INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Temporal Envelopes in Sine-Wave Speech Recognition

Li Xu

Communication Sciences and Disorders, Ohio University, USA
xul@ohio.edu

Abstract

There is a long debate on the relative importance of spectral and temporal cues in speech perception theories. On the one hand, the highly-intelligible sine-wave speech (SWS) has been viewed as a representation of the global spectral structure of the speech signal. On the other hand, there is accumulating evidence showing that the temporal aspects of speech without spectral details provide sufficient speech understanding. The present study explored whether the temporal envelopes imbedded in the SWS contribute to its intelligibility. In the experiments, both SWS and natural speech signals were processed with noise and tone vocoders to remove the spectral details but to preserve the temporal envelopes. Twenty-two normal-hearing, native English-speaking adult listeners participated in sentence recognition tasks. Speech recognition performance of vocoder-processed SWS was slightly inferior to that of vocoder-processed natural speech but both reached plateau performance at 6-8 channels. Acoustic analysis further indicated that the temporal envelopes of the SWS were almost identical to those of the natural speech, with a mean correlation coefficient r = 0.949 across all sentences. The results provide strong evidence that the SWS represents both spectral and temporal structures of the speech and that the temporal envelopes imbedded in SWS carry important information for speech recognition.

Index Terms: speech recognition, vocoder, temporal envelope, sine-wave speech

1. Introduction

Despite the enormous progress that has occurred in perceptual, neurophysiological and imaging research on speech in the last three decades, our understanding of the mechanisms that underlie both the intelligibility of speech and the perceptual organization of speech are still fragmentary. In particular, a coherent account of intelligibility of both sine-wave speech (SWS) and vocoded speech is not yet available. SWS consists of a number of sinusoids, typically three that track the frequencies of the first three formants of speech [1]. It lacks broadband formant structure, fundamental frequency (F0), or any clear distinction between periodic and aperiodic excitation that are traditionally believed to be important for speech perception. And yet, normal-hearing listeners can understand SWS from 60% to nearly 100% depending on the sentence materials [2-4].

The nature of the phonetic information represented in SWS is not clear [5]. The cues for the perception of SWS have been thought to be predominately spectral and have been given labels such as "global spectral structure" or "spectral skeleton" [3, 6, 7]. On the other hand, a recent report has suggested that both SWS and vocoded SWS contain more-or-less identical information about spectrotemporal dynamics of speech [8]. Such an observation was based on a vocoder processing with 33 channels. When we examine the time waveforms and spectrograms of the natural speech and SWS and 4-channel noise- and tone-vocoder processed signals (Fig. 1), for example, the spectral dynamics are very different among the original speech signal (either natural or SWS) and the noise- and tone-vocoder processed signals (column-wise comparisons in Fig. 1). What is striking is that the temporal features of the SWS signals (either original or vocoder processed) bear noteworthy resemblance to those of the natural speech (row-wise comparisons in Fig. 1). Therefore, it is important to quantify the amount of temporal envelope information in SWS.

From the first "talking machine" (aka vocoder) invented by Dudley [9] to the multichannel cochlear implants [10], there is mounting evidence that speech information can be conveyed in the temporal envelopes of the speech signal [11]. Certain features of consonants are well preserved within a single band (most strongly, manner of articulation) but others are very weak (place of articulation) [12, 13]. The spectral resolution required for speech recognition has been studied using a spectral smearing technique [14, 15]. It has been found that a reduced spectral resolution has a minimal effect on speech recognition in quiet, even for smearing that simulates auditory filters six times broader than in normal hearing [16, 17]. These results suggest that approximately five bands of spectral information would be sufficient for speech recognition in quiet. Using the vocoder technique, research has shown that normal-hearing listeners achieve good speech perception with the temporal envelopes carried by 4–8 sinusoids or narrow bands of noise depending on the speech materials [11, 18-23].

The aims of this study are to determine the amount of temporal envelope information in SWS and to explore whether the temporal envelope information imbedded in the SWS contributes to its intelligibility. The natural speech and the SWS signals were subject to vocoder processing with 1 to 8 channels and the processed sentences were presented to normal-hearing listeners for recognition. Acoustical analysis was performed to determine the similarities in temporal envelope between natural speech and SWS.

2. Method

2.1. Subjects

A total of 22 normal-hearing adults (15 females and 7 males) were recruited from the Ohio University student population. They were between 18 and 33 years of age. Normal-hearing status was determined based on pure-tone hearing threshold < 20 dB HL at octave frequencies between 250 and 8,000 Hz. Subjects were native speakers of English and had no prior experience with the City University of New York (CUNY) sentences [24], vocoder-processed speech, or SWS.

http://dx.doi.org/10.21437/Interspeech.2016-171

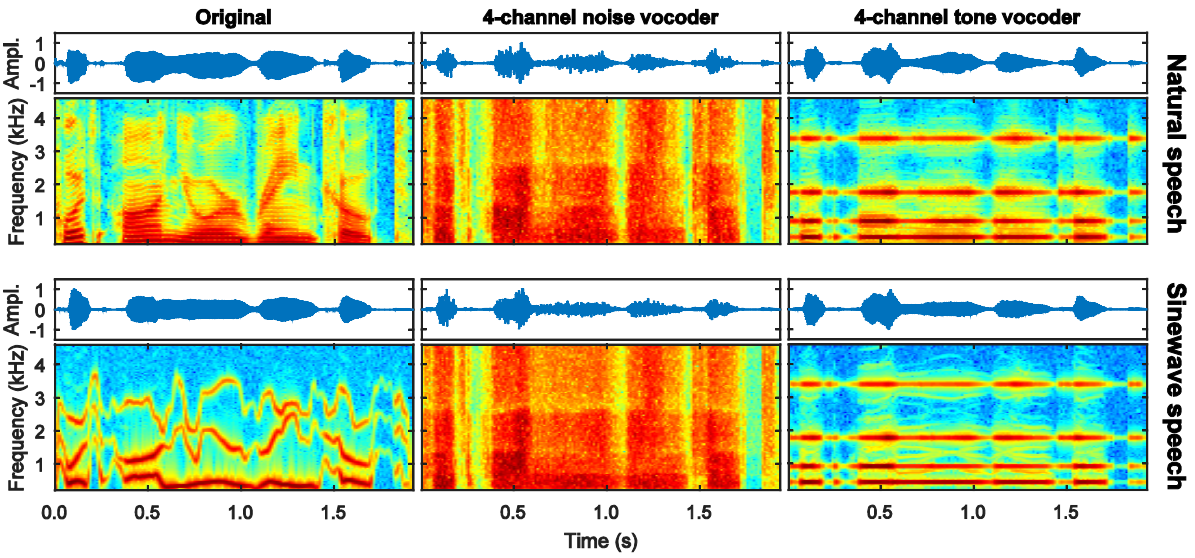

Figure 1: *Examples of waveforms and spectrograms of natural speech (upper left), sine-wave speech (lower leftt), 4-channel noise-vocoder processed natural speech (upper middle), 4-channel noise-vocoder processed sine-wave speech (lower middle), 4-channel tone-vocoder processed natural speech (upper right), and 4-channel tone-vocoder processed sine-wave speech (lower right). Spectrograms are in narrowband format with colors showing energy associated with particular time and frequency. The sentence is CUNY Sentence #4 of List #6, "Put on your raincoat," spoken by a male talker.*

2.2. Stimuli

Speech materials used in the experiments were CUNY sentences recorded at a 44.1-kHz sampling rate from a male talker whose mean voice F0 was 133 Hz [24]. The SWS stimuli were generated using a PRAAT routine [25] that extracted the center frequency of the formants of the original CUNY sentences based on Linear Predictor Coefficient (LPC) analysis. The extracted formants were then replaced with sinusoidal replicas. The amplitude and frequency of the center frequencies were measured every 10 ms of the speech signal. When there were no formants in the original speech, for example, prior to the release of a voiceless plosive or affricate, a simple linear interpolation was used to create the sinusoids in SWS. The sine-wave sentences as well as the original sentence materials were subjected to the vocoder signal processing as described below.

The noise-excited vocoder [11, 19, 21, 26] or the tone-excited vocoder [27] was realized using custom MATLAB (MathWorks, Natick, MA) software. The speech signals within the frequency range of 250 to 4,600 Hz were first passed through a bank of analysis filters (third-order elliptic bandpass filters) that varied from one to eight (1, 2, 3, 4, 6, and 8) in the number of channels. The Greenwood formula [28] was used to divide the frequency bands based on equal distance on the basilar membrane of the cochlea. The temporal envelope was extracted from each analysis band by half-wave rectification and low-pass filtering (second-order Butterworth filter). The cutoff frequency of the low-pass filter used in the noise-excited vocoder was set at 160 Hz whereas that used in the tone-excited vocoder was set at 40 Hz in order to reduce the amount of sidebands produced by amplitude modulation. In the noise-excited vocoder, a white noise that had passed through each of the same analysis filters was modulated by the temporal envelope of the corresponding band. In the tone-excited vocoder, a sinusoid at the center frequency of each band was modulated by the temporal envelope of the corresponding band. Lastly, the modulated noise bands or the sinusoids were summed and stored in the computer for later presentations.

2.3. Procedures

The speech signals were presented binaurally through a supra-aural headphone (Sennheiser, HD 265) in a sound attenuating booth (IAC, New York). Subjects were able to adjust the intensity of the signal to the most comfortable level of their choice. A custom MATLAB program was used to present the acoustic stimuli and to record subjects' responses in the graphical user interface. Subjects were instructed to type the sentence they had heard in a text box on the computer screen. Subjects were allowed to listen to each sentence as many times as they felt necessary to best understand the sentence. On average, a stimulus was repeated 4~5 times.

The 22 subjects were randomly divided into two groups (11 each). One group participated in the noise vocoder experiment and the other tone vocoder experiment. For both experiments, there were six channel conditions (1, 2, 3, 4, 6, and 8 channels) for the two types of speech materials (vocoder-processed natural speech and vocoder-processed SWS). Each condition was assigned randomly with one CUNY sentence list that consisted of 12 sentences so that no sentence was used more than once. In total, 144 typed responses were collected from each subject (12 sentences × 6 channel conditions × 2 speech types). All subjects were trained with the vocoder-processed natural speech and SWS to familiarize themselves with the test materials. Note that the training used different CUNY sentence lists from the ones used in the test. The training session used 18 vocoder-processed natural speech sentences and 18 vocoder-processed SWS sentences (6 sentences for each of the following channel conditions: 2, 4, and 8 channels). Feedback was provided for training but not for the test. The order of the conditions was fully randomized to reduce any order effects. Subjects took approximately two hours to complete the experiment.

2.4. Acoustic analysis of the temporal envelope

Acoustic analysis based on the correlational approach [29] was performed to examine the correlation between temporal envelopes of the naturally spoken sentences and sine-wave replicas of the sentences. For each of the 144 sentences used in the present study, the correlation coefficients across all channels in all channel conditions were averaged to represent the correlation of that sentence. We also computed the averaged correlation coefficients between the natural speech and the SWS when the sentences were different sentences. The longer sentence of the two was truncated to match the length of the shorter one. There were 10,296 possible permutations of all the 144 sentences.

3. Results

3.1. Sentence recognition of noise- and tone-vocoder processed SWS

Figure 2 (upper left) shows sentence recognition scores for noise-vocoded natural speech and SWS materials. The scores improved as a function of number of channels (i.e., noise bands). Individual variation under both conditions was small as indicated by the standard deviations. Under the conditions of 2, 3, 4, and 6 channels, the mean scores differed by 11.2 percentage points on average (pairwise comparisons after Bonferroni correction, $p < 0.0083$). A logistic regression as in [17] was used to fit the noise-vocoded natural speech recognition data (r^2 =0.994) and the noise-vocoded SWS recognition data ($r^2 = 0.990$) (Fig. 2, lower left). The two sets of functions were close with those of the noise-vocoded SWS data being shifted to the right slightly. On the logistic regression curves for the group mean data, the number of channels required to achieve 50% correct recognition was 2.0 and 2.6 for noise-vocoded natural speech and noise-vocoded SWS, respectively and the difference was statistically significant (paired t test, $t(10) = -5.64$, $p = 0.0002$). The regression slopes at 20-80% correct recognition were 40.4 and 26.3 percentage points/channel, respectively and the difference was statistically significant (paired t test, $t(10) = 5.47$, $p = 0.0003$).

In order to eliminate the potential effects of the inherent fluctuations in noise on the temporal envelope [30], we used tone vocoder processing (Fig. 1, right panels) in this part of the experiment. Here, the three time-varying sinusoids in the SWS were replaced by a number of constant-frequency sinusoids. As in the noise vocoder, the temporal envelopes of the constant-frequency sinusoids were derived from the frequency bands in the original signals with the center frequencies equal to the sinusoids. Sentence recognition data were obtained with both tone-vocoded natural speech and tone-vocoded SWS from a different group of 11 normal-hearing, native English-speaking subjects. Figure 2 (upper right) shows the scores for tone-vocoded sentence recognition. Under the conditions of 2, 3, 4, and 6 channels, the mean scores differed by 19.2 percentage points on average (pairwise comparisons after Bonferroni correction, all $p < 0.05$). Again, a logistic regression was used to fit the tone-vocoded natural speech data (r^2 =0.978) and the tone-vocoded SWS data ($r^2 = 0.994$) (Fig. 2, lower right). For the group mean data, the number of channels required to

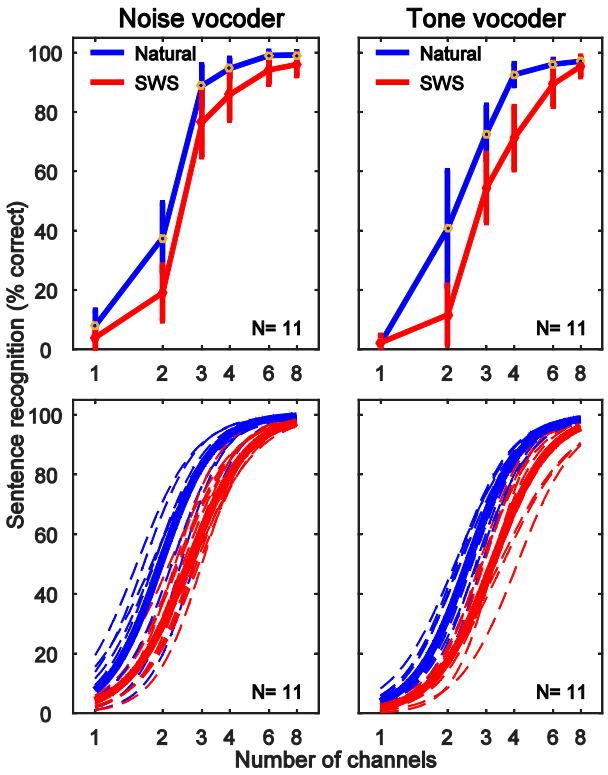

Figure 2: *Upper panel: Mean speech recognition performance of noise vocoder processed speech as a function of number of channels. The natural speech and vocoder-processed sine-wave speech (SWS) are represented by blue and red line, respectively. The error bars represent SDs. Lower panel: Logistic regression of the individual recognition data (dashed lines) and the group mean recognition data (thick solid lines).*

achieve 50% correct recognition was 2.5 and 3.2 for tone-vocoded natural speech and tone-vocoded SWS, respectively and the difference was statistically significant (paired t test, $t(10) = -7.33$, $p = 0.00003$). The regression slopes at 20-80% correct recognition were 30.8 and 22.4 percentage points/channel, respectively and the difference was also statistically significant (paired t test, $t(10) = 5.14$, $p = 0.0004$).

The speech recognition performance as a function of number of channels with the tone vocoder (Fig. 2 right) showed a similar trend to that with the noise vocoder (Fig. 2, left). However, the performance with the tone vocoder was lower than that with the noise vocoder for both natural speech and SWS. The differences in number of channels to reach 50% correct and the differences in the regression slopes at 20-80% correct between the noise and tone vocoders for either natural speech or SWS were all statistically significant (t test, all $p < 0.05$).

3.2. Acoustic analysis of the temporal envelope in natural and SWS signals

As shown in Fig. 1 (middle and left panels), the waveforms and spectrograms of vocoder-processed natural speech and SWS appear to resemble each other. Figure 3a shows the temporal envelopes (lowpass filtered at 40 Hz) of the same sentence that

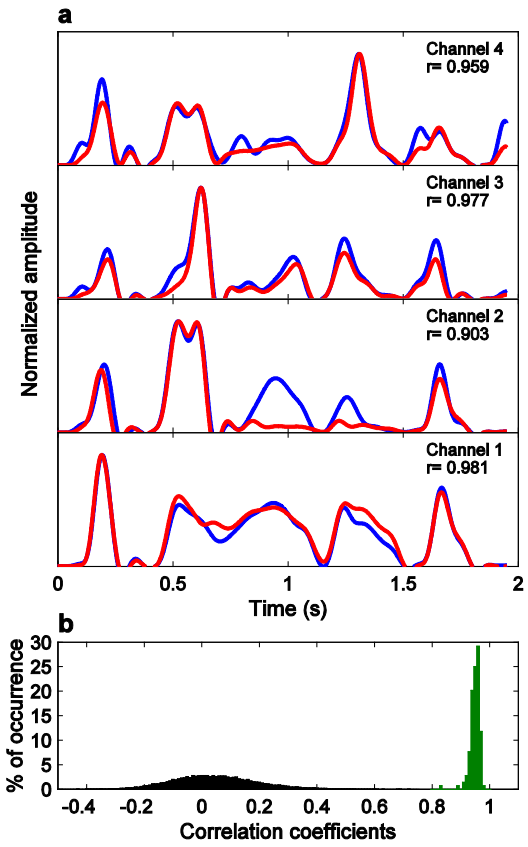

Figure 3: *Similarity of temporal envelope of the natural speech and the sine-wave replica. (a). Temporal envelopes of the example sentence, "Put on your raincoat", extracted from a 4-channel vocoder. Channels 1 through 4 were the frequency bands from low to high. The normalized temporal envelopes of the natural speech and SWS are represented by blue and red lines, respectively. The envelopes were extracted by a half-wave rectification and a lowpass filtering at 40 Hz. The correlation coefficients (r) between the two envelopes were then computed. The values were shown in the upper right corners. (b). Distribution of the correlation coefficients from the correlation analysis. The histogram plotted in green shows the distribution of the correlation coefficients between the natural speech and the SWS when the sentence was the same sentence (N= 144). The histogram plotted in black shows the distribution of the correlation coefficients between the natural speech and the SWS when the sentences were different sentences (N= 10,296).*

was plotted in Fig. 1, as an example. In each of the four channels, the envelopes of the two types of speech (i.e., natural and SWS) were highly correlated. Averaged correlation coefficients (r) across all channels represented the resemblance of the temporal envelopes. In this example, the averaged r value was 0.955. The mean r value across all the 144 CUNY sentences that were used in the present study under different channel conditions was 0.949 (SD = 0.017). On the other hand, when the correlation was computed for natural and SWS stimuli that were of two different sentences (N= 10,296), the overall mean r value reduced to 0.0472 (SD = 0.148) (Fig. 3b).

4. Discussion and Summary

SWS and vocoder speech adopt distinct signal processing (Fig. 1) and yet both are highly intelligible. Contemporary speech perception theories have not attempted to reconcile the different mechanisms of speech perception of the two types of speech. Some researchers have categorized the SWS and vocoder speech as containing global spectral structure and global amplitude structure of speech, respectively [3, 6, 7]. Results from the present study show that after the spectral details of the SWS are removed during vocoder processing, speech recognition using only the temporal envelopes of the sinusoids yields intelligibility slightly inferior to that of the vocoder-processed natural speech (Fig. 2). For both vocoder-processed natural speech and vocoder-processed SWS, sentence recognition reached plateau performance at 6 to 8 channels.

The two different types of vocoder (i.e., tone and noise vocoders) used in the present study produce a similar sentence-recognition performance between natural speech and SWS (Fig. 2). The differences in recognition performance around 3 or 4 channels between noise and tone vocoders are predictable because the tone vocoder in the present study was implemented with a lower envelope cutoff (40 Hz) whereas that of the noise vocoder was much higher (160 Hz). Such minor differences in sentence intelligibility between the two types of vocoder using different envelope cutoffs were reported by [30]. It is interesting to note though, that in the tone vocoder conditions, approximately 50% sentence recognition was achieved when the time-varied frequency contours of the 3 sinusoids of the SWS were replaced by 3 c onstant-frequency sinusoids. This result further confirms that the temporal envelope contained in the SWS can provide sentence intelligibility.

The acoustic analysis shows that the temporal envelopes of the SWS closely resemble the envelopes of the corresponding frequency bands of the natural speech (Fig. 3). The mean correlation coefficient between the envelopes of the natural speech and SWS of the CUNY sentences was approximately 0.95. Such results indicate that the temporal envelope contained in the SWS might contribute to its intelligibility. Thus, the present study provides evidence to support a unified explanation of the mechanisms underlying the perception of sine-wave and vocoder speech. That is, temporal envelope information conveyed in a small number of frequency channels provides good sentence recognition for either natural speech or SWS.

In summary, we have shown that SWS contains both spectral and temporal structures for speech recognition. The temporal structure of SWS closely resembles that of the natural speech. While it is conceivable that listeners use both spectral and temporal cues for speech recognition, further research will be needed to evaluate the perceptual weights of the two cues in SWS recognition.

5. Acknowledgements

The author is grateful to Natalie Bevilacqua, Heather Gradisek, Marisol Gliatas, Emily Hahn, Bethany Mendez, Alexa Patton, and Ning Zhou for their technical and editorial assistance.

5. References

[1] R. E. Remez, P.E. Rubin, D.B. Pisoni, and T.D. Carrell, "Speech perception without traditional speech cues," *Science*, vol. 212, pp. 947-949, 1981.

[2] Y. M. Feng, L. Xu, N. Zhou, G. Yang, and S. K. Yin, "Sine-wave speech recognition in a t onal language," *The Journal of the Acoustical Society of America*, vol. 131, pp. EL133-138, 2012.

[3] S. Nittrouer and J. H. Lowenstein, "Learning to perceptually organize speech signals in native fashion," *The Journal of the Acoustical Society of America*, vol. 127, pp. 1624-1635, 2010.

[4] R. E. Remez, P. E. Rubin, S. M. Berns, J. S. Pardo, and J. M. Lang, "The perceptual organization of speech," *Psychological Review*, vol. 101, pp. 129-156, 1994.

[5] J. M. Hillenbrand, M. J. Clark, and C. A. Baer, "Perception of sinewave vowels," *The Journal of the Acoustical Society of America*, vol. 129, pp. 3991-4000, 2011.

[6] S. Nittrouer, J. H. Lowenstein, and R. R. Packer, "Children discover the spectral skeletons in their native language before the amplitude envelopes," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 35, pp. 1245-1253, 2009.

[7] S. Nittrouer, J. Kuess, and J. H. Lowenstein, "Speech perception of sine-wave signals by children with cochlear implants," *The Journal of the Acoustical Society of America*, vol. 137, pp. 2811-2822, 2015.

[8] S. Rosen, and S. N. C. Hui, "Sine-wave and noise-vocoded sine-wave speech in a tone language: Acoustic details matter," *The Journal of the Acoustical Society of America*, vol. 138, pp. 3698-3702, 2015.

[9] H. Dudley, "The vocoder," *Bell Labs Rec*, vol. 18, pp. 122-126, 1939.

[10] G. Clark, *Cochlear Implants: Fundamentals and Applications*. New York: Springer, 2003.

[11] R. V. Shannon, F. G. Zeng, V. Kamath, J. Wygonski, and M. Ekelid, "Speech recognition with primarily temporal cues," *Science*, vol. 270, pp. 303-304, 1995.

[12] S. Rosen, "Temporal information in speech: acoustic, auditory and linguistic aspects," *Philosophical Transactions of the Royal Society B*, vol. 336, pp. 367-373, 1992.

[13] D. J. Van Tasell, S. D. Soli, V. M. Kirby, and G. P. Widin, "Speech waveform envelope cues for consonant recognition," *The Journal of the Acoustical Society of America*, vol. 82, pp. 1152-1161, 1987.

[14] A. Boothroyd, B. Mulhearn, J. Gong, and J. Ostroff "Effects of spectral smearing on phoneme and word recognition," *The Journal of the Acoustical Society of America*, vol. 100, pp. 1807-1818, 1996.

[15] E. Villchur, "Electronic models to simulate the effect of sensory distortions on speech perception by the deaf," *The Journal of the Acoustical Society of America*, vol. 62, pp. 665-674, 1977.

[16] T. Baer and B. C. Moore, "Effects of spectral smearing on the intelligibility of sentences in noise," *The Journal of the Acoustical Society of America*, vol. 94, pp. 1229-1241, 1993.

[17] T. Baer and B. C. Moore, "Effects of spectral smearing on the intelligibility of sentences in the presence of interfering speech," *The Journal of the Acoustical Society of America*, vol. 95, pp. 2277-2280, 1994.

[18] P. C. Loizou, M. Dorman, and Z. Tu, "The number of channels needed to understand speech," *The Journal of the Acoustical Society of America*, vol. 106, pp. 2097-2103, 1999.

[19] L. Xu, Y. Tsai, and B. E. Pfingst, "Features of stimulation affecting tonal-speech perception: Implications for cochlear prostheses," *The Journal of the Acoustical Society of America*, vol. 112, no. 1, pp. 247-258, 2002.

[20] R. V. Shannon, Q. J. Fu, and J. Galvin 3rd, "The number of spectral channels required for speech recognition depends on the difficulty of the listening situation,"*Acta Oto-laryngologica Supplementum*, pp. 50-54, 2004.

[21] L. Xu, C. S. Thompson, and B. E. Pfingst, "Relative contributions of spectral and temporal cues for phoneme recognition," *The Journal of the Acoustical Society of America*, vol. 117, pp. 3255-3267, 2005.

[22] F. G. Zeng, K. Nie, G. S. Stickney, Y. Y. Kong, M. Vongphoe, A. Bhargave, C. Wei, and K. Cao, "Speech recognition with amplitude and frequency modulations," *Proceedings of National Academy of Sciences*, vol. 102, pp. 2293-2298, 2005.

[23] B. J. Kim, S.-A. Chang, J. Yang, S. H. Oh, and L. Xu, "Relative contributions of spectral and temporal cues to Korean phoneme recognition," *PLoS ONE*, vol. 10, no. 7, pp. e0131807, 2015.

[24] A. Boothroyd, L. Hanin, and T. Hnath, "Sentence test of speech perception: Reliability, set equivalence and short term learning," in, Vol. New York, Speech and Hearing Sciences Research Center, City University of New York, 1985.

[25] C. Darwin, "Sine-wave speech produced from them automatically using a script for the Praat program," http://www.lifesci.sussex.ac.uk/home/Charles_Darwin/SWS (Last viewed March 15, 2016.)

[26] L. Xu, and B. E. Pfingst, "Spectral and temporal cues for speech recognition: implications for auditory prostheses," *Hearing Research*, vol. 242, pp. 132-140, 2008.

[27] M. F. Dorman, P. C. Loizou, and D. Rainey, "Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs," *The Journal of the Acoustical Society of America*, vol. 102, pp. 2403-2411, 1997.

[28] D. D. Greenwood, "A cochlear frequency-position function for several species--29 years later," *The Journal of the Acoustical Society of America*, vol. 87, pp. 2592-2605, 1990.

[29] F. G. Zeng, K. Nie, S. Liu, G. Stickney, E. Del Rio, Y. Y. Kong, and H. Chen, "The dichotomy in auditory perception between temporal envelope and fine structure cues," *The Journal of the Acoustical Society of America*, vol. 116, pp. 1351-1354, 2004.

[30] P. Souza, and S. Rosen, "Effects of envelope bandwidth on the intelligibility of sine- and noise-vocoded speech," *The Journal of the Acoustical Society of America*, vol. 126, pp. 792-805, 2009.