Assessing the perceptual contributions of vowels and consonants to Mandarin sentence intelligibility

Fei Chen,[a] Lena L. N. Wong, and Eva Y. W. Wong
Division of Speech and Hearing Sciences, The University of Hong Kong, Prince Philip Dental Hospital, 34 Hospital Road, Hong Kong, Hong Kong
feichen1@hku.hk, llnwong@hku.hk, evawyw09@hku.hk

Abstract: This study investigated the perceptual contributions of vowels and consonants to Mandarin sentence intelligibility. Mandarin sentences were edited using a noise-replacement paradigm to preserve various amounts of segmental information and presented to normal-hearing listeners to recognize. The vowel-only Mandarin sentences yielded a remarkable 3:1 intelligibility advantage over the consonant-only sentences. This advantage is larger than that obtained with English sentences, suggesting that vowels may have a greater contribution to sentence intelligibility in Mandarin than in English. Although providing information redundant to contributions from vowel centers, a little vowel-consonant boundary transition would significantly improve the intelligibility of the consonant-only Mandarin sentences.

PACS numbers: 43.71.Gv, 43.71.Es [AC]
Date Received: April 29, 2013 **Date Accepted:** June 17, 2013

1. Introduction

Vowels and consonants are two categories of speech sounds existing in all languages. Vowels are characterized by a relatively open vocal tract with sustained voicing in production and low frequency energy and long duration, whereas consonants are characterized by complete or partial vocal tract constriction in production and high frequency energy and short duration. A number of studies have attempted to investigate the perceptual contributions of vowels and consonants to speech (e.g., word and sentence) intelligibility (e.g., Owens *et al.*, 1968; Cole *et al.*, 1996; Cutler *et al.*, 2000; Bonatti *et al.*, 2005; Owren and Cardillo, 2006; Kewley-Port *et al.*, 2007; Fogerty and Kewley-Port, 2009). These studies have produced evidence of greater contributions by consonants under some conditions (e.g., Owens *et al.*, 1968; Cutler *et al.*, 2000; Bonatti *et al.*, 2005; Owren and Cardillo, 2006) and greater contributions by vowels under other conditions (e.g., Cole *et al.*, 1996; Kewley-Port *et al.*, 2007; Fogerty and Kewley-Port, 2009).

Cole *et al.* (1996) replaced vowel or consonant segments with speech-shaped noise in sentences taken from the TIMIT corpus. This process is known as the "noise-replacement paradigm." They found that the vowel-only sentences (consonants replaced) led to a remarkable 2:1 intelligibility advantage over the consonant-only sentences (vowels replaced). Kewley-Port *et al.* (2007) also confirmed the 2:1 intelligibility advantage of vowels for young normal-hearing (NH) listeners at 70 dB sound pressure level (SPL) and for elderly hearing-impaired (HI) listeners at 95 dB SPL, although elderly HI listeners had overall poorer performance than young NH listeners. Fogerty and Kewley-Port (2009) investigated how the relative perceptual contributions of consonants and vowels were affected by the co-articulation information across consonant-vowel (C-V) boundaries using the noise-replacement paradigm. In their study, glimpse

[a]Author to whom correspondence should be addressed.

windows were defined as the preserved speech signal intervals and contained proportional amounts of vowel information at C-V boundaries (i.e., VP) that was either added to consonants (i.e., C + VP) or deleted from vowels (i.e., V-VP). Their results once again confirmed the 2:1 intelligibility advantage of vowels over consonants in (English) sentences. Besides it was found that sentence intelligibility increased linearly under the C + VP conditions with the increase in VP, and the intelligibility under the V-VP conditions was unaffected until 30% VP was replaced by noise.

However, as all of the preceding studies used sentences from the TIMIT corpus, which is an English sentence database, it is unclear whether their findings may be applicable to other languages, such as Mandarin Chinese. Although both Mandarin and English contain vowels and consonants, they are different in many aspects. First of all, Mandarin is a tonal language in which lexical tone carries important information for distinguishing the meaning of each Mandarin word. The four lexical tones in Mandarin are characterized by the fundamental frequency (F0) contours of the voiced segments, and an identical syllable with different lexical tone conveys different meaning (Howie, 1976). In contrast, F0 contour in English conveys no lexical meaning. Second, syllable structure in English is rather complex in that consonant clusters can appear in both onset and coda of a syllable. However, Mandarin has no consonant cluster, and its syllable is basically simple, i.e., consonant-vowel. Li *et al.* (2000) proposed that there are just about 415 permutations of Mandarin syllable in common use. Even when lexical tones are considered, there are only about 1200 syllables in Mandarin (Howie, 1976). However, there are far more permutations of syllables in English because of the existence of consonant clusters. There are around 10 000 non-homophonous monosyllables in English according to the CMU Pronouncing Dictionary (Carnegie Mellon University Pronouncing Dictionary, http://www.speech.cs.cmu.edu/cgi-bin/cmudict). Hence the probability that a listener could accurately identify a syllable based on the vowel information available would be lower in English because of the existence of the large number of syllable permutations.

The purpose of the present study is to investigate the segmental contributions to sentence intelligibility in Mandarin. More specifically, this study will assess the perceptual contributions of vowels, consonants, vowel-consonant boundaries, and vowel centers to the intelligibility of Mandarin sentences. As the main acoustic differences between vowels and consonants are universal across all languages, it is hypothesized that, as in English, vowels in Mandarin would make a greater contribution to sentence intelligibility than consonants. However, the presentation of vowels in Mandarin provides listeners with more information than with "just" vowel information as is the case for English. The vowels also carry the main cues (e.g., F0 contour) about the lexical tones, and vowel and tone information together are likely to restrict lexical competition more than vowel information alone can do in other languages. Hence we hypothesize that compared to the intelligibility advantage of vowels in English, the vowel-only sentences might yield a more pronounced intelligibility advantage in Mandarin. In addition, to study the perceptual contributions of the co-articulation information across C-V boundaries and the information at vowel centers, the present work will examine the effects of adding various amounts of initial vowel portions to consonants or deleting various amounts of vowel information from both onsets and offsets of vowels to the intelligibility of Mandarin sentences.

2. Methods

2.1 Subjects and materials

Twenty (nine male and 11 female) young NH native Mandarin listeners were paid to participate in the experiment. The participants' ages ranged from 19 to 32 yr with majority being students from The University of Hong Kong.

The sentence materials were extracted from the Mandarin speech perception (MSP) corpus (Fu *et al.*, 2011). One hundred sentences from the 10 lists of the database

were used in this experiment, and each sentence contained seven monosyllabic words. The distribution of vowels, consonants, and tones within each list was phonetically balanced and followed their statistical distribution in around 3500 commonly used words in spoken Mandarin (Fu *et al.*, 2011). Note that there were 35 vowels [i.e., 6 simple vowels (/a/, /o/, /e/, /i/, /u/, /ü/), 13 complex vowels (/ai/, /ei/, /ao/, /ou/, /ia/, /ie/, /iao/, /iou/, /ua/, /uo/, /uai/, /uei/, /üe/), and 16 compound nasal vowels (/an/, /en/, /ang/, /eng/, /ong/, /ian/, /in/, /iang/, /ing/, /ion/, /uan/, /uen/, /uang/, /ueng/, /üan/, /ün/), see Yin and Felley (1990)] and 21 consonants [i.e., 2 nasals (/m/, /n/), 6 plosives (/b/, /p/, /d/, /t/, /g/, /k/), 6 affricates (/z/, /c/, /zh/, /ch/, /j/, /q/), 6 fricatives (/f/, /s/, /sh/, /r/, /x/, /h/), and 1 lateral (/l/)] in the MSP sentences, and all vowels and consonants were used according to the international standard *Scheme of the Chinese Phonetic Alphabet* (Yin and Felley, 1990). Vowel-consonant boundaries for the 35 vowels and 21 consonants in the MSP sentences were specified in two steps: (1) First labeled manually by an experienced phonetician based on the highly salient and abrupt acoustic landmarks (e.g., onset of F0 contour) observed in the spectrograms shown by PRAAT (a computer software for the acoustic analysis of speech), and (2) later verified by another experienced phonetician (the detected disagreements between the two phoneticians were manually corrected). The vowel/consonant boundaries for the MSP sentences can be downloaded at the website: http://www.speech.hku.hk/MSP_VC_phn/MSP_VC_phn.html.

2.2 Signal processing

Two signal processing strategies were used to create the testing stimuli in this study. The first strategy preserved the whole consonants and some proportion of vowel transitions with the remaining vowel portions replaced by noise. This type of stimulus is denoted as $C+VP_p$, where p is a factor controlling the proportion of vowel transition relative to the vowel duration. The second strategy preserved various portions of vowel centers but replaced the consonants and some vowel portions from vowel onset and offset by noise. This type of stimulus is denoted as V-VP_p, where p is a factor controlling the proportion of vowel segment to be replaced by noise at two edges (i.e., onset and offset). When implementing the noise-replacement paradigm in the preceding two strategies, all those segments within a sentence were substituted by a speech-shaped noise scaled to -16 dB relative to the level of the intact sentence (Fogerty and Kewley-Port, 2009).

For the $C+VP_p$ strategy, five vowel proportions were chosen, including $p=0$, 0.1, 0.2, 0.3, and 0.4, to yield 5 C + VP conditions. For the V-VP_p strategy, the same vowel proportions were chosen to be deleted from two vowel edges (i.e., onset and offset), yielding 5 V-VP conditions preserving 100%, 80%, 60%, 40%, and 20% of centered vowel segments, respectively. Figure 1 shows the schematic of these 10 C + VP and V-VP conditions. Note that when the proportion factor p equals to 0, the C + VP condition returns to the consonant-only (or C-only) condition, and the V-VP condition returns to the vowel-only (or V-only) condition. Figure 1 shows that the C + VP conditions only preserve the initial vowel transitions because the Mandarin words are characterized by their monosyllabic CV structure. The V-VP conditions in Fig. 1 delete both initial and final transitions as we are intended to investigate the contribution of vowel centers to Mandarin sentence intelligibility. Note that the durational amount that is changed for a given p value is not the same between the $C+VP_p$ and V-VP_p conditions.

2.3 Procedure

The experiment was conducted in a sound-proof booth, and stimuli were played to listeners through a circumaural headphone at a comfortable listening level. All participants attended a practice session (i.e., with feedback of sentence meaning) to listen to 40 noise-replaced sentences before the experiment, so as to get familiar with the experiment procedure and the noise-replaced sentences. Each listener participated in a total of 10 (=2 strategies $\times$ 5 values of the proportion factor p) testing conditions in

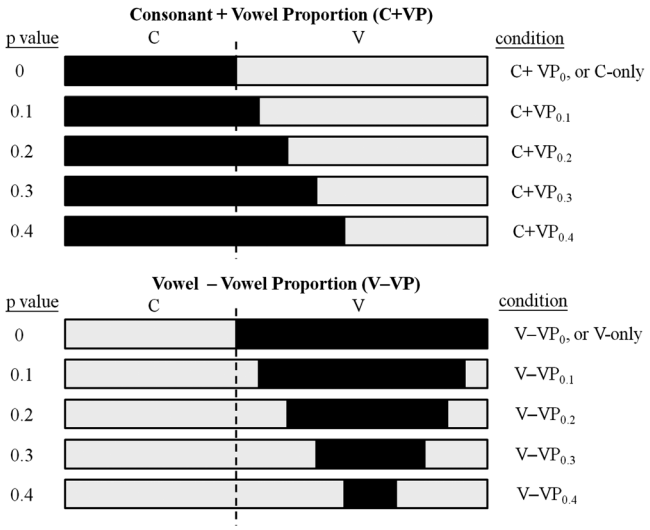

Fig. 1. Schematic of the C + VP and V-VP noise-replacement conditions for a Mandarin word. Dashed line displays the boundary between consonant and vowel. Black bars indicate the speech portions presented, and white bars indicate noise-replaced portions.

Mandarin sentence recognition test. The order of the 10 testing conditions was randomized across listeners. There were 10 sentences per condition, and no sentence was repeated across all conditions. Participants were allowed to listen to each stimulus for three times at maximum and were instructed to repeat all the words they could recognize. Sentence intelligibility score was calculated by dividing the total number of correctly recognized words by the total number of words in each testing condition.

3. Results

Figure 2 shows the mean scores of Mandarin sentence recognition for all testing conditions. First, it is seen that the intelligibility score of the V-only sentences is 99.0% [in Fig. 2(b)], indicating that listeners could recognize almost all words contained in the vowel-only (with consonants replaced by noise) Mandarin sentences. On the other hand, the intelligibility score of the C-only sentences is 34.1% [in Fig. 2(a)], which is significantly lower ($p < 0.05$) than the intelligibility score of the V-only condition (i.e., 99.0%) according to the Wilcoxon signed-rank test. It is interesting to see that the

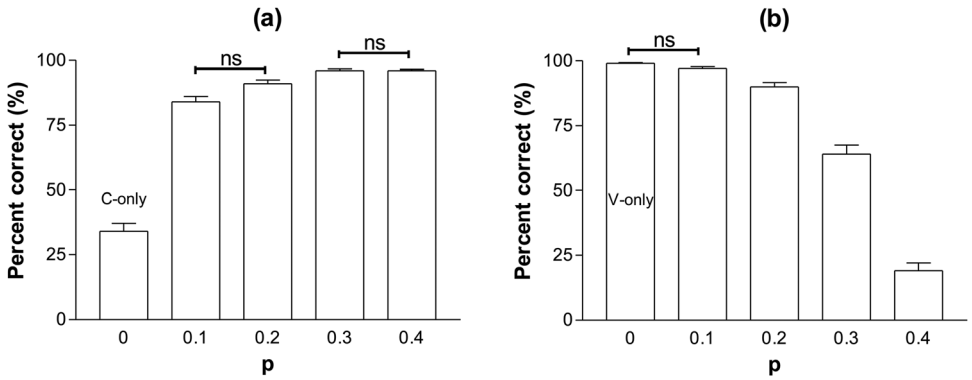

Fig. 2. Mean Mandarin sentence recognition scores as a function of proportion factor p in (a) the C + VP condition and (b) the V-VP condition. The error bars denote 1 standard error of the mean. ns, the difference of paired intelligibility scores is not significant ($p > 0.005$).

performance level of the C-only condition in Mandarin in this study is near to that of the C-only condition in English in Fogerty and Kewley-Port (2009), i.e., 34.1% vs 30.5%. It is the V-only condition where there is a difference between the Mandarin and English sentences. This indicates an increase in the information-carrying capacity of the vowels not just a difference in the relative contribution of consonants and vowels between the two languages.

Figure 2(a) gives the mean intelligibility scores of the $C + VP_p$ conditions as a function of proportion factor p. It is seen that preserving consonants and a small portion (i.e., $p = 0.1$ or 10% VP) of vowel transition would significantly increase the ineligibility score from 34.1% in the $C + VP_0$ (or C-only) condition to 83.8% in the $C + VP_{0.1}$ condition. Statistical significance was determined by using the percent correct score as the dependent variable, and signal processing condition (controlled by the proportion factor p) as the within-subject factor. One-way analysis of variance (ANOVA) with repeated measures indicated a significant effect of signal processing condition ($p < 0.001$). Multiple paired comparisons with Bonferroni correction were run between the intelligibility scores across the $C + VP$ conditions in Fig. 2(a). The Bonferroni-corrected statistical significance level was set at $p < 0.005$ ($\alpha = 0.05$). Figure 2(a) shows that almost all differences of paired intelligibility scores are significant ($p < 0.005$), except those of two pairs, i.e., [$C + VP_{0.1}$ (83.8%), $C + VP_{0.2}$ (90.4%)] and [$C + VP_{0.3}$ (96.4%), $C + VP_{0.4}$ (96.0%)].

Figure 2(b) shows the mean intelligibility scores of the V-VP_p conditions as a function of proportion factor p. It is seen that deleting 10% vowel portion (i.e., $p = 0.1$) from both edges of a vowel only slightly decreases the intelligibility score from 99.0% in the V-VP_0 (or V-only) condition to 97.4% in the V-$VP_{0.1}$ condition. Statistical significance was determined by using the percent correct score as the dependent variable, and signal processing condition (controlled by the proportion factor p) as the within-subject factor. One-way ANOVA analysis with repeated measures indicated a significant effect of signal processing condition ($p < 0.001$). Multiple paired comparisons with Bonferroni correction were run between the intelligibility scores across the V-VP conditions in Fig. 2(b). The Bonferroni-corrected statistical significance level was set at $p < 0.005$ ($\alpha = 0.05$). Figure 2(b) shows that almost all differences of paired intelligibility scores are significant ($p < 0.005$) except that of one pair, i.e., [V-VP_0 (99.0%), V-$VP_{0.1}$ (97.4%)].

4. Discussion and conclusions

Results in Fig. 2 show that vowels contribute more than consonants to Mandarin sentence intelligibility (i.e., V-only condition versus C-only condition); this is consistent with previous findings obtained in studies with English (e.g., Cole *et al.*, 1996; Kewley-Port *et al.*, 2007). This suggests that the intelligibility advantage of vowels is not restricted to English but also found in Mandarin, which is a tonal language. This may be attributed to the fact that the characteristics inherent in the production of vowels and consonants are universal or not language-specific, so that the intelligibility advantage of vowels over consonants can be achieved in different languages. However, the present results also suggest that other language-specific characteristics (e.g., syllable structure) may help to augment the impact of these "universal" characteristics. Further studies with other languages are still warranted to confirm this.

The present work also shows a notable difference in the ratio of segmental contributions to sentence intelligibility between English and Mandarin. In English, vowels were found to have a 2:1 intelligibility advantage over consonants (e.g., Cole *et al.*, 1996; Kewley-Port *et al.*, 2007). However, Mandarin vowels were found to have a 3:1 intelligibility advantage over consonants in this study (i.e., 99.0% vs 34.1%). This improved intelligibility advantage of vowels over consonants in Mandarin may be attributed to the following three primary factors: (1) The lexical tone information conveyed by the F0 contour in vowel segments, (2) the greater proportion of vowels compared to consonants in the phonemic inventory, and (3) vowels' accounting for a greater proportion of the total sentence duration.

First, tone information is important for lexical contrasts in Mandarin. Perceptual cues for tone identification such as F0 contour, amplitude contour, and vowel duration are mainly located in vowel segments (e.g., Howie, 1976; Chen and Loizou, 2011). Hence, vowels in Mandarin may have a more important role in speech intelligibility because of their additional role of carrying lexical tones. On the other hand with the absence of consonant clusters, consonants are of less importance for maintaining phonemic contrasts in Mandarin than in English. Second, there are 21 consonants and 35 vowels in Mandarin according to the classification used by the MSP corpus (Fu *et al.*, 2011), but 32 consonants and 20 vowels in English according to the classification used by Kewley-Port *et al.* (2007) and Fogerty and Kewley-Port (2009). With more vowels but fewer consonants in Mandarin, Mandarin vowels seem to be more important than English vowels for phonemic contrasts. As a result, they may yield a greater intelligibility advantage than English vowels. Third, in English, the percentage of vocalic intervals across the entire sentence was found to be smaller than that of consonantal intervals with almost 10% (e.g., Fogerty and Kewley-Port, 2009). However, the average duration of vowels across the entire sentence was 66.3%, while that of consonants was 25.9% in Mandarin. This predominantly long duration of vowels across sentences may also partially account for their large contribution to Mandarin sentence intelligibility.

To some extent, the present findings (i.e., the intelligibility advantage of vowels over consonants, and the language-specific intelligibility advantage of vowels) are consistent with previous results regarding to the cross-language similarities and differences in the contribution of vowels and consonants in speech perception. Cutler *et al.* (2000) found that vowels constrained lexical selection less tightly than consonants did, independent of language-specific phoneme repertoire and of relative distinctiveness of vowels. Bonatti *et al.* (2005) suggested that consonants and vowels were more tied to word identification and grammar, respectively, in continuous speech. In addition, speech materials and language-specific vowel-consonant ratio would also impact the relative contribution of vowels and consonants in speech perception (e.g., Owren and Cardillo, 2006). One of the reasons for the difference between isolated-word (e.g., Owren and Cardillo, 2006) and sentence perception is that the sentence context itself strongly constrains lexical competition. Hence listeners may retrieve or predict sentences' meaning by using *a priori* knowledge, language experience and contextual cues involved in a top-down processing. Note that future study is needed to investigate the effects of neighborhood density and transitional probabilities to speech perception in the context of noise-replacement manipulation.

For the C + VP testing conditions, results in this study share common finding with those obtained with English sentences by Fogerty and Kewley-Port (2009), i.e., providing transitional information across C-V boundaries could increase the sentence intelligibility under the consonant-dominant conditions (i.e., C + VP). However, unlike the pattern found in the English study by Fogerty and Kewley-Port (2009) where the intelligibility increase was linear as a function of the added vowel proportion, the trend observed in this study was non-linear, as seen in Fig. 2(a). Sentence intelligibility increased significantly when a small portion of vowel transition (i.e., $p = 0.1$ or 10% VP) was added to the C-only condition. Then the increase in intelligibility was not significant until 20% more VP was added to the $C + VP_{0.1}$ condition (i.e., in the $C + VP_{0.3}$ condition). This suggests that the transitional information across C-V boundaries is rich with acoustic information of vowels and is redundant to the intelligibility information at vowel centers. As a result, there was a large intelligibility increase (i.e., from 34.1% in the $C + VP_0$ condition to 83.8% in the $C + VP_{0.1}$ condition) when 10% VP was added to the C-only condition but the increase became smaller when more VP was added. This hypothesis was also supported by the finding that sentence intelligibility did not decrease significantly when 20% vowel portion was deleted from vowel onset and offset [i.e., $p = 0.1$ in Fig. 2(b)].

It is seen in Fig. 2(b) that the Mandarin sentences in the vowel-only condition have a high recognition accuracy (up to 99.0%), and the intelligibility does not decrease significantly even when 20% vowel portion is removed [i.e., $p = 0.1$ in Fig. 2(b)]. This

finding is different from that in English studies because the vowel-only condition never had more than 90% intelligibility in those studies (e.g., Cole *et al.*, 1996; Kewley-Port *et al.*, 2007). This suggests that the vowel-only condition is sufficient for listeners to identify Mandarin sentences with consonants replaced by noise. Even when two 10% vowel portions were deleted from both edges of a vowel [i.e., $p = 0.1$ in Fig. 2(b)], the intelligibility did not significantly decrease. Furthermore, the intelligibility of Mandarin sentences could still be maintained at 89.6% even when only 60% vowel portion was preserved at vowel centers [i.e., $p = 0.2$ in Fig. 2(b)], indicating that vowel centers have an important contribution to Mandarin sentence intelligibility. However, these vowel centers should not be viewed as "steady-state," as significant dynamic information is conveyed, particularly due to lexical tone.

In conclusion, the present work assessed the perceptual contributions of vowels and consonants to Mandarin sentence intelligibility based on the noise-replacement paradigm. Consistent with previous findings on the segmental contributions to sentence intelligibility in English, the present results showed that vowels contributed more than consonants to Mandarin sentence intelligibility, rendering a remarkable 3:1 intelligibility advantage of the vowel-only sentences over the consonant-only sentences. This advantage is, however, larger than that obtained with English sentences, suggesting that vowels may have a greater contribution to sentence intelligibility in Mandarin than in English. In addition, although providing information redundant to contributions from vowel centers, a small portion of vowel-consonant boundary transition (i.e., 10% vowel portion) may significantly improve the intelligibility of the consonant-only Mandarin sentences.

Acknowledgments

This research was supported by Faculty Research Fund (Faculty of Education) and Seed Funding for Basic Research, The University of Hong Kong, and by General Research Fund, The Hong Kong Research Grants council. This study was the basis for the Bachelor's thesis of the third author (E.Y.W.W.).

References and links

Bonatti, L., Peña, M., Nespor, M., and Mehler, J. (**2005**). "Linguistic constraints on statistical computations: The role of consonants and vowels in continuous speech processing," Psychol. Sci. **16**, 451–459.

Chen, F., and Loizou, P. (**2011**). "Predicting the intelligibility of vocoded and wideband Mandarin Chinese," J. Acoust. Soc. Am. **129**, 3281–3290.

Cole, R., Yan, Y., Mak, B., Fanty, M., and Bailey, T. (**1996**). "The contribution of consonants versus vowels to word recognition in fluent speech," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 853–856.

Cutler, A., Sebastian-Galles, N., Soler-Vilageliu, O., and Van Ooijen, B. (**2000**). "Constraints of vowels and consonants on lexical selection: Cross-linguistic comparisons," Mem. Cognit. **28**, 746–755.

Fogerty, D., and Kewley-Port, D. (**2009**). "Perceptual contributions of the consonant-vowel boundary to sentence intelligibility," J. Acoust. Soc. Am. **126**, 847–857.

Fu, Q. J., Zhu, M., and Wang, X. S. (**2011**). "Development and validation of the Mandarin speech perception test," J. Acoust. Soc. Am. **129**, EL267—EL273.

Howie, J. M. (**1976**). *Acoustical Studies of Mandarin Vowels and Tones* (Cambridge University Press, Cambridge, UK).

Kewley-Port, D., Burkle, T. Z., and Lee, J. H. (**2007**). "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.

Li, Z., Tan, E. C., McLoughin, I., and Teo, T. T. (**2000**). "Proposal of standards for the intelligibility tests of Chinese speech," IEE Proc. Vision Image Signal Process. **147**, 254–260.

Owens, E., Talbot, C. B., and Schubert, E. D. (**1968**). "Vowel discrimination of hearing-impaired listeners," J. Speech Hear. Res. **11**, 648–655.

Owren, M. J., and Cardillo, G. C. (**2006**). "The relative roles of vowels and consonants in discriminating talker versus word meaning," J. Acoust. Soc. Am. **119**, 1727–1739.

Yin, B., and Felley, M. (**1990**). *Chinese Romanization: Pronunciation and Orthography* (Sinolingua, Beijing, China).