A Speech Processing Strategy based on Sinusoidal Speech Model for Cochlear Implant Users

Sungmin Lee[*], Sara Akbarzadeh[*], Satnam Singh[*], and Chin Tuan-Tan[*†]
[*]Erik jonsson school of engineering and computer science, University of Texas at Dallas, Richardson, USA
E-mail: Sung.Lee@utdallas.edu Tel: 1-901-337-2574
[†] School of behavioral and brain science, University of Texas at Dallas, Richardson, USA
E-mail: Chin-Tuan.Tan@utdallas.edu

***Abstract*— In sinusoidal modeling(SM), speech signal, which is pseudo-periodic in structure, can be approximated by sinusoids and noise without losing significant speech information. A speech processing strategy based on this sinusoidal speech model will be relevant for encoding electric pulse streams in cochlear implant (CI) processing, where the number of channels available is limited. In this study, 5 normal hearing(NH) listeners and 2 CI users were asked to perform the task of speech recognition and perceived sound quality rating on speech sentences processed in 12 different test conditions. The sinusoidal analysis/synthesis algorithm was limited to 1, 3 or 6 sinusoids from the sentences low-pass filtered at either 1 kHz, 1.5 kHz, 3 kHz, or 6 kHz, re-synthesized as the test conditions. Each of 12 lists of AzBio sentences was randomly chosen and process with one of 12 test conditions, before they were presented to each participant at 65 dB SPL (Sound Pressure Level). Participant was instructed to repeat the sentence as they perceived, and the number of words correctly recognized was scored. They were also asked to rate the perceived sound quality of the sentences including original speech sentence, on the scale of 1 (distorted) to 10 (clean). Both speech recognition score and perceived sound quality rating across all participants increase when the number of sinusoids increases and low-pass filter broadens. Our current finding showed that three sinusoids may be sufficient to elicit the nearly maximum speech intelligibility and quality necessary for both NH and CI listeners. Sinusoidal speech model has the potential in facilitating the basis for a speech processing strategy in CI.**

I. Introduction

Speech is acoustically a complex signal to transfer a variety of meaningful messages in communication. Despite the complex nature in speech, it can simply be represented by a small number of sinusoidal components that peak in original spectrum without much degradation in perceiving speech content. This strong concept of sinusoidal modeling [1, 2] enables scientists to study the mechanism of speech perception without any pre-assumption of the acoustic characteristics of speech elements, for instance, vowel and consonant. There is no restricted bound as in how the sinusoidal components are extracted for reconstruction of speech. The speech content can be reasonably well-reserved with just a small number of sinusoidal components. Several studies have investigated the perceptual outcomes associated with the unique approach of SM in decomposing input speech, and its role as a front-end processing block [3,4,5] in analysis/synthesis systems. Assuming speech signal typically being greater in intensity than noise, SM basically eliminates less-intense spectral regions and can be seen as a noise suppression or spectral enhancement that improves signal to noise ratio (SNR) [6]. Timms [3] resynthesized the consonants and vowels with 16 and 8 sinusoidal components extracted by SM and found that the speech perception outcome with NH listeners actually was degraded with fewer sinusoidal components, which may not support the role of SM as a noise suppression technique. However, they found that this overly sparse representation of speech does facilitate a better perceptual outcome for listener, which can be seen as a meaningful pre-processing in the perspective of spectral reduction, spectral transposition and temporal modification that are currently found in assistive hearing technology.

The capability of SM to select a small number of perceptually meaningful spectral/sinusoidal components (or sometime channels) in the context of speech, actually draws a similar analysis/synthesis scheme that is commonly deployed in current assistive hearing devices (for instance, hearing aid (HA) and cochlear implant (CI)), where the number channels for acoustic or electric are greatly reduced. Particularly for CI, which is one of the common prosthetic device designed to restore hearing sensation for listeners with severe to profound hearing loss and for those who are not able to receive much benefit from hearing aid. Fundamentally, CI extracts speech envelopes from the output of a limited number of band-limited channels, and each of that is used to modulate electrical pulses for direct auditory nerve stimulation via the associated intracochlear electrode. There is a strong conceptual similarity between CI users listening to speech via electric stimulation and NH listening to speech processed using SM. The fact that NH listener is able to achieve similar perceptual outcome with original speech and sparsely represented speech processed using SM has inspired current study to further examine such process in retaining perceptual meaningful speech information. We hypothesize that the SM process will reveal the underlying auditory perception mechanism in retrieving sufficient speech information with minimum representation of the signal. Parameters, like the minimum number of sinusoidal components and the frequency range over which the components are selected, for retaining speech information will provide us a different insight to understand the efficacy of current assistive hearing device in delivering perceptual meaningful speech information, which may also lead to different design methodology to add to the current technology.

II. Method

A. Sinusoidal Modeling

Periodic signals can be approximated by a sum of sinusoids whose frequencies, magnitudes and phases can be uniquely determined to match the signal. One way of obtaining these sinusoids is to perform Fourier transform on the signal and identify the spectral components that peak out in magnitude from the spectrum. In sinusoidal modeling, a similar concept is extended for longer and complex signal like speech, which is pseudo-periodic, which selects the spectral components in short time Fourier transform (STFT) over the whole duration of signal.

$$STFT\ \{x(n)\}(m, \omega) \equiv X(m, \omega) = \sum_{n=-\infty}^{\infty} x_n w_{n-m} e^{-i\omega n} \qquad (1)$$

In STFT (1), x_n and w_n represent the signal to be transformed and the window function, respectively, in time domain. X is the Fourier transform of the windowed signal which represents the phase and magnitude of the signal over time m and frequency ω . Spectrogram of a signal shows the magnitude squared of STFT of that signal as a function of time.

$$Spectrogram\ \{x(n)\}(m, \omega) \equiv |X(m, \omega)|^2 \qquad (2)$$

Figure 1 shows the spectrogram of the sentence ‘She has your dark suit in greasy wash water all the year’ from TIMIT data set. Spectral components of higher magnitude are shaded darker than those of lower magnitude. The sampling frequency of the speech signal was 16 kHz and the spectrogram was constructed with the window length for STFT kept at 256 samples and moving in step of 128 samples.

In sinusoidal modeling, spectral components that peaks in STFT magnitude processed at each time interval (128 samples) and exceeds a pre-defined threshold were identified. Before the identified spectral components were selected as the sinusoids for re-synthesis of speech, a process of eliminating the spectral components, which exceed a pre-defined threshold in magnitude and frequency when compared to the spectral components identified in the earlier time interval, from selection was performed. In Figure 2, the final sinusoids selected from the remaining identified spectral components were plotted as red dots overlaying on the original spectrogram. The red lines indicate the continuity of the selected sinusoids over a period of time. Finally, the speech was resynthesized with only the selected sinusoids that lie on these red lines.

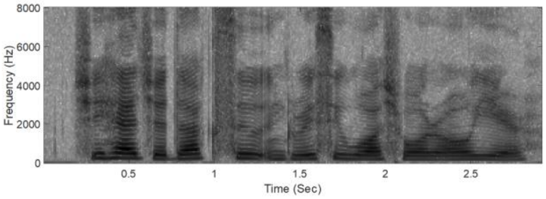

Figure 1. Spectrogram of a sample sentence

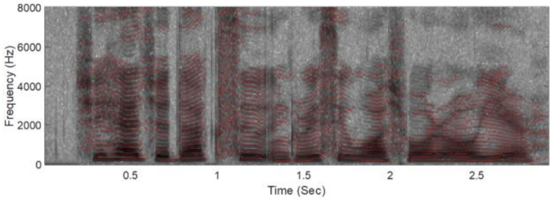

Figure 2. Selected spectral peaks in red for the sample sentence

B. Subjects

Six healthy normal hearing (NH) listeners (mean age: 25 years; age range: 22–28 years) identified with a self-report and hearing screening audiometry within 20 dB HL at octave frequencies from 250 to 8000 Hz, and two cochlear implant recipients (age: 61 and 65 respectively) with profound hearing loss participated in this study. All participants were native speakers of American English recruited from the Dallas, TX area. Listeners were compensated for their participation. The experimental protocol used in this study was approved by institutional review board of the University of Texas at Dallas.

C. Stimuli

This study aimed at exploring the feasibility of SM as a speech processing strategy in CI system. AzBio sentence test [7] which was developed for the purpose of pre- or post-operative CI evaluation was mainly used. The AzBio test consists of 15 lists in which each list contains 20 sentences. Among the 15 sets of original AzBio lists, 12 were randomly selected for modifications using above mentioned SM processing. The 12 lists were resynthesized with 12 different combinations (3*4) of the number of sinewaves and bandwidth using MATLAB R2017a. The three sinusoidal components (1, 3, and 6) and four bandwidths having cut-off frequency of low-pass filters at 1k, 1.5k, 3k, and 6 kHz were employed. A sampling frequency was 16 kHz and window length for short-time FFT was 256 samples. Figures 3 and 4 represent spectrograms of the sample sentence with varying number of selected spectral components (Figure 3 A: 6, B: 3, and C: 1) and bandwidths over which the spectral components were selected (Figure 4 A: 0-1 kHz, B: 0-1.5 kHz, C: 0-3 kHz, and D: 0-6 kHz) respectively.

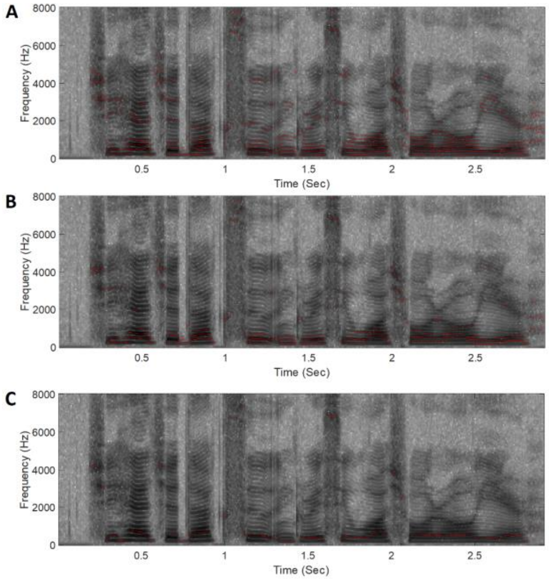

Figure 3. Different number of selected spectral peaks (A: 6, B: 3, and C: 1) in red for the sample sentence

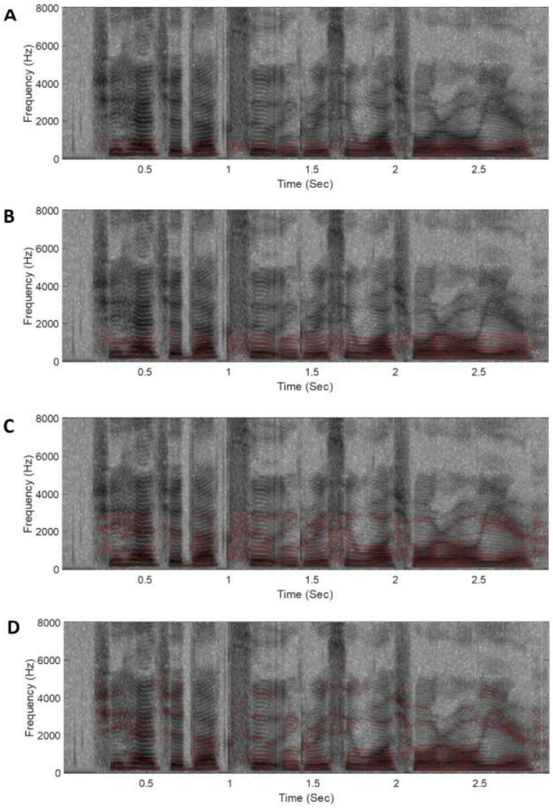

Figure 4. Selected spectral peaks in red for different bandwidths (A: 1 kHz, B: 1.5 kHz, C: 3 kHz, and D: 6 kHz)

D. *Procedures*

Each of 12 lists of the AzBio sentences were randomly chosen and processed with one of the 12 test conditions. Listener was seated in the middle of sound-treated booth. The speech stimuli were presented to each listener at 65 dB SPL throughout Sennheiser HD600 headphones for NH listeners and frontal speaker for CI users. The order of presentation for 12 lists of speech sentences was randomized for each individual. Listeners were instructed to carefully listen to sentences and repeat them. They were allowed to guess if they are unsure about the sentences. Their responses were scored in a word level in percentage.

After the speech recognition test, perceived sound quality of the sentence was also rated on a scale of 1 to 10, with 1 being the most distorted and 10 being the most natural. One of the sentences was chosen from the list 7 in AzBio test and processed with the 12 test conditions for this sound quality evaluation. The sound quality evaluation tasks were conducted twice for each participant in two different randomized orders. Average of the two ratings for each condition was computed and used in subsequent analysis. The whole procedure took an hour for each listener to complete.

III. Result and Discussion

A. *Listeners with normal hearing*

Speech recognition score

Overall, there is an increasing trend in the speech recognition score when the number of selected sinusoidal components increase and the bandwidths over which the sinusoidal components were selected. Figure 5 represents speech recognition scores as a function of the number of components (A) and bandwidths (B). A two-way repeated measure ANOVA was performed with the number of sinusoidal components and the bandwidth over which the sinusoidal components were selected as two within subject factors. The analysis showed significant effect with both the number of sinusoidal components [$F(2,10)=185.8$, $p < 0.01$] and the bandwidths [$F(3,15)=158.56$, $p < 0.01$]. Pair-wise comparisons for the number of sinusoidal components with Bonferroni adjustment showed that 1 vs. 6 and 1 vs. 3 comparisons were significantly different ($p<0.05$), but 3 vs. 6 pair was not ($p=0.058$). As for bandwidths, the pair-wise comparison showed that all the pairs were statistically different ($p<0.05$) except for 1 vs. 1.5k Hz ($p=0.061$) and 3 vs. 6 kHz comparisons ($p=0.155$). The analysis also found that there was a significant interaction between the two independent variables [$F(6,30)=15.345$, $p < 0.01$].

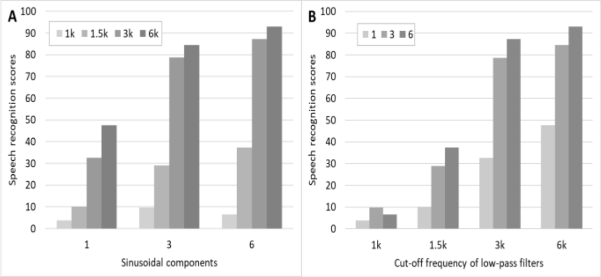

Figure 5. Speech recognition scores of NH participants for 12 test conditions shown as a function of sinusoidal components (A) and bandwidths (B)

Speech quality rating

Speech quality rating result also showed that participants' rating increases when the number of sinusoidal component increases and the bandwidth broadens. Figure 6 shows speech quality rating as a function of the number of components (A) and bandwidths (B). A two-way repeated measure ANOVA was performed with the number of sinusoidal components and the bandwidth over which the sinusoidal components were selected as two within subject factors. The analysis found significant effects with both the number of sinusoidal component [$F(2,10)=2.4613$, $p<0.01$] and the bandwidth [$F(3,15)=40.093$, $p<0.01$]. Likewise, Bonferroni adjusted pairwise comparisons with number of sinusoidal components found 1 vs. 6 and 1 vs. 3 comparisons ($p<0.05$) were significantly different, but 3 vs. 6 pair was not ($p=1.00$). As for bandwidths, the pair-wise comparison showed that all the pairs were statistically different except for 1 vs. 1.5k Hz ($p=0.551$) and 3 vs. 6 kHz pairs ($p=0.352$). The analysis also found that there was a significant interaction effect between the two independent variables [$F(6,30) = 3.302$, $p < 0.05$].

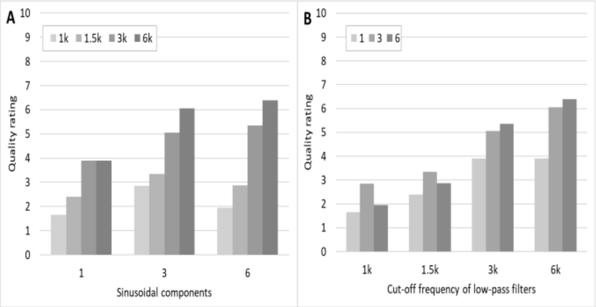

Figure 6. Speech quality rating of NH participants for 12 test conditions shown as a function of sinusoidal components (A) and bandwidths (B)

B. Listeners with CI

Due to the small sample size, ANOVAs were not performed. Instead, we were able to observe a similar trend in their speech recognition score and sound quality rating as their NH counterparts (Figures 7 and 8). We also observed that speech recognition scores with CI listeners were generally lower in value than those for NH listeners in most of the test conditions. To reach higher scores over 90%, NH participants would need 6 sinusoidal components, while CI participants reached approximately 70% with 6 sinusoidal components. However, CI participants rated higher sound quality than their NH counterparts in almost all the test conditions. Current trend of the outcomes suggest that CI participant may have greater preference for sparsely representation of speech. In further study we will work towards identifying the number of sinusoidal components for CI listeners to reach the highest possible speech recognition score.

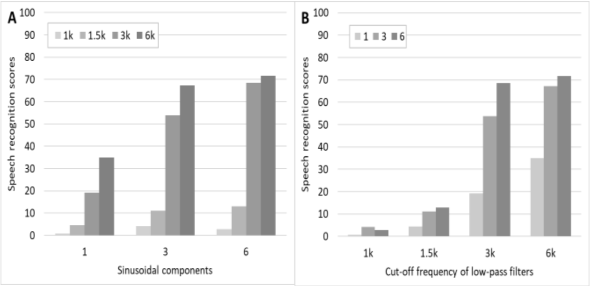

Figure 7. Speech recognition scores of CI participants for 12 test conditions shown as a function of sinusoidal components (A) and bandwidths (B)

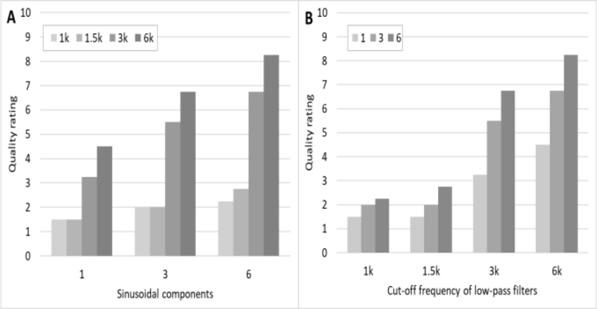

Figure 8. Speech quality rating of CI participants for 12 test conditions shown as a function of sinusoidal components (A) and bandwidths (B)

IV. Discussion

Interestingly, speech recognition scores for both NH and CI participants with 6 sinusoidal components were lower than those with 3 sinusoidal components, when the bandwidth is limited to 1 kHz. Clearly, there is not much perceptual meaningful speech information to retain in that frequency range of 0 – 1 kHz. Increasing the number of sinusoidal components in such a frequency range will just cause more confusion with highly redundant information and eventually reduce speech intelligibility instead. This observation agrees with the classical theory of acoustic/phonetic landmarks in speech perception. For instance, the first three formant frequencies (F1, F2 and F3) which are usually regarded as the necessary landmarks for perceiving voiced speech, are distributed over a frequency range that is wider than 1 kHz. Only in some voiced speech, F1 and/or even F2 can be located in the bandwidth below 1 kHz. Further investigation will also focus to avoid retaining redundant speech information.

Although one would hypothesize the underlying auditory perception mechanism between NH and CI participants are different, but the outcome of this study has shown that they are perceiving and rating the sound quality of speech processed by SM in a similar manner. Selecting 3 sinusoidal components over a frequency range of 0-3kHz to re-synthesize the speech will be sufficient in yielding a nearly maximum perceptual outcome that both NH and CI participants are able to achieve. One possible explanation is that SM is resynthesizing speech which retains speech information that is relevant to both NH and CI participants. An alternate explanation is that the way SM resynthesizes speech information may be relevant to CI in processing acoustic sound for electric stimulation to auditory nerve. Previous studies [8,9,10] have shown that speech perception performance for CI patients can vary significantly over a wide range. Our CI participants in current study are considered as good performer and their perceptual outcome may not fully reflect how CI patients perceive SM speech. Future study is required with greater sample size would be needed for a conclusive statement on the efficacy of SM in improving speech perceptual outcome for CI patients.

References

[1] R. J. McAulay, T. F. Quatlerl, “Speech analysis/synthesis based on a sinusoidal representation,” *IEEE Transactions on Acoustics, Speech, and Signal Processing,* vol. 34, no. 4, pp 744-754, Aug. 1986.

[2] R. J. McAulay, T. F. Quatlerl, “Pitch estimation and voicing detection based on a sinusoidal speech model.” In *ICASSP-90*, pp 249-252, Apr.1990.

[3] O. Timms, “Speech Processing Strategies Based on the Sinusoidal Speech,” *Doctoral dissertation,* 2003.

[4] K.S. Prasad, G.K. Ramaiah, M.B. Manjunatha. “Backend Tools for Speech Synthesis in Speech Processing,” *Indian Journal of Science and Technology*, vol. 10, no. 1, Jan. 2017.

[5] M. Asgari, I. Shafran, “Improvements to harmonic model for extracting better speech features in clinical applications,” *Computer Speech & Language, 47*, pp 298-313, Jan. 2018.

[6] J. M. Kates, “Speech Enhancement Based on a Sinusoidal Model,” *Journal of Speech and Hearing Research*, vol. 37 pp. 449-464, 1994.

[7] A. J. Spahr, M. F. Dorman, L. M. Litvak, S. Van Wie, R. H. Gifford, P. C. Liozou, L. M. Loiselle, T. Oakes, ans S. Cook, “Development and validation of the AzBio sentence lists,” *Ear Hear,* vol. 33, no. 1, pp 112-117, Jan-Feb 2012.

[8] A. C. Moberly, J. H. Lowenstein, S. Nittrouer, “Word recognition variability with cochlear implants:“Perceptual attention” versus “auditory sensitivity”,” *Ear and hearing*, vol. 37, no. 1, pp 14, Jan. 2016.

[9] G. Feng, E. M. Ingvalson, T. M. Grieco-Calub, M. Y. Roberts, M. E. Ryan, P. Birmingham, D. Burrowes,N. M. Young, P. C. Wong , “Neural preservation underlies speech improvement from auditory deprivation in young cochlear implant recipients,” *Proceedings of the National Academy of Sciences*, p. 201717603, Jan. 2018.

[10] D. James, K. Rajput, J. Brinton, U. Goswami, “Phonological awareness, vocabulary, and word reading in children who use cochlear implants: Does age of implantation explain individual variability in performance outcomes and growth?,” *Journal of Deaf Studies and Deaf Education*, vol. 13, no. 1, pp 117-137, Aug. 2007.