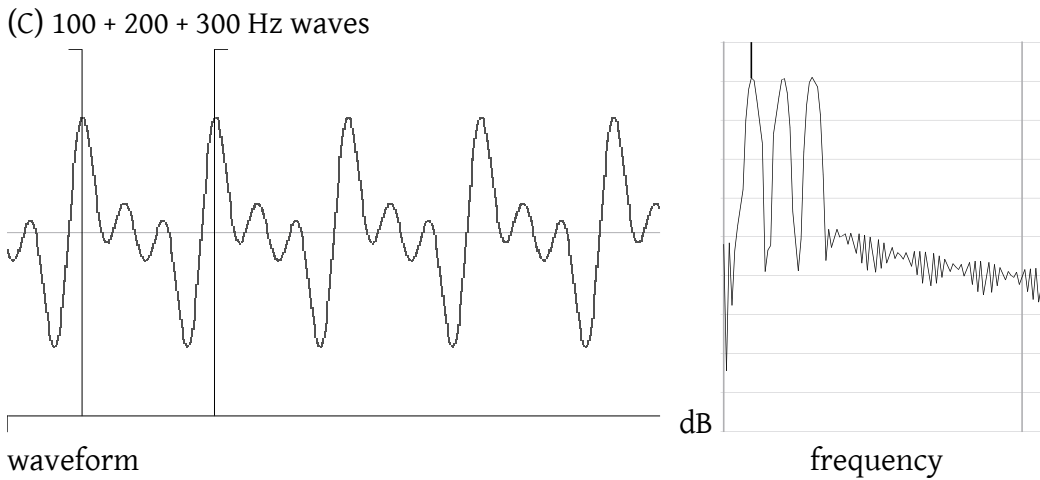


Fig. 2. Waveforms and power spectra of three waves, all with a 100 Hz fundamental. In all three instances $\omega=10$ ms.

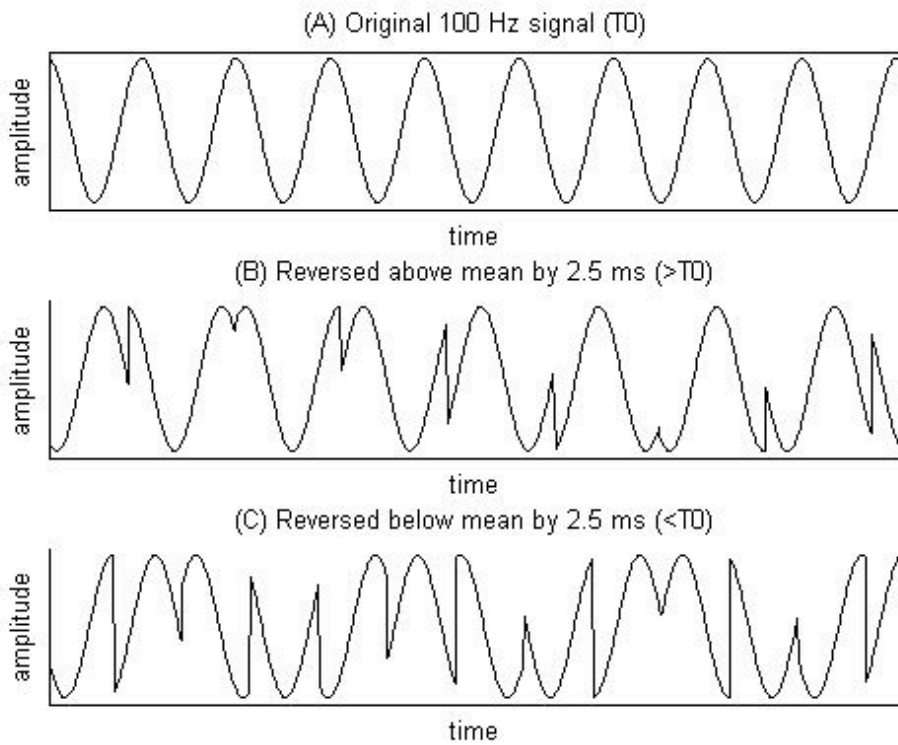


Fig. 3. Comparison of reversal envelopes on the pitch of a 100 Hz sine wave.

If the fundamental frequency or related harmonics closer to the fundamental are corrupted by the reversal, as seen in Figures 3 and 4, harmonics higher in the

frequency spectrum may provide assistance in recovering the F0. Cepstral analysis—the Fourier transform of a prior Fourier transform of the wave—provides a means for identifying the pitch from higher component frequencies. Nevertheless, a cepstral analysis of such a degraded environment as acoustic reversal is not immune to corruption, as seen in Figure 5. Note that in the bottom right panel of Figure 5, more peaks appear and the peaks are not aligned with 100, 200 or 300 Hz. What remains at issue is how human cognition resolves conflicting psychophysical evidence (Plack & Oxenham, 2005b).

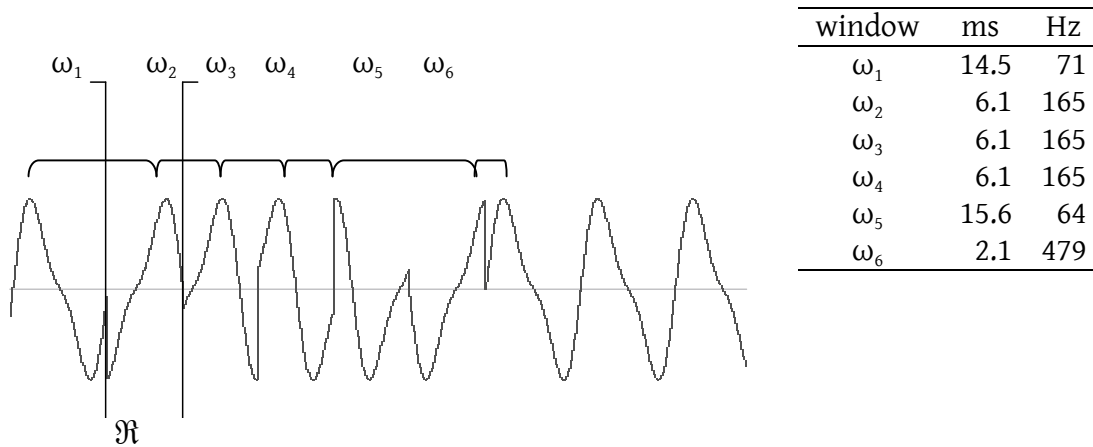


Fig. 4. Variability in the peak-to-peak calculation of pitch by milliseconds between the next highest peaks in a reversed waveform. The vertical lines mark the edge of one of the reversed portions of the signal (ϑ).

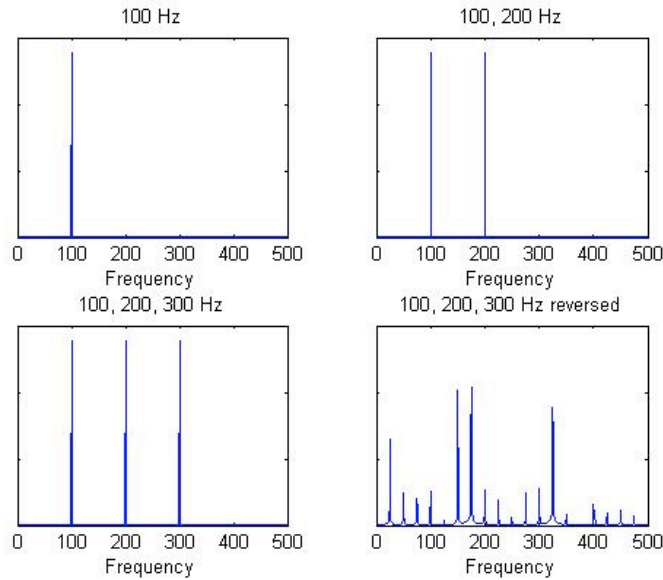


Fig. 5. Cepstral peaks of four waves, one simple wave, two complex waves and a reversed example of the wave with three components with a reversal window size of 8 ms.

Beyond issues related to window size, we now turn to a possible confound with lexical access. We adopt a model (Figure 6) where intonation perception follows lexical processing, and where lexical processing follows prelexical mediation of the input. In the case at hand, the input is ‘corrupted’ by the reversal process. If there is a prelexical processing, then we expect it to benefit from lexical feedback to assist in resolution of the problem. Below we describe an experiment modeled on English speakers’ knowledge that intonation patterns on single words convey precise semantic interpretations (Kiesling, 2004).

To investigate the interaction between knowledge of intonation and the time-reversal methodology we developed two dimensions of the stimuli: the window size of the time-reversal and a word vs. nonce-word status. First, Manipulating the window size of the time-reversal is important because, as demonstrated in Figure 3, there is the potential of different impact on the F0 of the stimuli based on the relationship between

the F0 of the original stimuli and the size of the time-reversal window. Second, by manipulating the pitch of a word, it is possible that the stress of a word would be affected also by the reversal. If this is the case, then lexical access can also be affected by the small envelope reversal. Comparing subjects' reactions to word vs. nonce-word stimuli in the context of upstream processing (Fig. 6) will provide information about whether lexical access interacts with the time-reversal methodology.

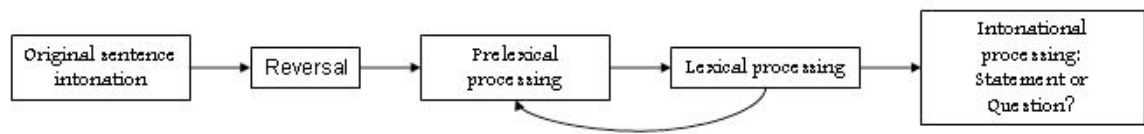


Fig. 6. Flow of information through listener’s decision-making process.

METHOD

Several factors influenced token selection, including: words versus nonce-words, final primary stress, final secondary stress and non-final stress; number of syllables in the word when two syllables or more, or monophthong versus diphthong when the word only had one syllable. Using such factors tied to phonology and lexical access facilitates testing listeners’ ability to perform cognitive tasks in spite of input signal corruption.

Materials

A male speaker of American English was recorded producing words and nonce words within a declarative carrier phrase (“Please say ___ again.”) so that no token was in a phrase initial or phrase final position. Target tokens included monosyllables with a

short or long vowel ([ɛ]: *pet, bet, debt, *ket, *zet, *het*; [eɪ]: *pate, bait, date, *vate, *zate, *shate, *hoat*), disyllabic words with initial or final stress (*contrast, produce, convert, *hoatvert, *hoatent, *hoatact* as CONtrast or conTRAST, etc.) and trisyllabic words with secondary or tertiary stress on the final syllable (*delegate, affiliate, moderate, *hoaterate*). The speaker was instructed to read the disyllabic and trisyllabic words with the same vowel quality in each syllable regardless of the stress pattern (i.e., without reduction of any vowel). A female speaker was also recorded saying the four contrastive disyllabic and four contrastive trisyllabic words (*contrast, produce; delegate, moderate*), although no instruction was given for controlling vowel quality. These tokens were used as distracters to the test tokens because of the pitch and quality contrast with the male tokens. Following Yrttiaho et al. (2008), real speech was selected over synthetic speech because research suggests the brain responds differently to real and synthetic speech (Belin, Zatorre, Lafaille, Ahad, & Pike, 2000; Gunji et al., 2003; Hewson-Stoate, Schonwiesner, & Krumbholz, 2006).

Each word was excised from the declarative carrier phrase and down sampled to 11,025 Hz. Following previous work, consecutive fixed-duration periods of the speech stream were reversed without smoothing the relation between the reversed periods. Alignment to the nearest zero-crossing was not performed precisely because the small difference between the reversal window and the actual pitch period would have been neutralized in many of the tokens. Additionally, the acoustic ‘clicking’ that may result from an irregular zero-crossing was not a factor in the perception of these sounds. Listeners heard a word time-reversed 2.5 msec below and 2.5 msec above one speaker’s mean pitch of 104 Hz.

Analysis

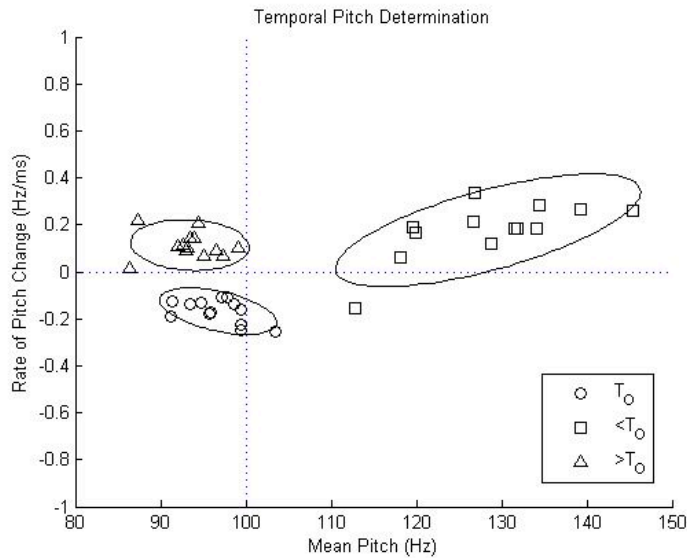
Two methods of calculating the pitch patterns produced by the two reversal windows make different predictions about how listeners may process degraded intonation contours. For an analogue calculation to a time-varying computation by listeners the Robust Algorithm for Pitch Tracking (Talkin, 1995; RAPT) generated a smoothed pitch contour for the vocalic portions of the tokens. The RAPT contour was compared to a smoothed pitch contour calculated by a cepstral algorithm. Data calculations were computed using Speech Filing System (Huckvale, 2004). From these two contours, pitch characteristics were calculated for each token, including mean, standard error of the mean and the rate of change (ROC_{f_0}). These descriptive statistics are shown in Table 1.

Table 1
Descriptive statistics for original and reversed tokens for temporal (RAPT) and spectral (Cepstral) methods of analysis

	RAPT		Cepstral	
	Pitch (Hz)	Rate of Pitch Change (Hz/ms)	Pitch (Hz)	Rate of Pitch Change (Hz/ms)
Original, T0	96.8 (1.0)	-0.107* (0.014)	102.1 (3.0)	0.008 (0.107)
Over Original, >T0	128.4* (2.5)	0.177 (0.015)	116.8* (2.2)	-0.191 (0.105)
Under Original, <T0	93.4 (1.0)	0.114 (0.015)	94.4 (1.2)	0.075 (0.040)

Note—standard errors are shown in parentheses. For each measure, the means are significant, where $p < 0.05$, except for Rate of Pitch Change for the cepstral method. A significant Scheffé post-hoc analysis difference for one of the three groups on the three significant measures is shown below with *.

A. Waveform Calculation



B. Spectral Calculation

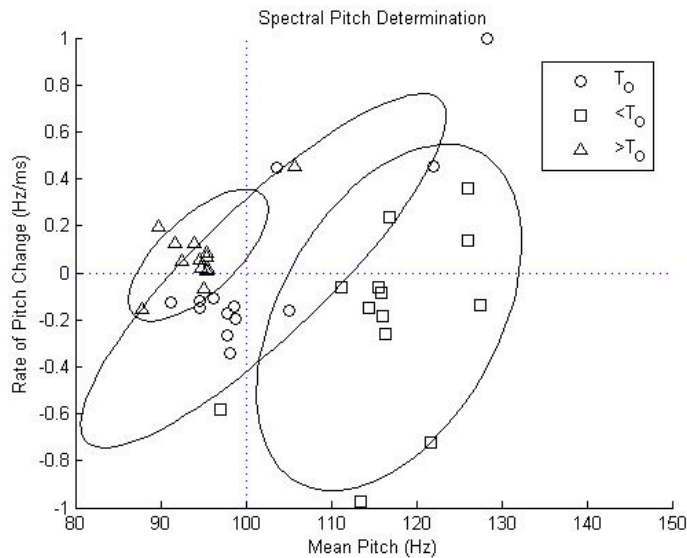


Fig. 7. Rate of Change_{f₀} relative to mean F0 calculated by a waveform method (A) and spectral method (B) for monosyllable real and nonce words. T₀ is the pitch window of each non-reversed token (=1/f). Tokens above the horizontal reference line are considered to have rising pitch.

Since pitch height influences the interpretation of a rise or fall (Studdert-Kennedy & Hadding, 1973), a ratio of ROC_{f₀} to mean F0 was also used in the comparison. Figure 7 shows the token's ROC_{f₀} relative to mean F0 for both the RAPT and cepstral methods of pitch calculation. Non-reversed tokens were also measured and are shown

in order to contextualize the reversals. Note that in the temporal calculation of pitch shown in Figure 7(A), the non-reversed tokens are relatively low in mean F0 and have a negative ROC_{f_0} indicating that all of the input tokens are statements with a fall in pitch. The tokens with a smaller-than pitch reversal (<TO) have a higher pitch because the reversal places pulses closer to other pulses while the tokens above pitch (>TO) have a mean pitch that approximates the non-reversed tokens. Both reversals produce a rise in pitch rather than a fall as indicated by the positive rate of pitch change. The spectral calculation of pitch shown in Figure 7(B) shows one important difference, namely that the smaller-than pitch tokens (<TO) that have a higher pitch, have a negative ROC_{f_0} .

Participants

Twenty-one speakers of English with no reported hearing difficulty participated in this study. All subjects consented to participation following

Procedure

Tokens were presented in random order and monaurally through headphones while two choices were presented on the computer screen using Praat software (Boersma & Weenink, 2008). The task for our subjects was to determine whether a sentence was a statement or a question. On the left was the word followed by a period and on the right was the word followed by a question mark (e.g., *produce.* and *produce?*). Hearing the time-reversed utterances, subjects performed a forced choice decision task of whether the utterance was a statement or a question. Responses were captured by a mouse click on the screen. Each subject was presented with a total of 88 sound files. For analysis purposes, however, responses to reversal pairs spoken by the male native

speaker of American English were subject to an analysis of variance (33 reversed above pitch and 33 reversed below pitch).

RESULTS

Mean percent perceived as question per token was submitted to a four-factor ANOVA (see Table 2): word size (1, 2, 3 syllables) x window size (<TO, >TO) x stress (non-final, final) x lexical status (word, non-word). Bilateral interactions were examined prior to main effects. We found significant interactions of word size with window size $[F(2,51) = 3.8, MS_e = 828, p < .05]$ and the position of stress in a di- or tri-syllabic word or number of vowel qualities (mono- or diphthong) if monosyllable $[F(2,51) = 9.3, MS_e = 2,000, p < .05]$. A significant main effect for a factor not related to the two interactions, was found for lexical status, or the difference between words and non-words $[F(1,51) = 6.5, MS_e = 1,400, p < .05]$ where words are more likely to be perceived as statements than non-words. Because of this outcome, we next examine window size for word size for all three word sizes. Following that we examine stress for polysyllable tokens because the decision process making process may be influenced by the lexical decision differences for monosyllable tokens and those tokens larger than one syllable.

Table 2
Results of ANOVA for perception as question

Source	d.f.	F	Mean Sq.
Word size (WdS)	2	16.5*	3,562
Window size (WnS)	1	6.1*	1,321
Stress (ST)	1	11.3*	2,443
Lexical status (Lex)	1	6.5*	1,400
WdS x WnS	2	3.8*	828
WdS x ST	2	9.3*	2,000
WdS x Lex	2	0.7	147
WnS x ST	1	1.1	239
WnS x Lex	1	0.0	1
ST x Lex	1	2.8	601
Error	51		216
Total	65		

Note--* $p < 0.05$

Window Size

Response data was submitted to a one-factor ANOVA for each word size where the factor was window size. The difference between the mean response by listeners to tokens of the two window sizes <TO ($M = 60\%$ as question) and >TO ($M = 38\%$ as question) was significant, [$F(1,24)=15.7$, $MS_e = 199.4$, $p < .05$]. But for disyllables, the difference ($M = 27\%$ and $M = 29\%$, respectively) was not significant [$F(1,22)=0.01$, $MS_e = 589.8$, $p > .05$], as was also the case for trisyllables ($M = 26\%$ and $M = 19\%$, respectively) [$F(1,14)=1.08$, $MS_e = 189.1$, $p > .05$]. Figure 8 shows that for word and reversal size, only one syllable words with a below pitch reversal (< TO) are perceived more often as statements than questions. All other word and reversal sizes reflect a question bias.

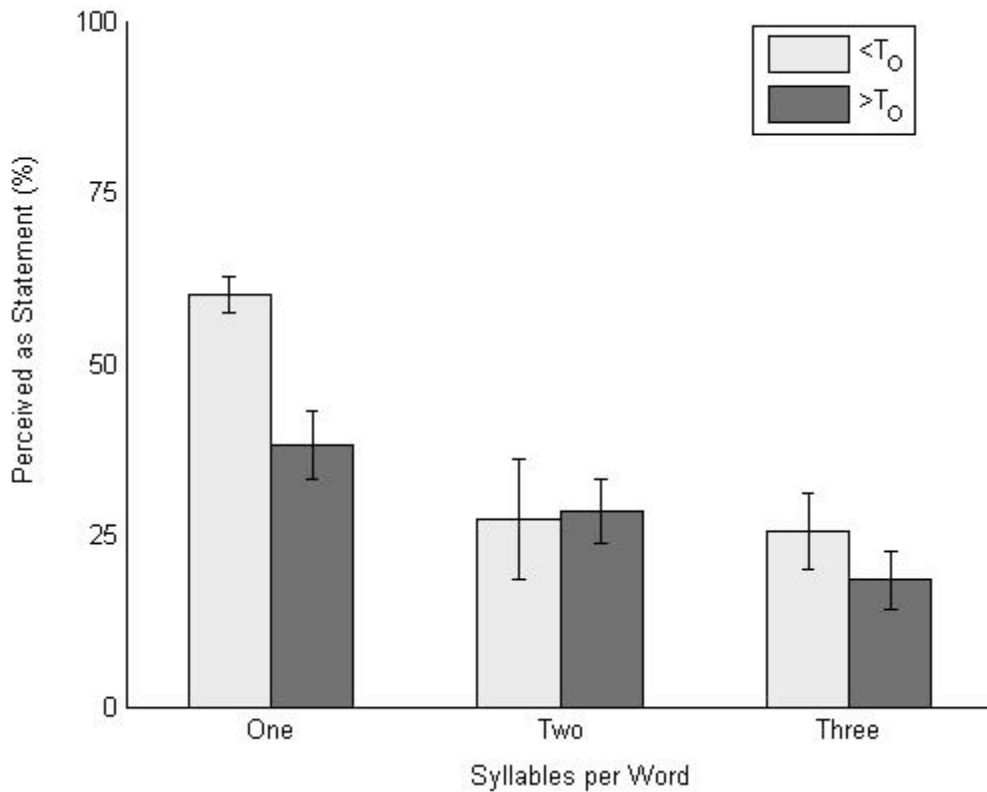


Fig. 8. Percentage of tokens perceived as questions with word size and reversal size as factors.

Stress and Lexical Access

Response data was submitted to a two-factor ANOVA for monosyllabic and polysyllabic words in order to examine the interaction between stress, intonation and lexical access: stress x lexical status. We assume that lexical access is distinct for monosyllabic tokens as compared to polysyllabic since, in the latter instance but not the former, stress is not contrastive in English. For monosyllables, no main effect or interaction between factors was found ($p > .05$). Thus, the percent for which monosyllable tokens were judged by listeners to be questions was no different between monophthongal words ($M = 47.6\%$), monophthongal non-words ($M = 50.0\%$), diphthongal words ($M = 45.2\%$) and diphthongal non-words ($M = 52.4\%$).

For polysyllabic words, a significant interaction was found between stress and lexical access [$F(1,36) = 6.33$, $MS_e = 1,500$, $p < 0.05$]. Figure 9 shows that nonce-words with final stress ($M = 51.8\%$) do not show a question bias while the other three groups of words do demonstrate the bias (i.e., non-words with non-final stress [$M = 16.1\%$], and words with non-final and final stress [$M = 14.7\%$ and $M = 25.4\%$, respectively]). Thus, listeners responded unequivocally to the polysyllabic words as questions when the tokens were words. However, the non-words split between tokens with non-final stress (like CONtrast) which pattern like words, and tokens with final stress (like conTRAST) to which listeners equivocated around chance. Submitting the polysyllabic words to individual one-factor ANOVAs found that words did not show a significant difference by stress [$F(1,22) = 3.0$, $MS_e = 233$, $p > 0.05$], whereas the one-factor ANOVA for non-words did find a significant difference by stress [$F(1,14) = 20.9$, $MS_e = 244$, $p < 0.05$].

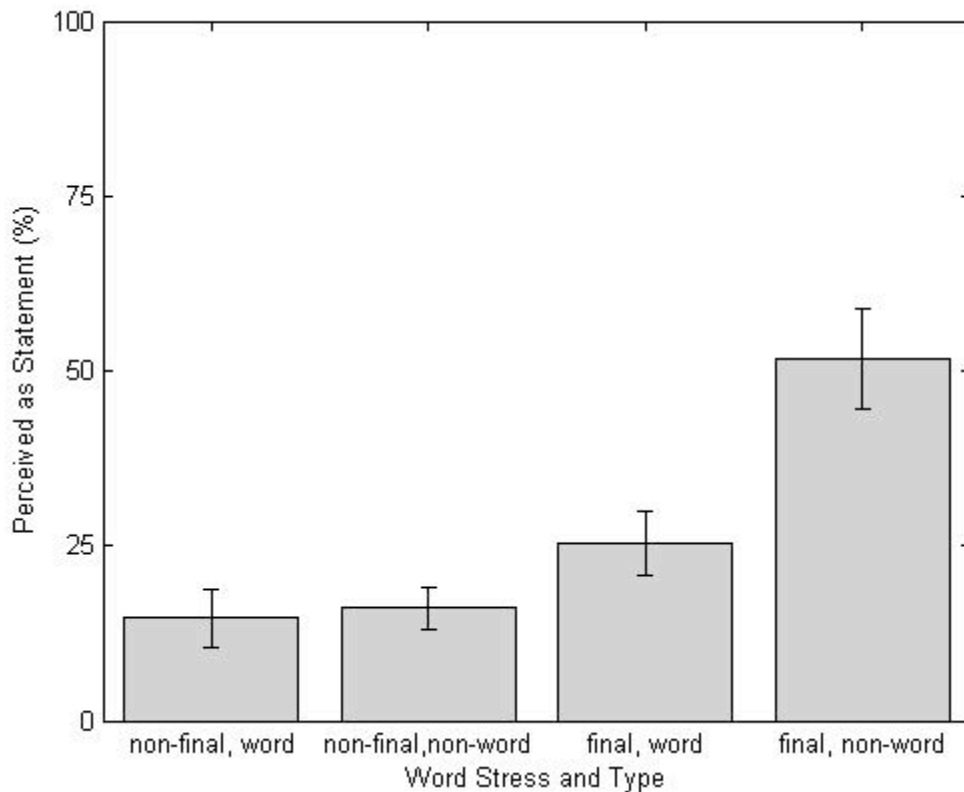


Fig. 9. Perceptual relation between stress position and lexical status for two or three syllable words.

DISCUSSION

Listeners clearly demonstrate that small time reversals are sufficient to change the perceived intonation pattern of the word. Not all information contained in the multi-channelled speech stream can be accurately recovered from time-reversed stimuli. When subjects are faced with an intonation perception task, they do not reconstitute all aspects of the speech stream accurately in the way suggested by Saberi & Perrott's (1999) large reversal windows. The experiment described in this paper highlights two aspects of the complex relation between prosody (stress, intonation) and lexical access shown in Figure 6. First, monosyllabic tokens highlight how

preprocessing by a prelexical level is fooled by the reversal, in other words, cognition for certain tokens cannot recover the original statement from the reversal technique. This is a direct signal processing effect. Second, a cognitive task related to intonation is influenced by the lexical feedback to the prelexical processor in the polysyllabic words when stress is in word-final position (see Figure 6). Such a finding complicates the direct signal processing effect by revealing upstream processing after signal processing effects should have occurred.

Monosyllabic tokens show that there is no lexical feedback because the pressure exerted on lexical organization or operations related to stress are not present for such tokens. Therefore, we can focus on window size issues. There are a few plausible explanations for the fact that monosyllabic words do show intonation reconstitution effects while other types of words do not. One is that intonation is generally ignored for monosyllabic words that suggest a general 'statement bias' by the subjects. Another is that intonation is processed in both temporal and spectral manners with a bias towards temporal processing for polysyllabic words and a bias towards spectral processing in monosyllabic words. Because there is no contrastive stress, information about monosyllable tokens proceeds to upstream processes resolving intonation. Thus, when temporal processing fails, there is increased reliance on spectral information.

Polysyllabic nonce-words with non-final stress show some aspects of intonation reconstitution. For this class of tokens, lexical feedback occurs because there is contrastive stress. The final rise in pitch for non-words with word-final pitch is difficult to parse. The listener does not hear initial stress so either the token could be perceived as initial stress with an extremely high intonation pitch rise, or the token could be

perceived as a final stress word with a statement intonation. These words do not show a clear bias towards the 'statement' intonation but instead show an at-chance type behavior. A plausible explanation for this is that since the stimulus is a nonce-word it is more difficult to interpret the final rise in F0 in these stimuli. The cause of this rise could be intonation indicating a question or it could be the presence of a final stress. Without lexical access to indicate whether the final syllable is stressed or not, the subjects appear to just guess for these types of words. This is a very interesting result since the lexicon may not be arranged with regard to stress pattern (Friedrich, Kotz, Friederici, & Alter, 2004; Slowiaczek, Soltano, & Bernstein, 2006; Soto-Faraco, Sebastian-Galles, & Cutler, 2001) yet the subjects do appear to be sensitive to the stress pattern.

Further investigation of small window time-reversals is crucial to our understanding of the robustness of speech perception. From the modest results reported here, we can cull two important caveats to current standard claims about time-reversed speech. First, the speech stream is multi-channeled containing (at least) lexical content, semantic content, affect and prosody. Second, this information is differentially disrupted based on the size of a reversal window. While it is important that listeners filter out noise from content-full information (Seyfarth, Cheney, & Bergman, 2005), it is also important for speakers and listeners to have a meaningful conversation. The time-reversal window methodology appears to be a very useful tool in investigating 'meaningfulness' in language because of the questions about intelligibility of speech, prosody and lexical access that have already been generated which are hallmarks of interesting inquiry.

REFERENCES

- Belin, P., Zatorre, R. J., Lafaille, P., Ahad, P., & Pike, B. (2000). Voice-selective areas in human auditory cortex. *Nature*, 403(6767), 309-312.
- Boersma, P., & Weenink, D. (2008). Praat (Version 5.030). Amsterdam: Institute of Phonetic Sciences.
- Bollinger, D. (1958). A theory of pitch accent in English. *Word*, 14, 109-149.
- de Cheveigne, A. (2005). Speech perception models. In C. J. Plack & A. J. Oxenham & R. R. Fay & A. N. Popper (Eds.), *Pitch: Neural coding and perception* (pp. 169-233). New York: Springer.
- Fourakis, M. (1991). Tempo, Stress, and Vowel Reduction in American English. *Journal of the Acoustical Society of America*, 90(4), 1816-1827.
- Friedrich, C. K., Kotz, S. A., Friederici, A. D., & Alter, K. (2004). Pitch modulates lexical identification in spoken word recognition: ERP and behavioral evidence. *Cognitive Brain Research*, 20(2), 300-308.
- Gunji, A., Koyama, S., Ishii, R., Levy, D., Okamoto, H., Kakigi, R., & Pantev, C. (2003). Magnetoencephalographic study of the cortical activity elicited by human voice. *Neuroscience Letters*, 348(1), 13-16.
- Hess, W. (1983). *Pitch determination of speech signals: Algorithms and devices*. New York: Springer-Verlag.
- Hewson-Stoate, N., Schonwiesner, M., & Krumbholz, K. (2006). Vowel processing evokes a large sustained response anterior to primary auditory cortex. *European Journal of Neuroscience*, 24(9), 2661-2671.

- Hoen, M., Meunier, F., Grataloup, C. L., Pellegrino, F., Grimault, N., Perrin, F., Perrot, X., & Collet, L. (2007). Phonetic and lexical interferences in informational masking during speech-in-speech comprehension. *Speech Communication, 49*(12), 905-916.
- Huckvale, M. (2004). *Speech Filing System Release 4.6/Windows*. London: University College.
- Kiesling, S. (2004). Dude. *American Speech, 79*, 281-305.
- Ladd, D. R. (1996). *Intonational Phonology*. England: Cambridge U Press.
- Lakshminarayanan, K., Ben Shalom, D., van Wassenhove, V., Orbelo, D., Houde, J., & Poeppel, D. (2003). The effect of spectral manipulations on the identification of affective and linguistic prosody. *Brain and Language, 84*(2), 250-263.
- Liberman, A. M. (1996). *Speech: A special code*. Cambridge, MA: MIT Press.
- Majewski, W., & Blasdell, R. (1969). Influence of Fundamental Frequency Cues on Perception of Some Synthetic Intonation Contours. *Journal of the Acoustical Society of America, 45*(2), 450-&.
- McQueen, J. M., Cutler, A., & Norris, D. (2006). Phonological abstraction in the mental lexicon. *Cognitive Science, 30*(6), 1113-1126.
- Mirman, D., McClelland, J. L., & Holt, L. L. (2006). An interactive Hebbian account of lexically guided tuning of speech perception. *Psychonomic Bulletin & Review, 13*(6), 958-965.
- Pierrehumbert, J. B. (1985). English Intonation - Its Representation and Realization. *Journal of Psycholinguistic Research, 14*(6), 596-597.
- Pierrehumbert, J. B. (1987). *The phonology and phonetics of English intonation*. Bloomington, Ind.: Indiana University Linguistics Club.

- Plack, C. J., & Oxenham, A. J. (2005a). Overview: the present and future of pitch. In C. J. Plack & A. J. Oxenham & R. R. Fay & A. N. Popper (Eds.), *Pitch: Neural coding and perception*. (pp. 1-6). New York: Springer.
- Plack, C. J., & Oxenham, A. J. (2005b). The psychophysics of pitch. In C. J. Plack & A. J. Oxenham & R. R. Fay & A. N. Popper (Eds.), *Pitch: Neural coding and perception* (pp. 7-55). New York: Springer.
- Saberi, K., & Perrott, D. R. (1999). Cognitive restoration of reversed speech. *Nature*, 398, 760.
- Scott, S. K., & Johnsrude, I. S. (2003). The neuroanatomical and functional organization of speech perception. *Trends in Neurosciences*, 26(2), 100-107.
- Seyfarth, R. M., Cheney, D. L., & Bergman, T. J. (2005). Primate social cognition and the origins of language. *Trends in Cognitive Sciences*, 9(6), 264-266.
- Shannon, R. V. (2005). Speech and music have different requirements for spectral resolution. *Auditory Spectral Processing*, 70, 121-134.
- Slowiaczek, L. M., Soltano, E. G., & Bernstein, H. L. (2006). Lexical and metrical stress in word recognition: Lexical or pre-lexical influences. *Journal of Psycholinguistic Research*, 35, 491-512.
- Smith, Z. M., Delgutte, B., & Oxenham, A. J. (2002). Chimaeric sounds reveal dichotomies in auditory perception. *Nature*, 416, 87-90.
- Soto-Faraco, S., Sebastian-Galles, N., & Cutler, A. (2001). Segmental and suprasegmental mismatch in lexical access. *Journal of Memory and Language*, 45(3), 412-432.
- Studdert-Kennedy, M., & Hadding, K. (1973). Auditory and linguistic processes in perception of intonation contours. *Language and Speech*, 16, 293-313.

- Talkin, D. (1995). A robust algorithm for pitch tracking (RAPT). In W. B. Kleijn & K. K. Paliwal (Eds.), *Speech Coding and Synthesis*. New York: Elsevier.
- Wang, X. Q. (2007). Neural coding strategies in auditory cortex. *Hearing Research*, 229(1-2), 81-93.
- Yrttiaho, S., Tiitinen, H., May, P. J. C., Leino, S., & Alku, P. (2008). Cortical sensitivity to periodicity of speech sounds. *Journal of the Acoustical Society of America*, 123(4), 2191-2199.

¹ Intelligibility can have a range of interpretations, ranging from a listener identifying a) human speech from non-human sounds or music, to b) a native or known language from non-native or unknown language, and to c) whether words or sentences have known semantic content.