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

Factors Affecting the Intelligibility of Sine-wave Speech

Fei Chen [1], *Daniel Fogerty* [2]

[1] Department of Electrical and Electronic Engineering, Southern University of Science and Technology, Shenzhen, China
[2] Department of Communication Sciences and Disorders, University of South Carolina, USA
fchen@sustc.edu.cn, dfogerty@mailbox.sc.edu

Abstract

Studies on sine-wave speech (SWS) perception suggest that formants contain sufficient information for sentence intelligibility. This study further investigated the effects of amplitude modulation, number of sine-waves, and vowel resonance in SWS recognition. Results showed that Mandarin sentences synthesized using frequency trajectories of the first two formants were highly intelligible with additional contributions from formant amplitude modulation. However, amplitude modulation significantly contributed to intelligibility when only the vowels were preserved. The present work demonstrates that the intelligibility of Mandarin SWS can be largely attributed to the frequency transition of the first two formants and is susceptible to temporal interruption.

Index Terms: Speech intelligibility, sine-wave speech.

1. Introduction

The speech signal contains many acoustic cues for speech perception, several of which provide overlapping or complementary information. This high-dimensionality of speech is demonstrated, in part, by preserved speech recognition even when some acoustic cues are lost or removed. Hence, many studies have investigated the amount and type of acoustic information that is required for speech recognition, particularly for listening in quiet. Sine-wave speech (SWS) is designed to sparsely code the acoustic structure of speech by using sine-waves that are frequency and amplitude modulated by the formants. Other acoustic cues known to be important for speech understanding, e.g., harmonic structure and fundamental frequency (F0) contour, are discarded during this processing. Feng et al. reported the recognition performance of Mandarin SWS and found that Mandarin-speaking listeners could receive a high recognition rate for sine-wave sentences [1]. Sine-wave based speech synthesis has provided a useful framework to study the importance of formants to speech recognition [e.g., 1, 2, 3]. However, some questions remain regarding the contribution of formant information in SWS recognition. The objective of this work is to assess the relative importance of acoustic cues for SWS recognition.

The perceptual organization of these sine-wave replicas of speech was first investigated by Broadbent and Ladefoged using only two sine-waves: F1 and F2 [4]. They demonstrated perceptual fusion of these sine-waves on the bases of common frequency and amplitude modulation. However, even after more than half a century, a number of questions remain regarding how these sparse acoustic constituents of speech contribute to intelligibility. Specifically, this study measured the contribution of F3 commonly included in SWS, dissociated the contributions of frequency and amplitude modulation cues of the synthesized sine-wave speech, and investigated how these formant cues were represented in vowels across the sentence.

First, sine-wave speech is commonly synthesized with at least the first three formants, i.e., F1, F2 and F3 [e.g., 1, 2, 3]. However, F1 and F2 are sufficient for evoking speech processes, even when presented to separate ears [4]. Still unresolved is the degree to which this sparse representation is sufficient for providing meaningful information for successful sentence recognition. Earlier studies have demonstrated that the first two formants provide sufficiently good vowel identification [e.g., 5], particularly when capturing intrinsic vowel dynamics [6]. Furthermore, vowels carry important perceptual information for sentence intelligibility [e.g., 7, 8]. Therefore, how effective are the first two formants at independently conveying meaningful information for sentence intelligibility?

Second, the formant trajectories used in synthesizing SWS contain both frequency-varying and amplitude-varying information of formants. What is the relative importance of these two time-varying acoustic properties (i.e., amplitude and frequency modulation of formants) for the intelligibility of sine-wave speech? Carrell and Opie studied the effect of amplitude co-modulation (by a 100 Hz triangular waveform) on the intelligibility of sine-wave speech, and found a contributing effect of amplitude modulation to improve the intelligibility of sine-wave speech [3]. While many studies demonstrate the importance of temporal envelope (i.e., amplitude fluctuation) for speech perception, speech stimuli synthesized to preserve primarily the frequency modulations (i.e., the temporal fine structure, TFS) of the acoustic stimulus have also been found to be fairly intelligible [9]. This motivates the present work to examine the independent contribution of formant frequency trajectories (i.e., the amplitudes of formants are set to a constant level) to the intelligibility of sine-wave speech.

Third, the formants are primarily represented in the vowels where they have the greatest energy, although sine-wave synthesis also includes information from consonant oral, nasal, and fricative resonance. The importance of acoustic cues present during vowels has had significant recent attention, primarily through the use of a segment replacement paradigm that preserves either the vowels or the consonants in the sentence [e.g., 7]. Kewley-Port et al. reported that vowel-only sentences outperformed consonant-only sentences two-to-one [7]. Chen et al. demonstrated that preserving vowel segments can lead to almost perfect recognition of Mandarin sentences [8]. These results suggest an important role for acoustic properties present during vowels for sentence intelligibility.

http://dx.doi.org/10.21437/Interspeech.2016-4

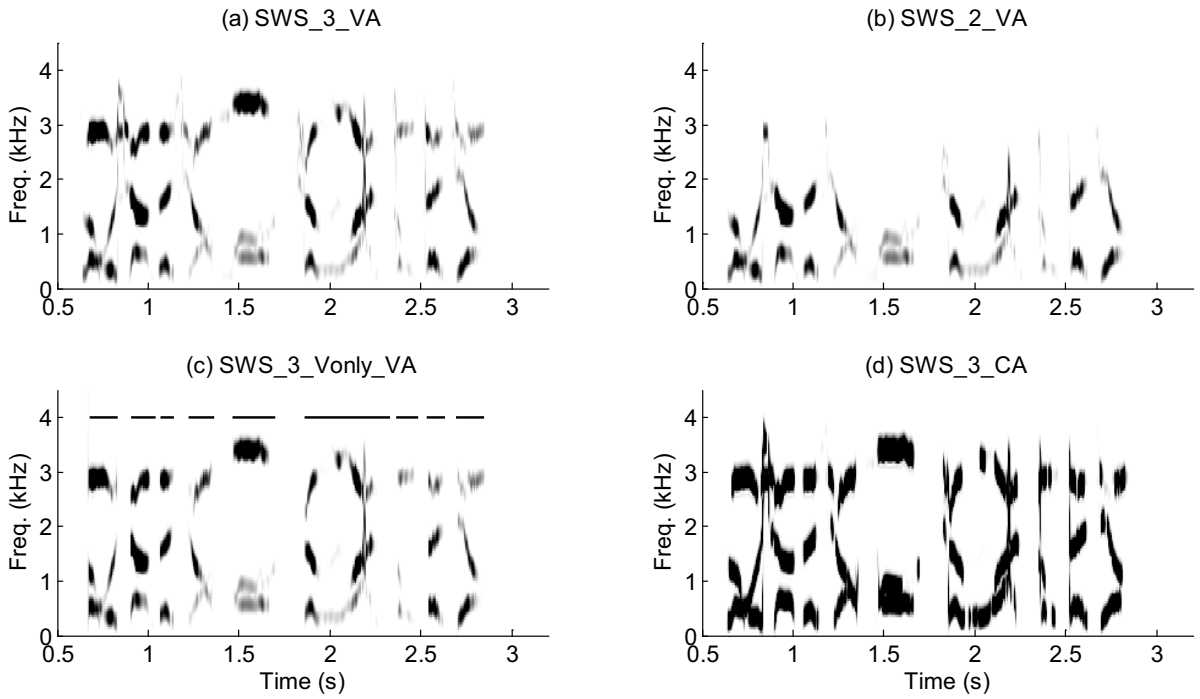

Figure 1. *Example spectrograms of sine-wave speech ('lou2 xia4 de1 xiao3 mao1 zheng3 wan3 dou1 zai4 jiao4' in Mandarin, or 'The cat on the ground floor shouted all the night' in English) synthesized with (a) the first three formants in condition SWS_3_VA, (b) the first two formants in condition SWS_2_VA, (c) the vowel segments of the first three formants in condition SWS_3_Vonly_VA, and (d) the first three formants with constant amplitude in condition SWS_3_CA. The solid lines in panel (c) delineate the vowel segments of the isolated words.*

Recent work has begun to define the general acoustic properties responsible for these vowel effects including the importance of amplitude modulation [e.g., 10]. This study specifically examines the contribution of the frequency and amplitude modulations of vowel formants to determine if these cues are sufficient for intelligibility in the absence of other vowel-related cues of harmonicity, formant bandwidth, (a)periodicity, etc. In addition, it determines if these cues, present during vowels, largely determine the intelligibility of sine-wave speech, even in the absence of formant resonance from other speech segments. This study will also assess the relative importance of formant frequency and formant amplitude to the intelligibility of vowel-only sine-wave speech.

2. Methods

2.1. Subjects and materials

Twelve (seven male) normal-hearing (NH) native Mandarin listeners (M=22 yrs, 19-25 yrs) were recruited by a convenient sampling from The University of Hong Kong. All subjects were paid for their participation in this study. The sentence materials were adopted from the Mandarin version of the Hearing in Noise Test (MHINT) [11]. There were 24 lists from the MHINT database, and each list composed of 10 ten-syllable Mandarin sentences. All the sentences were produced by a male speaker, with F0 ranging from 75 to 180 Hz.

2.2. Signal processing

This study generated six types of sine-wave speech containing different amounts of formant cues. When synthesizing the sine-wave speech, the formant trajectories with frequency-varying and amplitude-varying information (using 16-ms duration Hanning window with 50% overlapping between adjacent frames) were first extracted from the original speech signal based on linear predictive coding (LPC) analysis. The Matlab code to implement the above SWS processing is available at http://www.haskins.yale.edu/featured/sws/MATLAB/matlab.html [Last viewed 2 February 2016]. This study assessed the effects of three formant conditions on the intelligibility of sine-wave speech. More specifically, the sine-wave speech was synthesized with (1) the first three formants (i.e., F1, F2 and F3), and (2) the first two formants (i.e., F1 and F2). These two formant conditions are denoted as SWS_3 and SWS_2, respectively. In addition, based on the SWS_3 condition, this study further generated vowel-only sentences, i.e., preserving vowel segments while replacing the rest with silence (see [8] for more on vowel and consonant classification). This formant condition is denoted as SWS_3_Vonly. Figure 1 (a)-(c) show the spectrograms of an example sentence processed by the three formant conditions. The above-mentioned formant conditions used time-varying amplitude (i.e., condition 'VA') for the sinusoids to synthesize the sine-wave speech. The present work also examined the intelligibility of sine-wave speech synthesized by using constant sinusoid amplitude (i.e., condition 'CA') or an equal level for each sinusoid. Figure 1 (d) exemplifies the spectrogram of an example sentence processed by the SWS_3_CA condition.

2.3. Procedure

The experiment was conducted in a sound attenuating booth. Stimuli were played through a circumaural headphone binaurally at a comfortable listening level to listeners. Practice

(i.e., with feedback) of 40 non-experimental sentences (in condition SWS_3_VA) was given to listeners before the actual testing session. Each listener participated in a total of six experimental testing conditions [=3 formant conditions (i.e., SWS_3, SWS_2, and SWS_3_Vonly) × 2 amplitude conditions (i.e., varying-amplitude VA and constant-amplitude CA)], with each containing ten MHINT sentences. Condition test order was varied randomly across listeners and no sentence was repeated across conditions. Participants were allowed to listen to each stimulus for a maximum of three times and required to repeat as many words as they could recognize. The intelligibility score for each condition was computed as the ratio between the number of the correctly recognized words and the total number of words contained in each list of 10 MHINT sentences.

3. Results

Mean recognition scores of Mandarin sine-wave sentences for all conditions are shown in Table 1. Statistical significance was determined by using the percent recognition score as the dependent variable; the formant and amplitude conditions were the two within-subject factors. Note that for all statistical analyses below, the sentence recognition scores in Table 1 were first converted to rational arcsine units (RAU) using the rationalized arcsine transform (see Figure 2). Two-way analysis of variance (ANOVA) with repeated measures indicated a significant effect of formant condition (F[2, 22]=54.30, p<0.001), amplitude condition (F[1, 11]=23.29, p<0.001), and a non-significant interaction (F[2, 22]=1.16, p=0.331) between the formant and amplitude conditions.

Paired-comparison t-tests at the same formant condition showed a non-significant [t(11)=1.42, p>0.05] difference between the recognition scores of conditions SWS_3_VA and SWS_3_CA, a non-significant [t(11)=1.54, p>0.05] difference between conditions SWS_2_VA and SWS_2_CA, but a significant [t(11)=3.34, p<0.05] difference between conditions SWS_3_Vonly_VA and SWS_3_Vonly_CA. Thus, amplitude modulated formant information was only observed to impact performance when recognition was limited to the vowel segments.

Multiple paired-comparisons with Bonferroni correction were run among recognition scores at the same amplitude condition in Fig. 2. The Bonferroni-corrected statistical significance level was set at p<0.017 (α=0.05). Significant difference was observed between SWS_3 and SWS_2 conditions for either VA [t(11)=3.00, p<0.017] or CA testing [t(11)=3.67, p<0.017]. Compared to these full sentence conditions, performance was significantly poorer when sine-waves were presented only during the vowels, in both VA and CA conditions [i.e., t(11)=7.16, p<0.017 at VA condition, and t(11)= 7.37, p<0.017 at CA condition].

4. Discussion and conclusions

Given that many acoustic cues contained in the speech signal co-vary or are redundant for speech recognition, many studies have tried to identify the minimum necessary acoustic cues to support speech perception as well as the independent contributions of those acoustic properties [e.g., 9, 10]. Previous studies have demonstrated that sine-wave speech synthesized from the first three formants is highly intelligible [e.g., 1, 2, 3]. The present work further degraded the spectral and temporal specification of the SWS replicas by presenting only the first two formant trajectories. Experimental results showed that even only using the first two formant trajectories, the synthesized sine-wave speech was still highly intelligible, i.e., 88.7% for SWS_2_VA condition in Table 1. Consistent with previous work, we hypothesize that this ability to capture speech intelligibility from the first two formant trajectories may be attributed to the top-down processing commonly utilized in sentence recognition. That is, listeners can use their knowledge, language experience, and contextual information in speech to compensate the loss of acoustic cues when understanding processed or distorted speech. It is notable that sparse acoustic representation of speech, based on frequency modulation of the first two formants, is sufficient to evoke the perceptual organization of speech (see [2]) necessary for cognitive-linguistic processing.

When assessing the relative importance of formant frequency and formant amplitude to the intelligibility of sine-wave speech, this study found that formant frequency trajectory carried sufficient information to support high sentence intelligibility. Experimental results showed that in the SWS_3 and SWS_2 conditions, the differences between the recognition scores in the VA and CA conditions were not significant (see Fig. 2). In this regard, formant amplitude contributes minimally to the preserved high intelligibility of sine-wave speech that preserves time-varying formant frequency trajectories. The formant frequency trajectory captures important information from speech production, e.g., resonance in vocal tract. These results suggest that the peak resonant response of the vocal tract is particularly important for intelligibility, with additional contribution from relative amplitude changes in that resonant response. In addition, this study showed that the Mandarin speech synthesized with the first two formants (with time-varying frequency but constant amplitude) was highly intelligible, i.e., 84.1% in condition SWS_2_CA. This suggests that the intelligibility of Mandarin

Table 1. Sine-wave sentence recognition scores for all conditions.

	VA	CA
SWS_3	95.7±1.8	91.8±2.5
SWS_2	88.7±3.0	84.1±3.7
SWS_3_Vonly	70.7±4.3	53.3±8.0

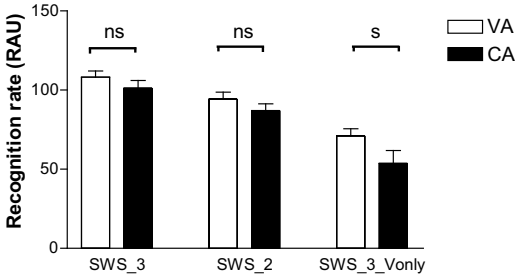

Figure 2. *Mandarin sine-wave sentence recognition scores (in RAU) for all conditions. The error bars denote ±1 standard error of the mean. 'ns' and 's' denote that the recognition scores (in RAU) are non-significantly (p>0.05) or significantly (p<0.05) different between conditions varying-amplitude 'VA' and constant-amplitude 'CA'.*

sine-wave speech can be largely attributed to the frequency trajectories of the first two formants.

However, the amplitudes of formants do play a significant role for sine-wave speech recognition when the speech is interrupted, e.g., segmental interruption in this study. The vowel-only sine-wave speech synthesized with both formant frequency and amplitude information was 70.7%; however when the formant amplitude information was removed, the recognition score notably declined to 53.3%. This finding is consistent with many previous results. That is, in adverse conditions, additional specification of the acoustic speech signal becomes necessary [e.g., 11]. In addition, these results are in line with other studies of vowel contributions to sentence intelligibility that highlight the importance of amplitude modulation of the vowels during temporal interruption [e.g., 10]. Further investigation is necessary regarding the relative contributions of these acoustic properties in challenging listening environments, e.g., in noise. While performance was degraded in the vowel-only conditions, it is notable that even these highly degraded vowels (i.e., three constant amplitude, frequency-varying sine-waves) resulted in higher sentence intelligibility scores than Mandarin sentences that preserve the full spectrum consonants (53% for SWS_3_Vonly_CA versus 34% for consonant-only sentences; [8]).

Carrell and Opie found that the time-varying sinusoidal sentences with constant sinusoid amplitude had much lower intelligibility scores compared to those with modulated amplitude (by a 100 Hz triangular waveform) [3]. However, the present work showed that the SWS_3_CA condition still led to a very high intelligibility score, i.e., 91.8%, possibly indicating that frequency modulation alone is sufficient for perceptually organizing the three sinusoids into one speech percept. It is not clear whether the difference between these studies could be attributed to a language difference. However, initial pilot testing with four native English-speaking NH adults with English HINT sentences resulted in a mean of 80% (SD = 19%) for SWS_3_VA and 35% (SD = 12%) for SWS_3_CA, $t(3)=10.5$, $p<0.01$. This suggests that performance in general may be lower for English materials and that amplitude modulation of the formants may be essential for maximum intelligibility of English SWS sentences. Further study is required to define these language differences. In addition to the importance of tonal contour for speech perception, Mandarin has a simple consonant-vowel syllable structure that differs significantly from the complex syllable structure of English that allows for consonant clusters. Hence, we postulate that English sine-wave speech may have more temporal interruption due to its consonant clusters (i.e., lacking formant information) than that in Mandarin, which may have a detrimental effect on understanding English sine-wave sentences. Furthermore, amplitude modulation may provide important information regarding the complex syllable structures of English. For example, amplitude cues may provide information regarding variation and transitions between oral, nasal, and fricative resonance, as well as non-resonant portions of speech. Further investigation is required to identify the type of speech information provided by amplitude modulation of the sine-wave formant analogue and its contribution to intelligibility, particularly as it is related to phonetic and phonotactic differences between English and Mandarin.

In conclusion, the present work suggested that the intelligibility of Mandarin SWS could be largely attributed to the frequency transition of the first two formants. Formant amplitude resulted in additional contribution to intelligibility relative to formant frequency trajectories. Nevertheless, formant amplitude does appear to contribute significantly to understanding of sine-wave sentences in adverse listening conditions, e.g., temporal interruption by only preserving vowel segments.

5. Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 61571213). This study was also supported by a grant from Neural and Cognitive Sciences Research Center, Southern University of Science and Technology, Shenzhen, China.

6. References

[1] Feng, Y. M., Xu, L., Zhou, N., Yang, G., and Yin, S. K., "Sinewave speech recognition in a tonal language," J. Acoust. Soc. Am., 131: 133–138, 2012.

[2] Remez, R. E., Rubin, P. E., Pisoni, D. B., and Carrell, T. D., "Speech perception without traditional speech cues," Science, 212: 947–949, 1981.

[3] Carrel, T. D. and Opie, J. M., "The effect of amplitude comodulation on auditory object formation in sentence perception," Percept Psychophys., 52: 437–445, 1992.

[4] Broadbent, D.E. and Ladefoged, P., "On the fusion of sounds reaching different sense organs," J. Acoust. Soc. Am., 29: 115–127, 1957.

[5] Peterson, G. and Barney, H., "Control methods used in a study of vowels," J. Acoust. Soc. Am., 24: 175–184, 1952.

[6] Hillenbrand, J. M., "Static and dynamic approaches to vowel perception," in Vowel inherent spectral change edited by G. S. Morrison and P. F. Assmann (Springer, Berlin, Heidelberg), pp. 9–30, 2013.

[7] Kewley-Port, D., Burkle, T. Z., and Lee, J. H., "Contribution of consonant versus vowel information to sentence intelligibility for young normal-hearing and elderly hearing-impaired listeners," J. Acoust. Soc. Am., 122: 2365–2375, 2007.

[8] Chen, F., Wong, L. L. N., and Wong, Y. W., "Assessing the perceptual contributions of vowels and consonants to Mandarin sentence intelligibility," J. Acoust. Soc. Am., 134: EL178–EL184, 2013.

[9] Smith, Z. M., Delgutte, B., and Oxenham, A. J., "Chimaeric sounds reveal dichotomies in auditory perception," Nature, 416: 87–90, 2002.

[10] Fogerty, D. and Humes, L. E., "The role of vowel and consonant fundamental frequency, envelope, and temporal fine structure cues to the intelligibility of words and sentences," J. Acoust. Soc. Am., 131: 1490–1501, 2012.

[11] Chen, F., Wong, L. L. N., and Hu, Y., "Effects of lexical tone contour on Mandarin sentence intelligibility," J. Speech Lang. Hear. Res., 57: 338–345, 2014.