Sine-wave speech recognition in a tonal language

Yan-Mei Feng

*Department of Otolaryngology, Shanghai Sixth People's Hospital, Institute of Otolaryngology,
Shanghai Jiao Tong University, Shanghai 200233, People's Republic of China
yanmeifeng2008@gmail.com*

Li Xu

*School of Rehabilitation and Communication Sciences, Ohio University, Athens, Ohio 45701
xul@ohio.edu*

Ning Zhou

*Kresge Hearing Research Institute, University of Michigan, Ann Arbor, Michigan 48109
nzhoujo@umich.edu*

Guang Yang and Shan-Kai Yin^{a)}

*Department of Otolaryngology, Shanghai Sixth People's Hospital, Institute of Otolaryngology,
Shanghai Jiao Tong University, Shanghai 200233, People's Republic of China
gyang321@hotmail.com, yinshankai@china.com*

Abstract: It is hypothesized that in sine-wave replicas of natural speech, lexical tone recognition would be severely impaired due to the loss of F0 information, but the linguistic information at the sentence level could be retrieved even with limited tone information. Forty-one native Mandarin-Chinese-speaking listeners participated in the experiments. Results showed that sine-wave tone-recognition performance was on average only 32.7% correct. However, sine-wave sentence-recognition performance was very accurate, approximately 92% correct on average. Therefore the functional load of lexical tones on sentence recognition is limited, and the high-level recognition of sine-wave sentences is likely attributed to the perceptual organization that is influenced by top-down processes.

© 2012 Acoustical Society of America

PACS numbers: 43.71.Gv [SGS]

Date Received: August 16, 2011

Date Accepted: November 10, 2011

1. Introduction

Remez *et al.* (1981) reported that sine-wave replicas of full-spectrum speech signals are sufficient to support perception of the linguistic message in the absence of traditional acoustic cues when the listeners were instructed to listen for speech. A single sentence “Where were you a year ago?” spoken by an adult male voice was used in that report. In such sine-wave speech, three sinusoids follow in time the frequencies and amplitudes of the first three formants of the natural utterances. Sine-wave speech lacks many of the acoustic features of natural speech that are thought to be important for speech perception, such as the broadband formant structures, formant frequency transitions, and harmonics of a common fundamental frequency (F0) (Remez *et al.*, 1981; Remez *et al.*, 1994). Several other studies have reported similar findings that sine-wave replicas of English sentences are highly intelligible, with recognition performance ranging from 60% to 100% correct depending on the difficulty levels of the sentence materials (Carrell and Opie, 1992; Barker and Cooke, 1999; Nittrouer and Lowenstein, 2010).

In tonal languages, F0 patterns of the vocalic part of monosyllables convey lexical meanings. Mandarin Chinese is one of the most widely spoken tonal languages.

^{a)} Author to whom correspondence should be addressed.

Fig. 1. (Color online) Narrowband spectrograms of the natural speech materials (upper panels) and the sine-wave replicas of the same speech materials (lower panels). The left small panels show the spectrograms of four tone tokens (i.e., /ma1/, /ma2/, /ma3/, and /ma4/) of a male voice. The right panels show the spectrograms of an MHINT sentence (i.e., /wo3 men2 dou1 zai4 xue2 xi2 dian4 nao3 da3 zi4/ or “We are all learning how to type on a computer” in English) of a male voice.

It has four F0 patterns (i.e., tones): (1) high-level, (2) rising, (3) falling and then rising, and (4) high-falling (Fig. 1, upper left panels). It is known that the F0 and its harmonics are the primary cues for lexical tone recognition (Liang, 1963; Lin, 1988; Fu *et al.*, 1998; Xu *et al.*, 2002). In the classic source-filter theory of speech production (Fant, 1960), the F0 is derived from the source (i.e., vocal fold vibration). The filter (i.e., the vocal tract configuration) gives rise to the formant peaks. In sine-wave speech processing, the F0 and its harmonics are absent, and only three sinusoids follow the peak values of the first three formants. As expected, the spectrograms of the sine-wave replicas of four tones of the same syllable are similar with only subtle differences. As shown in Fig. 1 (lower left panels), the first three tones had very similar spectrograms, whereas tone 4 showed a fairly large front peak in all three formants at the beginning of the utterance. Such a peak, however, was not a consistent observation across all syllables used in the present study. At the sentence level (Fig. 1, right panels), the first, second, and third formant contours in the sine-wave speech are independent of the F0 (or more easily visible, its harmonics) variation patterns and provide no consistent cues for F0 of the natural speech.

Therefore, we hypothesize that lexical tone recognition for sine-wave processed Mandarin Chinese monosyllabic words is poor. Assuming poor tone recognition, the second research question we sought to address is whether the linguistic messages of sine-wave speech at the sentence level can be retrieved with limited or no tone information. In the present study, two sets of experiments were conducted to test tone and sentence recognition performance with sine-wave processed Mandarin Chinese monosyllabic words and sentences.

2. Method

Forty-one native Mandarin-Chinese-speaking young adults (21 females and 20 males, aged 23–30 yr old) were recruited from the Shanghai Jiao Tong University graduate student population. They had normal hearing with pure-tone air-conduction thresholds

≤ 20 dB HL (re: ANSI, 2010) at octave frequencies from 250 to 8000 Hz in both ears. None of the subjects had a history of ear diseases or experience of listening to sine-wave speech. The use of human subjects in this study was reviewed and approved by the Institutional Review Board of Shanghai Sixth People's Hospital.

The Mandarin Chinese speech test materials consisted of tone and sentence stimuli. The tone test materials were adopted from Zhang (1990) in which there were 10 different monosyllables (i.e., /fu/, /ji/, /ma/, /qi/, /wan/, /xi/, /xian/, /yan/, /yang/, and /yi/) each with four tones. The resultant 40 monosyllabic words were recorded in a male and a female voice. The durations of the four different tones for the same syllable were all equal as described in detail in Xu *et al.* (2002). The sentence test materials were the Mandarin Chinese version of the Hearing in Noise Test (MHINT) (Wong *et al.*, 2007). Each list in MHINT contains 20 sentences, and each sentence consists of 10 keywords. The MHINT sentences represent a simple, conversational style of speech and are easily understood by native Mandarin-Chinese-speaking listeners with varied educational backgrounds (Wong *et al.*, 2007). Previous reports have shown that normal-hearing, native Mandarin-Chinese-speaking listeners could achieve nearly perfect recognition using these tone and sentence test materials in quiet conditions (Zhang, 1990; Xu *et al.*, 2002; Zhang *et al.*, 2010; Wang *et al.*, 2011).

Sine-wave replicas of the natural speech were created by extracting the center frequency of the first three formants of the original tone tokens and MHINT sentences using Linear Predictor Coefficient (LPC) analysis. The extracted formants were then replaced with sinusoid replicas. The process was automated by means of a script written by Darwin (2003) using the PRAAT scripting language (Boersma and Weenink, 2010). The spectrograms of the sine-wave replicas and those of the natural speech materials were compared to ensure that the contours of the sine waves matched the formant frequencies of the natural speech (see Fig. 1).

The experiments were conducted in a double-walled sound booth. The stimuli were presented at the most comfortable level for each subject, approximately 65 dB (A) [ranging from 62.5 to 67.5 dB (A)]. The level was set during the practice session. The loudspeaker for stimulus presentation was mounted 1 m in front of the subject at 0° azimuth. The tone test was in a four-alternative forced-choice format in which the subjects were required to choose one of the four tones after each presentation of a tone token. A custom graphical user interface written in MATLAB (Mathworks, Natick, MA) was used to present the stimuli and to record the responses during the experiment. A total of 80 sine-wave processed tokens (i.e., 10 syllables $\times$ 4 tones $\times$ 2 voices) were used for practice with feedback. The test consisted of 160 sine-wave processed tokens (i.e., 10 syllables $\times$ 4 tones $\times$ 2 voices $\times$ 2 repetitions) that were presented in a random order with no feedback. For the sine-wave sentence recognition test, two MHINT sentence lists were used for practice and two different lists were used for test. The sentence stimuli chosen at random from the lists were presented in quiet via the loudspeaker and the subjects were required to write down each word they had heard. The subjects were allowed to listen to the sentence for as many times as they desired. Feedback was only provided during practice. The order of the two tests (i.e., tone recognition and sentence recognitions tests) was chosen at random for each subject. It took approximately 1-2 h to complete the two tests for each subject.

3. Results and discussion

Sine-wave tone recognition performance in the group of 41 normal-hearing, native Mandarin-speaking subjects was only slightly above chance (which would be 25% correct) (Fig. 2). The mean performance was 32.7% correct with a range of 23.8% to 44.4% correct. A *t*-test indicated that the mean score was significantly different from chance of 25% correct ($t(40) = 9.84$, $P < 0.001$). While statistically significant, performance was only slightly better than chance and much poorer than what one would expect from natural speech. A binomial-variable analysis (Thornton and Raffin, 1978) revealed that for any individual subject, a tone-recognition score of $>35\%$ correct is

Fig. 2. Recognition scores of sine-wave Mandarin-Chinese tones and sentences. The three horizontal solid lines of the box represent the 25th percentile, median, and 75th percentile of the data. The whiskers show the range. The filled symbols represent individual data. The dashed line on the left indicates the chance level (i.e., 25% correct) for tone recognition.

significantly higher than the chance performance ($P < 0.05$). Thus, 12 of the 41 subjects (i.e., 29.3%) showed sine-wave tone recognition significantly higher than chance, whereas the remaining 29 subjects (i.e., 70.7%) did not.

The results of the sine-wave tone recognition test were consistent with the hypothesis that performance would be poor when using only formant peak information because the contours of the sine waves bear no clear relationship with the F0 contours (Fig. 1, left panels). Note that the tone tokens were equal in duration. Performance might have been higher if the duration cue was preserved (see also Xu *et al.*, 2002). Remez and Rubin (1984) showed that the lowest-frequency sinusoid that replicates the first formant has a dominant effect on the perception of the intonation patterns of the sine-wave sentences. In the example shown in Fig. 1 (lower left panels), the lowest sinusoids had a frequency around 1000 Hz, much higher than the typical F0, and their contours did not resemble those of the F0 or the harmonics of the original tone tokens except for tone 1 (Fig. 1, upper left panels). For some vowels (e.g., /i/ or /u/), however, the lowest sinusoids could be in the vicinity of F0.

The slightly-higher-than-chance performance may be attributed to the temporal amplitude cues in the sine-wave signals. Several studies have shown that temporal amplitude cues alone may yield tone-recognition scores as high as 50% to 70% correct in Mandarin Chinese (Whalen and Xu, 1992; Fu and Zeng, 2000; Xu *et al.*, 2002; Luo and Fu, 2004; Kong and Zeng, 2006). However, it is possible that the frequency modulations of the three sinusoids and the inharmonic relationship among them have made the amplitudes of the sinusoids less useful cues for tone recognition. It is also possible that the lower performance level in the current study could be attributed to the neutralization of duration in the lexical tones.

The confusion matrix of sine-wave tone recognition pooled across all subjects is shown in Table 1. The distributions of response error patterns were fairly even across the four tones, although there were relatively more responses for tone 1 but fewer for tone 4. This is consistent with our hypothesis that the sine-wave replication neutralizes the F0 information thus the monotone effect produces a preponderance of tone 1 perceptions. We also compared the sine-wave tone recognition of the male and female speakers. The overall mean scores were 34.6% and 31.2% correct for the male and female speaker, respectively. This difference, although fairly small, was statistically significant [paired *t*-test, $t(40) = 2.44$, $P < 0.05$].

Table 1. Confusion matrix for the sine-wave tone recognition. Data were pooled across 41 subjects.

		Responses (%)			
		Tone 1	Tone 2	Tone 3	Tone 4
Stimuli	Tone 1	43.2	21.6	14.1	21.1
	Tone 2	27.2	23.8	28.7	20.3
	Tone 3	23.4	24.8	34.7	17.1
	Tone 4	28.2	21.7	20.4	29.7
	Total	122.0	91.9	97.9	88.2

Despite the poor performance in sine-wave tone recognition, our subjects were very accurate in Mandarin-Chinese sine-wave sentence recognition. As shown in Fig. 2, the sine-wave sentence recognition scores were between 78% and 100% correct with a mean performance of 91.6% correct. A significant but weak linear correlation was found between sine-wave tone recognition and sentence recognition performance (Pearson $r = 0.385$, $P < 0.05$). Hence, sine-wave tone recognition performance accounted for less than 15% of the variance in sine-wave sentence recognition performance in Mandarin Chinese. Although the functional load or weighting parameter of lexical tone for recognizing speech at sentences levels could not be directly computed, it is conceivable that its role is limited. It has been shown that tones are probably not essential for comprehension at sentence level because the acoustic distinction between tones is compromised in natural running speech (e.g., Xu, 2006).

The processing mechanism of sine-wave speech recognition in Mandarin Chinese is probably not different from that in English. The temporospectral patterns in the sine-wave signals elicit phonetic perceptual organization (Remez, 2008). The process is likely to be strongly influenced by top-down processes. With knowledge of the syntax and semantics of their native language, listeners are able to make use of contextual information in the sentences to compensate for the poor tone information in the sine-wave speech signals. The speculation could be tested by comparing sine-wave speech recognition using high-probability versus low-probability sentences or even syntactically appropriate but semantically anomalous sentences in future studies.

Acknowledgments

The research was supported in part by Grant No. 81070778 from NSFC and NIH/NIDCD Grant No. R15-DC009504.

References and links

- ANSI S3.6–2010. (2010). *Specifications for audiometers* (Acoustical Society of America, New York).
- Barker, J., and Cooke, M. (1999). "Is the sine-wave speech cocktail party worth attending?" *Speech Commun.* **27**, 159–174.
- Boersma, P., and Weenink, D. (2010). "PRAAT: doing phonetics by computer (version 5.2.35) [computer program]." <http://www.praat.org> (Last viewed June 11, 2010).
- Carrel, T. D., and Opie, J. M. (1992). "The effect of amplitude comodulation on auditory object formation in sentence perception," *Percept Psychophys.* **52**, 437–445.
- Darwin, C. (2003). "Sine-wave speech produced automatically using a script for the PRAAT program," http://www.lifesci.sussex.ac.uk/home/Chris_Darwin/SWS/ (Last viewed December 8, 2011).
- Fant, G. (1960). *Acoustic Theory of Speech Production* (Mouton, The Hague).
- Fu, Q. J., and Zeng, F. G. (2000). "Identification of temporal envelope cues in Chinese tone recognition," *Asia Pac. J. Speech Lang. Hear.* **5**, 45–57.
- Fu, Q. J., Zeng, F. G., Shannon, R. V., and Soli, S. D. (1998). "Importance of tonal envelope cues in Chinese speech recognition," *J. Acoust. Soc. Am.* **104**, 505–510.
- Kong, Y. Y., and Zeng, F. G. (2006). "Temporal and spectral cues in Mandarin tone recognition," *J. Acoust. Soc. Am.* **120**, 2830–2840.
- Liang, Z. A. (1963). "The auditory perception of Mandarin tones," *Acta Physiol. Sinica* **26**, 85–91.

- Lin, M.C. (1988). "The acoustic characteristics and perceptual cues of tones in Standard Chinese," *Zhongguo Yuwen* **204**, 182–193.
- Luo, X., and Fu, Q. J. (2004). "Enhancing Chinese tone recognition by manipulating amplitude envelope: implications for cochlear implant," *J. Acoust. Soc. Am.* **116**, 3659–3667.
- Nittrouer, S., and Lowenstein, J. H. (2010). "Learning to perceptually organize speech signals in native fashion," *J. Acoust. Soc. Am.* **127**, 1624–1635.
- Remez, R. E. (2008). "Perceptual organization of speech," in *The Handbook of Speech Perception*, edited by D. B. Pisoni and R. E. Remez (Blackwell, Malden, MA).
- Remez, R. E., and Rubin, P. E. (1984). "On the perception of intonation from sinusoidal sentences," *Percept. Psychophys.* **35**, 429–440.
- Remez, R. E., Rubin, P. E., Berns, S. M., Pardo, J. S., and Lang, J. M. (1994). "On the perceptual organization of speech," *Psychol. Rev.* **101**, 129–156.
- Remez, R. E., Rubin, P. E., Pisoni, D. B., and Carrell, T. D. (1981). "Speech perception without traditional speech cues," *Science* **212**, 947–949.
- Thornton, A. R., and Raffin, M. J. (1978). "Speech discrimination scores modeled as a binomial variable," *J. Speech Hear. Res.* **21**, 507–518.
- Wang, W., Zhou, N., and Xu, L. (2011). "Musical pitch and lexical tone perception with cochlear implants," *Int. J. Audiol.* **50**, 270–278.
- Whalen, D. H., and Xu, Y. (1992). "Information for Mandarin tones in the amplitude contour and in brief segments," *Phonetica* **49**, 25–47.
- Wong, L. L., Soli, S. D., Liu, S., Han, N., and Huang, M. W. (2007). "Development of the Mandarin Hearing in Noise Test (MHINT)," *Ear Hear.* **28**, 70S–74S.
- Xu, L., Tsai, Y., and Pfingst, B. E. (2002). "Features of stimulation affecting tonal-speech perception: implications for cochlear prostheses," *J. Acoust. Soc. Am.* **112**, 247–258.
- Xu, Y. (2006). "Tone in connected discourse," in *Encyclopedia of Language and Linguistics*, edited by K. Brown, 2nd ed. (Elsevier, Oxford), Vol. 12, pp. 742–750.
- Zhang, H. (1990). "Design and preliminary application of the minimal auditory capabilities in Chinese people," *Chinese J. Otorhinolaryngol.* **25**, 79–82.
- Zhang, N., Liu, S., Xu, J., Liu, B., Qi, B., Yang, Y., Kong, Y., and Han, D. (2010). "Development and applications of alternative methods of segmentation for Mandarin Hearing in Noise Test in normal-hearing listeners and cochlear implant users," *Acta Otolaryngol.* **130**, 831–837.