Running head: Sine-wave speech in a tone language

Sine-wave and noise-vocoded sine-wave speech in a tone language: acoustic details matter

Stuart Rosen and Sze Ngar Catherine Hui

UCL Department of Speech, Hearing and Phonetic Sciences
2 Wakefield Street, London, WC1N 1PF, United Kingdom

stuart@phon.ucl.ac.uk; snc.hui@gmail.com

Abstract

Sine-wave speech (SWS) is a highly simplified version of speech consisting only of frequency- and amplitude-modulated sinusoids representing the formants. That listeners can successfully understand SWS has led to claims that speech perception must be based on abstract properties of the stimuli far removed from their specific acoustic form. Here it is shown, in bilingual Cantonese/English listeners, that performance with Cantonese SWS is improved by noise-vocoding, with no effect on English utterances. This manipulation preserves the abstract informational structure in the signals, but changes its surface form. The differential effects of noise-vocoding likely arise from the fact that Cantonese is a tonal language, hence more reliant on fundamental frequency (F0) contours for its intelligibility. SWS does not preserve tonal information from the original speech, but does have false tonal information signalled by the lowest frequency sinusoid. Noise vocoding SWS appears to minimise the tonal percept, which thus interferes less in the perception of Cantonese. It has no effect in English which is minimally reliant on F0 variations for intelligibility. Therefore, it is not only the informational structure of a sound that is important, but also on how its acoustic detail interacts with the phonological structure of a given language.

PACS numbers: 43.71.An, 43.71.Hw

I. INTRODUCTION

There is long-running interest in the degree to which listeners can understand speech which has undergone unusual degradations. One of the most prominent of these transformations is known as *sine-wave speech* (SWS). In SWS, the complexity of speech is reduced to a few sinusoids (typically 3-4). The frequencies and amplitudes of these sinusoids track the formants in the speech signal, the regions of increased energy in the spectrum arising from vocal tract resonances (Remez, Rubin, Pisoni, & Carrell, 1981). Although leading to a percept that is notably weird (often described as bird-like), SWS can be highly intelligible, at least when listeners expect to hear speech.

SWS has played a large role in theorising about the nature of the speech perception process, most notably in the work of Remez and his colleagues. Perhaps the crucial (and generally accepted) claim is that when listening to SWS '… a perceiver must attend to the dynamic spectrotemporal variation of an unnatural carrier to identify phonetic and superordinate linguistic properties' (Remez et al., 2011, p. 969). Whilst reinforcing the indisputable notion that it is dynamic spectral variation that is the *sina qua non* of speech intelligibility (Rosen & Iverson, 2007), there is an underlying notion that the detailed acoustic form of the carrier is irrelevant as 'none of the familiar acoustic products of vocal sound is present in the spectrum' of SWS (Remez et al., 2011, p. 969). A more general claim in a recent review states that , '… direct measures of the perceptual organization of speech … reveal the action of a perceptual function that is … indifferent to the detailed grain of sensation …' (Remez & Thomas, 2013, p. 214). In short, processed signals which differ in surface acoustic structure should be equally intelligible as long as the informational structure is preserved.

We attempted to address this issue with reference to SWS in a tone language, that is, a language in which different fundamental frequency contours (referred to as *tones*) signal different words. Consider Hong Kong Cantonese, in which there are six distinct lexical tones. The syllable [jau], for example, can take all six tones, and thus has six possible meanings, dependent upon the tone used: 'worry', 'paint', 'thin', 'oil', 'friend' and 'again' (Yip, 2002). SWS, however, retains no features conveying the original fundamental frequency, and the formants, reflecting the movement of the articulators, would be more-or-less identical in these 6 words. There would, therefore, be little or no acoustic differences between their sine-wave versions. Although there can be minor differences in duration and amplitude across tones, these cues are unreliable (Kuo, Rosen, & Faulkner, 2008).

It would thus seem that tonal information would simply be unavailable in Cantonese SWS, but the situation is, in fact, worse. SWS is perceived to have a distinct intonation that has been convincingly shown to be cued by the tone representing the first formant (Remez & Rubin, 1993; Remez & Rubin, 1984). Therefore, syllables with the same tone will be perceived to have a different fundamental frequency (F0) contour depending upon their first formant trajectories. It may well be that this trajectory matches one of the genuine Cantonese tones (typically providing a misleading cue to word identity), or does not map readily onto a tonal category (confusing the listener).

There have been few studies of SWS in a tone language. Feng, Xu, Zhou, Yang, & Yin (2012) tested speakers of Mandarin Chinese (which has 4 tones) with

SWS materials on tone identification (for 10 different monosyllables) and perception of simple everyday sentences. Performance was barely above chance for tone identification, but very high for sentences, leading Feng et al. to conclude that tones played very little role in sentence recognition. However, as is well known, no particular cue alone is essential for speech perception. There is a great deal of redundancy in speech, especially in simple sentences where sentence context helps to resolve lexical ambiguity. For example, the fact that speech remains highly intelligible after bandpass filtering between 1-4 kHz does not imply that information outside these frequency regions is irrelevant. In the same way, successful sentence recognition when tones are not available (or when the perceived tone presents wrong or confusing information), does not mean they would not be used in ordinary circumstances.

Here, we compared the perception of ordinary Cantonese SWS sentences to SWS that had been processed through a 33-band noise vocoder (Shannon, Zeng, Kamath, Wygonski, & Ekelid, 1995). When applied to SWS, noise vocoding results in 3 or 4 bands of noise varying in frequency and amplitude in the same way as the original sinusoids (Rosen, Wise, Chadha, Conway, & Scott, 2011). Figure 1 shows spectrograms and spectral sections of both versions of SWS, as well as natural speech. See the supplementary materials for audio examples.

[Figure 1 about here]

Noise-vocoding is, in fact, another popular way to degrade speech, often used to simulate the processing exacted by a cochlear implant. Our aim here was quite different, in trying to minimise the perception of any fundamental frequency contour. Note too that although noise vocoding has a big effect on the informational content of ordinary speech, especially as it is ordinarily done with a relatively small number of channels (for example, in smearing spectral detail and eliminating harmonicity), noise vocoding SWS with so many channels loses little or nothing of the informational content of the signal.

Why might noise-vocoding eliminate the tonal information in SWS? For one thing, noise-vocoding eliminates the clear pitch of the component sinusoids in the SWS complex, minimising any percept of a voice pitch contour. Furthermore, noise-vocoding causes the separate 'formants' to cohere together more firmly, making it extremely difficult to 'hear out' and isolate one of them, a process which appears to be crucial in hearing an intonation pattern in SWS (Remez & Rubin, 1993; Remez & Rubin, 1984). Therefore, if noise-vocoding reduces or eliminates the perception of a (misleading) voice pitch contour, performance in a tone language should be better for noise-vocoded SWS (NzVoc-SWS) than for ordinary SWS. Whilst acknowledging the role that intonation can play in intelligibility for even a non-tonal language like English (e.g., Laures & Weismer, 1999), we would expect English materials to be less affected by any percept of a false fundamental frequency contour.

Therefore, we tested bilingual Cantonese/English speakers in both Cantonese and English, comparing performance for SWS and NzVoc-SWS in each language. Insofar as noise vocoding minimised the percept of a fundamental frequency contour, we would expect performance in Cantonese to increase with noise vocoding but for performance in English not to change. If speech perception is based only on the abstract acoustic structure in a signal, as Remez and colleagues would have it, we expect no effects of noise vocoding in either language.

II. METHODS

A. Speech material

All speech materials were drawn from digital anechoic recordings at 22.05 kHz of the BKB sentence lists (Bench & Bamford, 1979), a corpus of simple everyday sentences scored on the basis of identified key words (e.g., *The **clown** had a **funny face***). Only sentences with 3 key words were used. The English sentences comprised four lists of 14 sentences recorded by a male speaker of standard Southern British English. Four different lists were translated into natural spoken Cantonese with matching keywords and then recorded by a 25 year old male Cantonese speaker, originally from Hong Kong.

B. Stimulus processing

***1. Creation of sine-wave speech (SWS)**:* Semi-automatic procedures (with extensive hand-checking and correcting) were used to track the frequencies and amplitudes of up to three formants every 10 ms. The English materials were created as part of earlier research, using special scripts implemented in a package for speech processing that become obsolete prior to this study (ESPS). Therefore, a different software package was used for Cantonese (WinSnoori - Laprie, 2007). Two different people oversaw the determination of the formant tracks, each a native speaker of the language concerned, with considerable training in phonetics and speech sciences. From these values, sine-wave versions of the original speech were constructed by synthesizing up to three independent sinusoids whose frequency and amplitude matched those of the original formants (Remez et al., 1981), using the same software for both languages.

2. Creation of noise-vocoded sine-wave speech (NzVoc-SWS): SWS sentences were passed through a noise-excited vocoder using essentially the technique described by Shannon et al. (1995). Envelope detection used full-wave rectification and low-pass filtering at 32 Hz after spectral analysis by a bank of 33 analysis filters over the frequency range from 70 Hz – 10.0 kHz, spaced using Greenwood's equation (Greenwood, 1990). As Figure 1 shows, SWS consists of three spectral components at most, each very narrow in the spectrum. Noise-vocoding broadens these, and also removes the quasi-periodicity in each of the SWS components. Both versions of SWS are very different in character from the natural speech, although NzVoc-SWS has features more reminiscent of the formants in natural speech than the sinusoids of SWS.

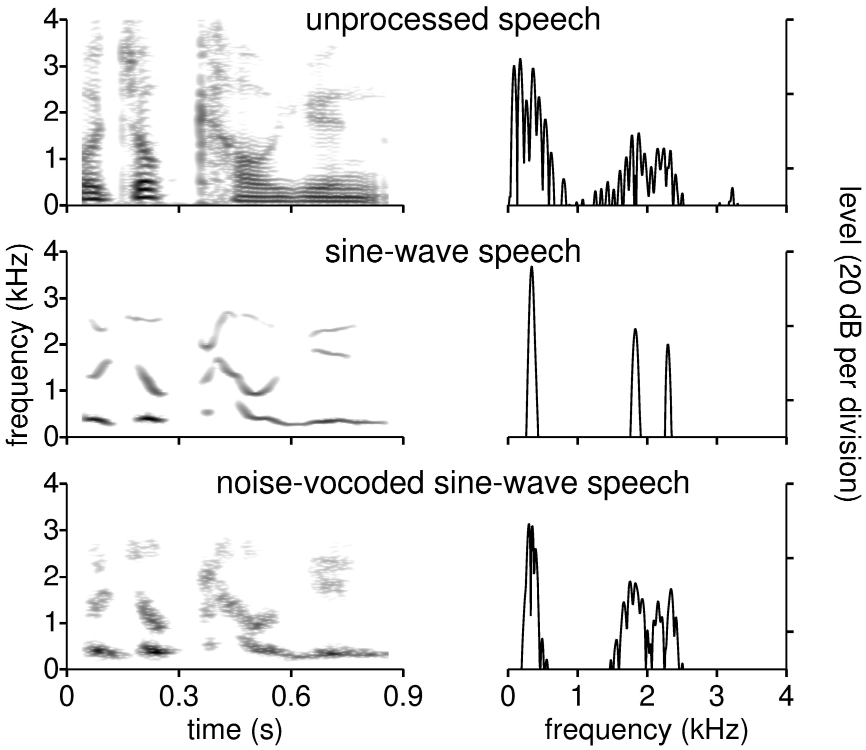

Figure 1: Narrowband Spectrograms (at left) and spectral cross sections (at right) of unprocessed Cantonese speech, and the two versions of sine-wave speech used in this study. The spectral slices are taken at the time value of 0.69 s. This figure depicts the first 900 ms of the sentence 個足球員 唔見咗一隻鞋 *meaning* ***The footballer lost a boot****, with the excerpt transcribed as* [kɔː3 tsʊk1 k^{h}ɐu4 jyːn4]. *Note in the spectral sections how the multiplicity of harmonics varying in overall height in the unprocessed speech has been simplified down to three sinusoidal components. Noise-vocoding the sine-wave speech broadens these components so that they have a spectral shape more similar to speech, but still lacking harmonics.*

C. Participants

A total of 29 Cantonese/English bilinguals took part. All reported having normal hearing, and had not participated previously in any study involving sine-wave or noise-vocoded speech. The listeners fell into two groups, depending upon whether Cantonese or English was their dominant language (hereinafter referred to as L1).

The Cantonese-dominant group contained seventeen native Cantonese speakers from Hong Kong who learnt English as a second language in school. They were aged between 18 and 22 years old, except for one listener aged 60. The English-dominant group contained twelve listeners who lived in London, and were aged between 20 and 27 years old, save for one aged 40. Ten of these were born in England, learnt Cantonese at home (or at Chinese schools) and use English as their main language. The other two, although not born in the UK, had lived in London for at least 3 years, and had used English as their main language at school.

Procedure

All tests were blocked by language, with the order of the languages tested counter-balanced across listeners. Within each block, SWS and NzVoc-SWS sentences were presented alternately, again with the order counter-balanced across listeners. The order of the sentences presented was fixed for all listeners, thus controlling for variations in intelligibility from sentence to sentence. Note that because the talkers were different for the two languages, as were the sentences, and because the SWS was constructed slightly differently, we cannot say anything about the relative intelligibility of the materials in the two different languages. As our primary focus is on performance *differences* arising from the interaction of noise vocoding and language type (tonal vs.non-tonal), this does not matter here.

All sentences were presented over headphones, under computer control, at a fixed and comfortable level. A practice session of eight items in the appropriate language preceded each test block. Here, for each item, a SWS or NzVoc-SWS sentence was presented, and the listener asked to repeat back what was said. No matter the response, the original unprocessed version of the sentence was played, followed by the processed and unprocessed versions again.

Listeners were then presented 40 sentences in the language under test (20 in each processing condition), with encouragement to guess even if they were not sure what was said. Each test sentence was played only once, and no feedback was given. Scoring was on the basis of a ‘loose’ key word method, meaning that only the root of the response word need match, without regard for tense, person or number. Audio recordings were made of all the listeners’ responses for checking although responses were also scored during the test.

III. RESULTS

The proportion of key words correct per language and processing condition was calculated for each listener, and is shown for each group in the boxplots of Figure 2. Inspection of the data revealed no evidence that the two older listeners performed differently than the younger ones, nor that the two English dominant bilinguals born outside London performed differently to those born in London.

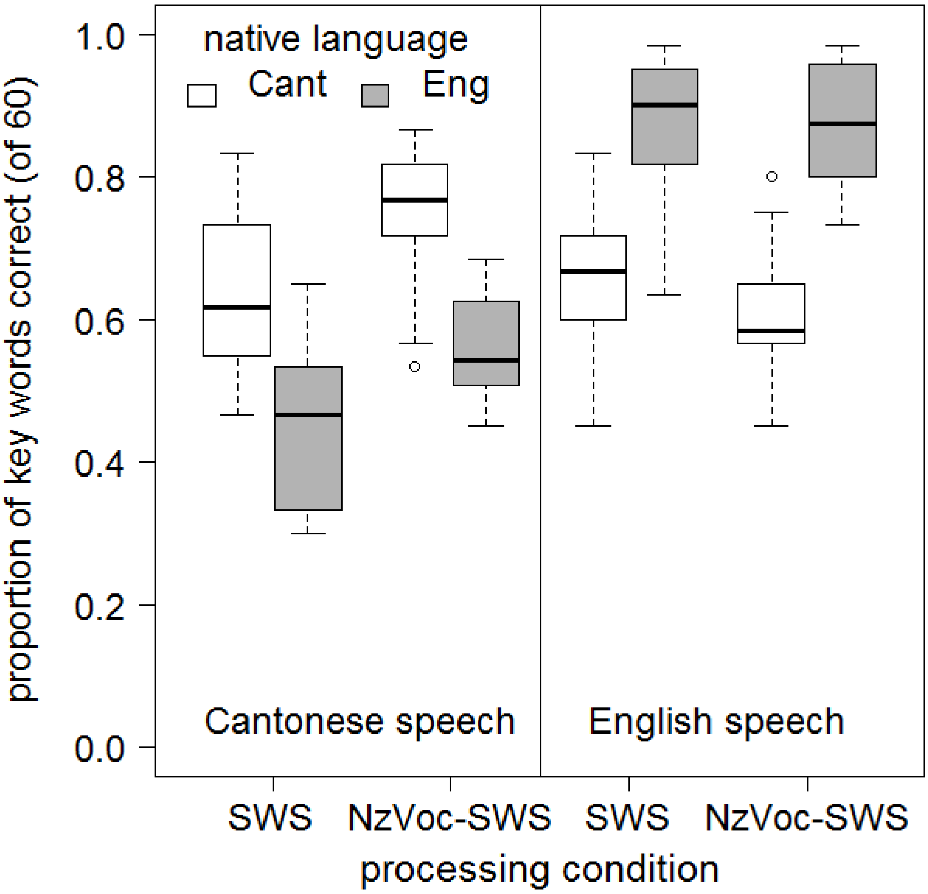

Figure 2: Boxplots of the individual scores for each processing condition and language, shown separately for the Cantonese-dominant and English-dominant participants.

Because the data are in the form of proportions, a set of mixed-effects logistic regression models was applied using the lme4 package of R (Bates, Maechler, Bolker, & Walker, 2014). Starting from a saturated model with *listener* as a random intercept, and predictors of test language (English vs. Cantonese), processing condition (SWS vs. NzVoc-SWS) and first language (L1), predictors were excised if they did not significantly improve the fit of the model. A complex pattern of interactions was obtained. Neither the highest, third-order interaction was significant

[χ^2=1.7, df=1, p=0.18], nor did processing condition interact with L1 [χ^2=0.1, df=1, p=0.81]. Test language interacted with processing condition, and separately with the listener's native language [χ^2>34.1, df=1, p<0.001 in both cases]. The latter interaction of L1 x test language is easy to understand, and expected, insofar as Cantonese L1 listeners performed better than the English L1 listeners for both versions of Cantonese speech, and vice versa for the English speech. The advantage for the English L1 listeners was about 25 percentage points for English, and that for the Cantonese L1 listeners for Cantonese about 18 percentage points

The interaction of test language with processing condition is the key effect regarding our main hypothesis, and arises because performance for all listeners appears to be unaffected by processing condition for the English test material, but is better for NzVoc-SWS than for SWS in Cantonese. A clearer way to display this difference is to calculate, for each listener, the advantage in performance for the noise vocoded speech by subtracting the score for SWS speech from the score for NzVoc-SWS separately for each test language (Figure 3). It can be clearly seen that Cantonese speech benefits more from being noise-vocoded than English speech does (by about 8 percentage points). Note that the relatively high scores obtained by the English L1 listeners for both versions of English speech may have constrained finding any differences between these conditions (medians of ≈0.9). On the other hand, 7 of the 12 listeners had a score for SWS English speech of ≤0.9 so had some leeway for improvement.

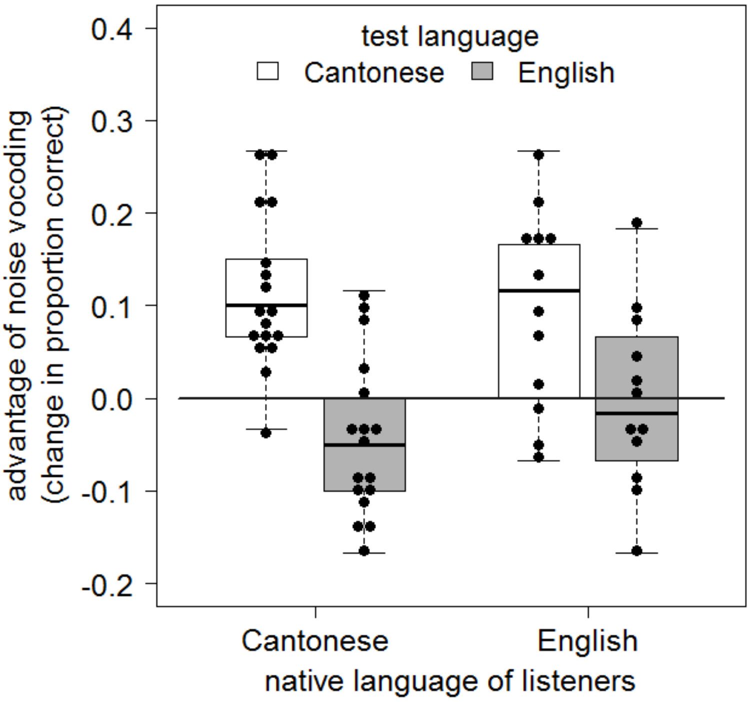

Figure 3: Boxplots of the improvement in performance for noise-vocoded SWS over SWS shown separately for each test language and group of listeners. The black circles show individual results.

IV. DISCUSSION

The results obtained are easily summarised. Least surprisingly, bilingual listeners dominant in Cantonese performed better with sentence materials in Cantonese than listeners dominant in English, and vice versa. More pertinent to our interests here, for both groups of listeners, performance with Cantonese SWS was improved by noise vocoding, but this extra transformation had no effect on the intelligibility of English SWS. This difference almost certainly arises from the fact that Cantonese is a tonal language, in which variations in fundamental frequency play an important role in speech intelligibility.

There is of course, still the question of exactly how noise-vocoding exerts its effects. Our explanation involves the assumption that noise vocoding eliminates the perception of the voice pitch contour, but we did not investigate this claim directly. However, there are at least two factors that support this notion. Presumably, the loss of periodicity within the lowest sinusoidal component contributes to a weaker pitch percept. Also, the spectral components cohere much better when noise-vocoded, at least subjectively. It would not be surprising that this would make it harder to 'hear out' the lowest component as a distinct perceptual feature signalling a melodic contour, especially with the noisier representation effected by noise vocoding.

There are at least two implications of our findings, one practical and one more theoretical. From a practical point of view, these results stand as a warning to researchers that it cannot be assumed that a particular degradation or transformation of speech will have the same impact in different languages. Although the essential aspects of the speech signal must be similar (as constrained by the human vocal apparatus and auditory pathway), languages differ greatly in the extent to which different features are exploited for intelligibility. Therefore, the special characteristics of each language must be considered in the light of any particular degradation.

From a more theoretical standpoint, these results show that it is not sufficient to talk only about the abstract informational structure of a signal without regard to its detailed acoustic form, because sounds have to be transduced and processed by the auditory system, which may affect the information perceived in the signal. The SWS and NzVoc-SWS contained more-or-less identical information about spectro-temporal dynamics (as confirmed by the fact that performance in English was not affected by the extra transformation), yet led to different performance in Cantonese.

Clearly ‘the detailed grain of sensation’ can be an important factor in speech perception.

Acknowledgements

We thank: Errol Hui for his beautiful voice, put to good use recording the sentences in Cantonese; Paul Iverson for the techniques used to develop the English sine wave speech and Clare Norman who did most of those measurements; Philip Rubin, Steve Frost, and Dan Ellis who wrote the SineWave Synthesis package from Haskins Laboratories, New Haven, CT, USA, parts of which were used to generate sine-wave speech; all the listeners for their willing participation. Tim Green, Paul Iverson and Katharine Mair commented helpfully on earlier versions of the manuscript, as did an anonymous reviewer and James Hillenbrand on a later one. Dawna Lewis provided her stimuli for interesting listening, and Kevin Tang transcribed the Cantonese sentence in Figure 1. Most of the work reported here was performed as part of the UCL Speech Sciences BSc thesis of Sze Ngar Catherine Hui.

References

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1-7 [On-line]. Available: http://cran.r-project.org/package=lme4 (Last viewed 16/09/2015).

Bench, J. & Bamford, J. (1979). *Speech-hearing Tests and the Spoken Language of Hearing-impaired Children*. (Academic Press, London), pp. 1-543.

Feng, Y. M., Xu, L., Zhou, N., Yang, G., & Yin, S. K. (2012). Sine-wave speech recognition in a tonal language. *Journal of the Acoustical Society of America, 131,* EL133-EL138.

Greenwood, D. D. (1990). A cochlear frequency-position function for several species - 29 years later. *Journal of the Acoustical Society of America, 87,* 2592-2605.

Kuo, Y. C., Rosen, S., & Faulkner, A. (2008). Acoustic cues to tonal contrasts in Mandarin: Implications for cochlear implants. *Journal of the Acoustical Society of America, 123,* 2815-2824.

Laprie, Y. (2007). WinSnoori – Free software for speech analysis. http://www.loria.fr/~laprie/WinSnoori/ (Last viewed 23/03/2007).

Laures, J. S. & Weismer, G. (1999). The effects of a flattened fundamental frequency on intelligibility at the sentence level. *Journal of Speech Language and Hearing Research, 42,* 1148-1156.

Remez, R. E., Dubowski, K. R., Broder, R. S., Davids, M. L., Grossman, Y. S., Moskalenko, M. et al. (2011). Auditory-Phonetic Projection and Lexical Structure in the Recognition of Sine-Wave Words. *Journal of Experimental Psychology-Human Perception and Performance, 37,* 968-977.

Remez, R. E. & Rubin, P. E. (1984). On the Perception of Intonation from Sinusoidal Sentences. *Perception & Psychophysics, 35,* 429-440.

Remez, R. E. & Rubin, P. E. (1993). On the Intonation of Sinusoidal Sentences - Contour and Pitch Height. *Journal of the Acoustical Society of America, 94,* 1983-1988.

Remez, R. E. & Thomas, E. F. (2013). Early recognition of speech. *Wiley Interdisciplinary Reviews-Cognitive Science, 4,* 213-223.

Remez, R., Rubin, P. E., Pisoni, D. B., & Carrell, T. D. (1981). Speech perception without traditional speech cues. *Science, 212:* 947-950.

Rosen, S. & Iverson, P. (2007). Constructing adequate non-speech analogues: what is special about speech anyway? *Developmental Science, 10,* 165-168.

Rosen, S., Wise, R. J. S., Chadha, S., Conway, E. J., & Scott, S. K. (2011). Hemispheric Asymmetries in Speech Perception: Sense, Nonsense and Modulations. *Plos One, 6(9),* e24672. doi:10.1371/journal.pone.0024672.

Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science, 270,* 303-304.

Yip, M. (2002). *Tone*. (Cambridge University Press, Cambridge, UK), pp. 1-376.