Published in final edited form as:

Speech Commun. 2017 September ; 92: 125–131. doi:10.1016/j.specom.2017.06.008.

The development and validation of the Closed-set Mandarin Sentence (CMS) test

Duo-Duo Tao^a, Qian-Jie Fu^b, John J. Galvin III^b, and Ya-Feng Yu^a

^aDepartment of Ear, Nose, and Throat, The First Affiliated Hospital of Soochow University, Suzhou 215006, China

^bDepartment of Head and Neck Surgery, David Geffen School of Medicine, UCLA, Los Angeles, CA, USA

Abstract

Matrix-styled sentence tests offer a closed-set paradigm that may be useful when evaluating speech intelligibility. Ideally, sentence test materials should reflect the distribution of phonemes within the target language. We developed and validated the Closed-set Mandarin Sentence (CMS) test to assess Mandarin speech intelligibility in noise. CMS test materials were selected to be familiar words and to represent the natural distribution of vowels, consonants, and lexical tones found in Mandarin Chinese. Ten key words in each of five categories (Name, Verb, Number, Color, and Fruit) were produced by a native Mandarin talker, resulting in a total of 50 words that could be combined to produce 100,000 unique sentences. Normative data were collected in 10 normal-hearing, adult Mandarin-speaking Chinese listeners using a closed-set test paradigm. Two test runs were conducted for each subject, and 20 sentences per run were randomly generated while ensuring that each word was presented only twice in each run. First, the level of the words in each category were adjusted to produce equal intelligibility in noise. Test-retest reliability for word-in-sentence recognition was excellent according to Cronbach's alpha (0.952). After the category level adjustments, speech reception thresholds (SRTs) for sentences in noise, defined as the signal-to-noise ratio (SNR) that produced 50% correct whole sentence recognition, were adaptively measured by adjusting the SNR according to the correctness of response. The mean SRT was -7.9 (SE=0.41) and -8.1 (SE=0.34) dB for runs 1 and 2, respectively. The mean standard deviation across runs was 0.93 dB, and paired t-tests showed no significant difference between runs 1 and 2 ($p=0.74$) despite random sentences being generated for each run and each subject. The results suggest that the CMS provides large stimulus set with which to repeatedly and reliably measure Mandarin-speaking listeners' speech understanding in noise using a closed-set paradigm.

Corresponding author. Ya-Feng Yu, Department of Ear, Nose, and Throat, The First Affiliated Hospital of Soochow University, 188 Shizi Street, Suzhou 215006, China. entdtao@163.com.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Keywords

speech perception; speech reception threshold; Mandarin speech recognition; speech intelligibility in noise; matrix sentence test

Introduction

Open-set sentence recognition tests administered in quiet or in noise are often used to diagnose hearing health status. One limitation to open-set testing is that a hearing professional must administer and score the test. Another limitation is that materials available for clinical or research testing is often limited to several hundred sentences, which can be problematic for research when repeated testing is needed (e.g., longitudinal studies, comparisons of signal processing strategies, etc.). One alternative to open-set sentence testing is the matrix sentence test, which was originally developed by Hagerman (1982) in Swedish, and further developed for many languages, including American English, British English, Dutch, German, French, Italian, Spanish, Turkish, and Russian (see Kollmeier et al., 2015 for a review). A matrix sentence test was also recently developed to measure perception of sung speech by Crew et al. (2015, 2016). A matrix test typically consists of five-word sentences that have the same syntax, but are not necessarily meaningful. There are generally five categories in fixed order (e.g., Name, Verb, Number, Adjective, and Object), each of which has 10 items, such that multiple sentences can be generated by randomly selecting a word from each category. A 5×10 matrix can generate 100,000 unique sentences, making the matrix test a powerful tool with which to repeatedly measure speech perception in clinical and research settings.

When adaptively measuring speech reception thresholds (SRTs), defined as the signal-to-noise ratio (SNR) that produces 50% correct sentence recognition in noise, is important for the slope of performance-intensity (PI) function to be relatively steep around the 50% correct point (Plomp and Milpen, 1979). If sentences are too difficult, or vary in terms of difficulty, the slope of the PI function will become shallower and the SRT will be more variable and less accurate. Context cues and predictability can aid word-in-sentence recognition for normal-hearing (NH) and hearing-impaired (HI) listeners (e.g., Pisoni et al., 1987; Boothroyd and Nitttrouer, 1998; Kong et al., 2015). The fixed syntax in matrix sentence tests may improve predictability, which would in turn sharpen the slope of PI function in noise. With low-predictability sentences, the slope will likely be shallower, and may not allow for 100% correct recognition in quiet, especially for HI listeners. Matrix sentence tests are often administered using a closed-set paradigm, thereby reducing some uncertainty associated with open-set testing. Thus, the familiar words, fixed syntax, and closed-set paradigm used in matrix sentence tests may improve predictability and reduce uncertainty, resulting in steep PI functions and better stability for adaptively measured SRTs.

Although Mandarin Chinese is the most widely spoken language in the world, there is currently no “gold standard” sentence recognition test. Relatively few speech materials been developed with which to evaluate intelligibility (Wong et al., 2007, Fu et al., 2011; Xi et al, 2012), and the speech materials are limited to a few hundred sentences, making it difficult to

repeatedly measure performance within subjects. As such, a matrix-like Chinese sentence test, given its potential to generate 100,000 unique sentences, would benefit both clinical and research practice. To date, there is no such test available for Mandarin Chinese. Different from previous matrix tests developed for Western and European languages (Kollmeier et al., 2015), Mandarin Chinese is a tonal language in which fundamental frequency (F0) cues are important for lexical tone recognition (Liang, 1963; Lin, 1988), and lexical tone recognition is important for sentence recognition (Fu et al., 1998). There are four major tonal patterns in Mandarin Chinese, characterized by their F0 contour: Tone 1 (flat F0), Tone 2 (rising F0), Tone 3 (falling-rising F0), and Tone 4 (falling F0). Depending on the tone, the same syllable can have different meanings: /ma/ may mean “mother”, “linen”, “horse”, “scold” for Tones 1, 2, 3, or 4, respectively. A fifth tone (neutral tone or Tone 0) is also occasionally used in Mandarin Chinese.

Intuitively, it seems important to consider the distribution of phonemes found in everyday speech when developing materials for speech testing. Previous matrix sentence tests have largely preserved the natural distribution of phonemes (Ozimek et al., 2010; Hochmuth et al., 2012). Mandarin Chinese poses greater difficulty than for non-tonal languages because the distribution of lexical tones must also be considered. Chinese hearing-impaired listeners, especially cochlear implant (CI) users may have poorer sensitivity to F0 information (and therefore lexical tones) than do NH listeners (Fu et al., 1998). As such, it is critical to preserve the natural distribution of phonemes and tones when developing Mandarin Chinese speech materials to evaluate speech performance in NH, HI, and CI listeners.

Here, we present the development and validation of the Closed-set Mandarin Sentence (CMS) test. The main goal for developing the CMS were to create a large stimulus set of sentences that could be used to reliably and repeatedly used to measure Mandarin speech understanding in noise using a closed-set paradigm, and that performance would be comparable to existing Mandarin sentence recognition tests. The CMS test materials were developed similarly to previous matrix tests (Akeroyd et al., 2015; Kollmeier et al. 2015), but with some notable differences. Similar to previous matrix tests, a fixed structure was used to generate sentences (Name, Verb, Number, Adjective, Object). Familiar words were used to allow for testing in adults and children. CMS materials were selected to reflect the natural distribution of phonemes (and in the CMS, lexical tones) found within the native language. The levels of words were adjusted to maintain equal intelligibility across categories. However, the CMS test differs from previous matrix tests in several ways. For example, Akeroyd et al. (2015) recommends that matrix test stimuli should allow for generation of semantically nonsensical sentences. In the CMS, words were selected to produce similar semantic sense, as the possibility of nonsensical sentences could affect intelligibility during testing. In the CMS test, the Adjective was intentionally fixed to be a Color and the Object to be a Fruit, as this was thought to improve semantic stability across randomly generated sentences. The CMS materials also contain the same number of syllables for each word within a category, while previous matrix sentence tests can be inconsistent in terms of the number of syllables across words within a category (Hochmuth et al., 2012). The words within each category for the CMS were recorded in isolation. In previous matrix tests (Kollmeier et al., 2015), words were recorded in tandem across categories to preserve co-articulation, which was thought to improve the naturalness of the

generated sentences. In the CMS, auxiliary words were included within the categories to reduce co-articulation effects across categories. To our knowledge, the CMS is the first closed-set sentence test developed for a tonal language. In this study, SRTs in noise for words-in-sentences and for whole sentences were measured in NH listeners using the CMS test materials.

2. Methods

2.1 Subjects

Ten adult native Mandarin-speaking Chinese NH listeners participated in the study (4 males and 6 females; mean age = 26 years; range = 22 – 30 years old). In compliance with ethical standards for human subjects, written informed consent was obtained from all participants before proceeding with any of the study procedures. This study was approved by Institutional Review Board in The First Affiliated Hospital of Soochow University, Suzhou, China. For all NH subjects, pure-tone thresholds did not exceed 20 dB HL at octave frequencies from 125 to 8000 Hz.

2.2 CMS speech materials

In developing the CMS speech materials, the first criterion was that the words should all be familiar and widely used in daily life. The second criterion was that the randomly generated sentence sets should be balanced in terms of the distribution of vowels, consonants, and lexical tones. Similar to previous matrix tests (Kollmeier et al., 2015), 10 commonly used words were selected for each of 5 categories: Name, Verb, Number, Color, and Fruit. The Color (adjective) and Fruit (object) categories were strictly specified to maintain semantic stability across the randomly generated sentences. This approach was different from that of some previous matrix studies (Kollmeier et al., 2015), in which the adjectives and objects were not strictly specified, leading to sentence constructs that might differ greatly in terms of semantic predictability. The distribution of vowels, consonants, and lexical tones from 3500 commonly used Mandarin Chinese words (Tang, 1995) was used to generate the distribution for the CMS. To best match the distribution from Tang (1995), it was necessary to adjust the distribution within the CMS, given the limited number of total key words (50). The number of targeted vowels and consonants was allowed to vary by ± 1 and the number of targeted tones was allowed to vary by ± 2 , relative to Tang (1995).

Table 1 shows the the words selected for each of the 5 categories, chosen according to the distribution constraints. The target words for Name, Verb, Number, and Color were monosyllables; the target words for Fruit were disyllables. Within each category, the number of syllables was kept constant. Auxiliary monosyllables were appended to the beginning of the Name target words and to the end of the Verb and Number target words; two monosyllables were added to the end of the Color target words. These auxiliary monosyllables were included to improve the naturalness of the randomly generated sentences, as these are commonly used in conversational Mandarin. The auxiliary words also helped to minimize co-articulation effects between subsequent syllables across categories. Note that the co-articulation between subsequent syllables was always preserved within categories since all the syllables in each word were produced naturally. Thus, the Name,

Verb, Number and Fruit categories each contained 2 syllables per stimulus and the Color category contained 3 syllables per stimulus.

Table 2 shows the distribution of consonants, vowels, and tones from Tang (1995) and from a randomly generated list of 10 sentences (all 50 words used). Pearson's Chi-square tests showed no significant difference between the CMS and Tang (1995) in the distributions for consonants (chi-square = 82.8; $df = 80$, $p = 0.394$), vowels (chi-square = 162.0; $df = 135$, $p = 0.056$), or tones (chi-square = 20; $df = 16$, $p = 0.220$).

After the 50 words were selected according to the above criteria, the stimulus set was recorded. All words were produced by a native Mandarin male speaker with standard Mandarin pronunciation. The speaker was instructed to pronounce each of the two- or three-syllable items within each category at a moderate speaking rate with a natural intonation and accent. Note that this approach was different from that used in previous matrix studies, in which pairs of words across categories were produced together to preserve co-articulations between words (Akeroyd et al., 2015; Kollmeier et al., 2015). Stimuli were recorded in a sound-treated booth using a high quality microphone (Neumann U87), connected to an audio input device (Edirol UA-25EX; sampling rate: 44.1 kHz; 24-bit resolution), connected to a computer. The original recordings were saved directly onto a hard drive.

Recordings were edited using Adobe Audition software. First, all recordings contained were high-pass filtered (80 Hz cut-off frequency). Next, stimuli (key word plus auxiliary word for the Name, Verb, Number, and Color categories; keyword for the Fruit category) were carefully edited such that there was minimal onset and offset silence around each stimulus. This editing was important to maintain minimal and consistent pauses between stimuli when they were strung into sentences (i.e., to avoid cues associated with words having an overly long period of silence). Next, the amplitude of all stimuli was normalized to have the same root-mean-square (RMS) level (65 dB). Test-specific noise was created by generating white noise then band-pass filtering according to the long-term spectrum averaged across all stimuli. Fig. 1 shows the speech spectrum averaged across all CMS words and the test-specific noise used during the experiment with CMS stimuli. All speech and noise stimuli were then down-sampled to 22,050 Hz with 16-bit resolution.

2.3 CMS test procedures

Before formally testing CMS sentence recognition in noise, the level of stimuli within each category was adjusted to produce equal intelligibility in noise, as in previous matrix tests (Ozimek et al., 2010; Hochmuth et al., 2012; Akeroyd et al., 2015; Kollmeier et al., 2015). To determine the amount of level adjustment needed for each category, word-in-sentence recognition was measured using a closed-set procedure for 6 signal-to-noise ratios (SNRs): 0, -3, -6, -9, -12, and -15 dB. As noted above, all stimuli were normalized to have the same RMS amplitude (65 dB); the level of the noise for each target SNR was calculated according to the long-term RMS of the randomly generated target sentence (which was the same for all target sentences). To maintain the target distribution of consonants, vowels, and tones, blocks of 10 randomly generated sentences were created in which all 50 words were used (i.e., each word presented once in each block). Two blocks were tested for each run and two runs were conducted for each SNR. As such, all 50 words were presented twice, but the

sentences were randomly generated and theoretically different across subjects and test runs. During testing, a sentence was randomly generated and presented at the target SNR; the onset of the noise was 500 ms before the onset of the first category stimulus and offset of the noise was 500 ms after offset of the last category stimulus. The subject responded by clicking on one of the buttons for each of the categories (Fig. 2), after which a new sentence was generated and presented to the subject. Performance was scored in terms of percent correct for each category. Sigmoid functions were fit to the performance-intensity (PI) functions for each category across the target SNRs; the SNR that produced 50% correct recognition was calculated for each category (speech reception threshold, or SRT_{cat}). Level adjustments were then made according to the difference in the SRT_{word} for a given category and the mean SRT_{word} across all categories. As such, these level adjustments produced similar intelligibility across categories.

After the category-level adjustments, SRTs for sentence recognition in noise were adaptively measured using a 1-up/1-down closed-set procedure, converging on the SNR that produced 50% correct whole sentence recognition in noise (Rule 1 from Chan et al., 2008). In this procedure, a correct response required all words in a sentence to be recognized; if the subject failed to correctly recognize even one word in a sentence, the response was scored as incorrect. Again, to maintain the target distribution of consonants, vowels, and tones, blocks of 10 randomly generated sentences were created in which all 50 words were used; two blocks were tested for each run. During testing, a sentence was randomly generated and presented at the target SNR, which was calculated according to the long-term RMS amplitude of the sentence. The onset of the noise was 500 ms before the onset of the first category stimulus and offset of the noise was 500 ms after offset of the last category stimulus. The subject responded by clicking on one of the buttons for each of the categories (Fig. 2), after which a new sentence was generated and presented to the subject. If the subject responded correctly for all categories, the SNR was reduced; if the subject did not respond correctly for all categories, the SNR was increased. The initial SNR was 0 dB (the highest fixed SNR tested for the previous word-in-sentence recognition), the initial step size was 4 dB, and the final step size was 2 dB (as in Plomp and Milpen, 1979). For each run, the SRT was calculated across the final 6 reversals in SNR; the mean SRT for each subject was calculated across test runs. If the test run failed to obtain 6 reversals within the 20 sentences presented, the test run was discarded. Note that 6 reversals were obtained within each test run in all subjects.

For all testing, speech and noise were presented in a sound field via a single loudspeaker; subjects were seated directly facing the loudspeaker at 1 m distance. Speech was presented at 65 dBA. Prior to formal testing, subjects were allowed to preview all the CMS words and to listen to a few CMS sentences in quiet to familiarize them with the test procedures.

3. Results

3.1 SRT_{cat} and category-level adjustments

Fig. 3 shows the mean percent correct (across subjects and the two test runs) for each category (different symbols) as a function of SNR; the lines show sigmoid fits to the data for each category, and the slope and 50% correct values for each category (SRT_{cat}) are shown in

the legend. A paired t-test showed no significant difference across test runs ($p=0.668$), so data were combined across test runs for each subject. For the sigmoid fits, r^2 values were 0.99 and p values were 0.001. The standard deviation across test runs was also calculated for each subject for each category and SNR; the mean standard deviation across all subjects and test runs was 5.48 dB. Test-retest reliability was calculated across all SNRs and categories; Cronbach's $\alpha = 0.952$, indicating excellent test-retest reliability. The SRT_{cat} was poorest with Name and best with Number. A two-way repeated-measures analysis of variance was performed on the data shown in Fig. 3, with category (Name, Verb, Number, Color, Fruit) and SNR (0, -3, -6, -9, -12, -15) as factors. Results showed significant effects for category [$F(4,180)=39.3$, $p<0.001$] and SNR [$F(5,180)=296.4$, $p<0.001$]; there was a significant interaction [$F(20,180)=6.7$, $p<0.001$]. Post-hoc Bonferroni pairwise comparisons showed that performance was significantly poorer with Name than with the remaining categories ($p<0.05$ in all cases), significantly poorer with Fruit relative to Number and Verb ($p<0.05$ in both cases), and significantly poorer with Color relative to Number ($p<0.05$).

The mean SRT_{cat} (across all categories) was -12.86 dB. Level adjustments were applied to each category relative to the mean SRT_{cat} across all categories. Accordingly, the level adjustments applied to each word in a category was +1.16, -0.44, -1.14, +0.16, and +0.26 dB for the Name, Verb, Number, Color, and Fruit categories, respectively.

3.2 Adaptively measured SRTs for whole sentence recognition in noise

Fig. 4 shows boxplots of SRTs for Run 1 (grey boxes) and Run 2 (white boxes) collected with and without the category level adjustments. Note that one subject was unable to complete this portion of the experiment, so data is shown for 9 subjects. With the category level adjustment, the mean SRT (across subjects) was -7.9 dB ($SE=0.41$) and -8.1 ($SE=0.34$) dB for Runs 1 and 2, respectively; the mean difference across test runs was 0.2 dB. A paired t-test showed no significant difference between test runs ($p=0.74$). The standard deviation was calculated between runs for each subject; the mean standard deviation (across subjects) between test runs was 0.93 dB.

With no level adjustment, the mean SRT (across subjects) was -8.9 dB ($SE=0.32$) and -7.8 ($SE=0.41$) dB for Runs 1 and 2, respectively; the mean difference across test runs was 1.1 dB. A paired t-test showed a significant difference between test runs ($p=0.04$). The standard deviation was calculated between runs for each subject. The standard deviation was calculated between runs for each subject; the mean standard deviation (across subjects) between test runs was 1.18 dB.

Discussion

The present data suggest that the CMS test may be a valuable tool with which to measure speech intelligibility in noise using a closed-set paradigm. The CMS test demonstrated good test-retest reliability for words-in sentences, and there was no significant difference in SRTs for whole sentence recognition in noise between test runs after the category level adjustments. Below we discuss the results in greater detail.

Word-in sentence recognition in noise significantly differed across categories (Fig. 3), with the poorest SRT_{cat} for Name and the best SRT_{cat} for Number. For the Spanish matrix test, the poorest SRT_{cat} was for Object and the best SRT_{cat} was for Name (Hochmuth et al., 2012); for the Polish matrix tests, the poorest SRT_{cat} was for Verb and the best SRT_{cat} was for Number (Ozimek et al., 2010). Idiosyncrasies in word selection as well as overall language differences may explain differences in relative intelligibility for specific categories across studies. Nevertheless, the range of level adjustments across categories was comparable to that in Ozimek et al., (2010), but smaller than in Hochmuth et al. (2012). It is important to note that for all tests in this study, two blocks of 10 randomly generated sentences (comprised of all 50 words, which were presented only once within a block) were tested for each subject and each test run. Because sentences were randomly generated for each block, no two 10-sentence blocks were (theoretically) the same within a run, across runs, and across subjects. The slopes of the sigmoid functions fit to the data in Figure 4 were comparable to those reported in Hochmuth et al. (2012), but shallower than in Puglisi et al. (2015).

For word-in-sentence recognition in noise (Fig. 3), which was measured for fixed SNRs, Cronbach's alpha showed excellent test-retest reliability across test Runs 1 and 2. For adaptively measured whole sentence recognition in noise with the category level adjustment (Fig. 4), the standard deviation across subjects was 1.2 dB for Run 1 and 1.0 dB for Run 2, and the mean standard deviation across test runs was 0.93 dB. As such, the inter-subject variability with runs was comparable to the intra-subject variability across runs. Under these circumstances, Cronbach's alpha test reliability would not be an appropriate measure. The mean standard deviation across test runs was comparable to that of Plomp and Milpen (1979), who suggested that a mean standard deviation ≤ 1 dB indicated good reliability across test lists. Thus, the present CMS materials appear to exhibit good reliability across the randomly generated test lists for both word-in-sentence and whole sentence recognition in noise.

With the category level adjustments, the mean adaptively measured CMS SRT for whole sentence recognition in noise (-7.9 dB) was well within the range reported by Kollmeier et al. (2015) for matrix tests with non-tonal languages (-6.2 to -9.7 dB). The present mean SRT were somewhat better than those reported for BKB-like Mandarin sentences in noise (-6 dB; Xi et al., 2012) and MHINT sentences in noise (-4.3 dB; Wong et al., 2007), possibly due to differences in test paradigm (closed- versus open-set), speech materials, talkers, etc.

While SRTs with the CMS materials were similar to those in previous matrix tests with non-tonal languages (Kollmeier et al., 2015), and while the CMS shares some of the approaches towards test development and validation with previous matrix tests, there are some notable differences. In previous matrix tests (Akeroyd et al., 2015; Kollmeier et al., 2015, successive words across categories were recorded together, then edited to preserve co-articulation between words. In the CMS, stimuli within categories were recorded in isolation. However, the auxiliary words used in the Verb, Number, and Color categories ensured that the transitions and co-articulation targets were consistent between categories. The recording and editing methods used in previous matrix studies resulted in several hundred sentences at

most, while the CMS allows for 100,000 sentences. This was an important consideration for the development of the CMS, as the available Mandarin speech tests (Wong et al., 2007; Fu et al., 2011; Xi et al., 2012) offer too few sentences for repeated testing as is often required for research. While the methods used in previous matrix tests may allow for more natural-sounding sentences, it is unclear how this might affect intelligibility in these tests. The somewhat large range in SRTs across languages (−6.2 to −9.7 dB) reported by Kollmeier et al. (2015) suggests that despite efforts to homogenize and refine matrix test development, sizeable differences remain across languages in terms of speech intelligibility in noise.

Conclusions

While the CMS test may not have been developed strictly according to methods described by Akeroyd et al. (2015), it offers the possibility to repeatedly and reliably test Mandarin speech understanding in noise in a closed-set paradigm using a very large stimulus set consisting of familiar words that reflect distribution of vowels, consonants, and lexical tones found in Mandarin Chinese. However, further studies are needed to validate the CMS test at multiple sites with larger number of subjects in targeted populations (e.g., children, hearing-impaired listeners, CI users, etc.).

Acknowledgments

We thank the subjects for their participation in this study. This work was supported by the National Institutes of Health [grant number R01-004993]; the National Natural Science Foundation of China [grant number 81600796]; the Natural Science Foundation of Jiangsu Province for Young Scholar [grant number 20160344]; and the Natural Science Foundation of Jiangsu Province for Colleges and Universities [grant number 16KJB320011].

Acronyms and abbreviations

CI	cochlear implant
CMS	coordinated Mandarin sentence
CRM	coordinate response measure
F0	fundamental frequency
FIST	French intelligibility sentence test
MHINT	Mandarin hearing in noise test
MSP	Mandarin speech perception
NH	normal hearing
ICRA	international collegium of rehabilitative audiology
RM ANOVA	repeated measures analysis of variance
RMS	root-mean-square
SNR	signal-to-noise ratio

SRT speech reception threshold

References

- Akeroyd MA, Arlinger S, Bentler RA, et al. International Collegium of Rehabilitative Audiology (ICRA) recommendations for the construction of multilingual speech tests. *Int J Audiol.* 2015; 54(Suppl 2):17–22. [PubMed: 25922886]
- Boothroyd A, Nitttrouer S. Mathematical treatment of context effects in phoneme and word recognition. *J Acoust Soc Am.* 1988; 84:101–114. [PubMed: 3411038]
- Chan JC, Freed DJ, Vermiglio AJ, et al. Evaluation of binaural functions in bilateral cochlear implant users. *Int. J. Audiol.* 2008; 47:296–310. [PubMed: 18569102]
- Crew JD, Galvin JJ, Fu QJ. Melodic contour identification and sentence recognition using sung speech. *J. Acoust. Soc. Am.* 2015; 138:EL347–351. [PubMed: 26428838]
- Crew JD, Galvin JJ, Fu QJ. Perception of Sung Speech in Bimodal Cochlear Implant Users. *Trends Hear.* 2016; 20doi: 10.1177/2331216516669329
- Fu QJ, Zeng FG, Shannon RV, et al. Importance of tonal envelope cues in Chinese speech recognition. *J. Acoust. Soc. Am.* 1998; 104:505–510. [PubMed: 9670541]
- Fu QJ, Zhu M, Wang X. Development and validation of the Mandarin speech perception test. *J. Acoust. Soc. Am.* 2011; 129:EL267–273. [PubMed: 21682363]
- Hagerman B. Sentences for testing speech intelligibility in noise. *Scand. Audiol.* 1982; 11:79–87. [PubMed: 7178810]
- Hochmuth S, Brand T, Zokoll MA, et al. A Spanish matrix sentence test for assessing speech reception thresholds in noise. *Int. J. Audiol.* 2012; 51:536–544. [PubMed: 22537033]
- Kollmeier B, Warzybok A, Hochmuth S, et al. The multilingual matrix test: Principles, applications, and comparison across languages: A review. *Int. J. Audiol.* 2015; 54(Suppl 2):3–16.A. [PubMed: 26383182]
- Kong YY, Donaldson G, Somarowthu A. Effects of contextual cues on speech recognition in simulated electric-acoustic stimulation. *J. Acoust. Soc. Am.* 2015; 137:2846–2857. [PubMed: 25994712]
- Li Y, Wang S, Su Q, et al. Validation of list equivalency for Mandarin speech materials to use with cochlear implant listeners. *Int J Audiol.* 2016; 14:1–10.
- Liang ZA. The auditory perception of Mandarin Tones. *Acta Phys. Sin.* 1963; 26:85–91.
- Lin MC. The acoustic characteristics and perceptual cues of tones in Standard Chinese. *Chinese Yuwen.* 1988; 204:182–193.
- Ozimek E, Warzybok A, Kutzner D. Polish sentence matrix test for speech intelligibility measurement in noise. *Int. J. Audiol.* 2010; 49:444–454. [PubMed: 20482292]
- Pisoni DB, Manous LM, Dedina MJ. Comprehension of natural and synthetic speech: effects of predictability on the verification of sentences controlled for intelligibility. *Comput. Speech Lang.* 1987; 2:303–320. [PubMed: 23226919]
- Plomp R, Mimpfen AM. Improving the reliability of testing the speech reception threshold for sentences. *Audiol.* 1979; 18:43–53.
- Puglisi GE, Warzybok A, Hochmuth S, et al. An Italian matrix sentence test for the evaluation of speech intelligibility in noise. *Int. J. Audiol.* 2015; 54(Suppl 2):44–50. [PubMed: 26371592]
- Tang YH. Statistical analysis of Mandarin Chinese. *J. Chengde Teachers' Coll. Nat.* 1995:66–76.
- Wong LL, Soli SD, Liu S, et al. Development of the Mandarin Hearing in Noise Test (MHINT). *Ear Hear.* 2007; 28:70S–74S. [PubMed: 17496652]
- Xi X, Ching TY, Ji F, et al. Development of a corpus of Mandarin sentences in babble with homogeneity optimized via psychometric evaluation. *Int. J. Audiol.* 2012; 51:399–404. [PubMed: 22201527]

Highlights

- Here, we present the Closed-set Mandarin Sentence (CMS) test.
- The CMS reflects the distribution of phonemes and lexical tones in Mandarin.
- The CMS can generate 100,000 unique sentences using 50 familiar words.
- The CMS exhibited excellent reliability for testing words-in-sentences in noise.
- The CMS exhibited good reliability for testing whole sentence recognition in noise.

Figure 1.
Speech spectrum averaged across all CMS words (red line) and test-specific noise (black line).

小王	捡了	一个	褐色的	金桔
小赵	卖了	两个	青色的	木瓜
小虎	切了	三个	粉色的	枇杷
小孔	种了	四个	银色的	香蕉
小周	拿了	五个	灰色的	樱桃
小邓	装了	六个	杏色的	草莓
小杜	看了	七个	黄色的	桂圆
小蔡	扔了	八个	绿色的	柚子
小谢	选了	九个	橙色的	菠萝
小田	抱了	数个	白色的	龙眼
继续				
小虎	种了	六个	黄色的	香蕉

Figure 2.
Response screen for the CMS test.

Figure 3.

Mean performance (across 2 test runs and across subjects) for each category for sets of 10 randomly generated sentences. The lines show sigmoid fits to the data; the slope and the SNR for 50% correct are shown in the legend. The error bars show the standard error.

Figure 4.

Boxplots of SRTs for whole sentence recognition in noise for Run 1 (grey boxes) and Run 2 (white boxes) for sets of randomly generated sentences with and without the category level adjustment. In each box, the solid line shows the median, the dashed line shows the mean, and the error bars show the 10th and 90th percentiles.

Table 1

Auxiliary and target words selected for each category for the CMS. The parentheses show the English translation of the Chinese word; the Pinyin translation is shown immediately below.

姓名 (Name)	动词 (Verb)	数字 (Number)	颜色 (Color)	水果 (Fruit)
小王 (Wang) xiao3 wang2	捡了 (picks up) jian3 le0	一个 (one) yi1 ge0	褐色的 (brown) he3 se4 de0	金桔 (kumquat) jin1 ju2
小赵 (Zhao) xiao3 zhao4	卖了 (sells) mai4 le0	两个 (two) liang3 ge0	青色的 (blue) qing1 se4 de0	木瓜 (papaya) mu4 gua1
小虎 (Hu) xiao3 hu3	切了 (cuts) qie1 le0	三个 (three) san1 ge0	粉色的 (pink) fen3 se4 de0	枇杷 (loquat) pi2 pa2
小孔 (Kong) xiao3 kong3	种了 (plants) zhong4 le0	四个 (four) si4 ge0	银色的 (silver) yin2 se4 de0	香蕉 (banana) xiang1 jiao1
小周 (Zhou) xiao3 zhou1	拿了 (takes) na2 le0	五个 (five) wu3 ge0	灰色的 (grey) hui1 se4 de0	樱桃 (cherry) ying1 tao2
小邓 (Deng) xiao3 deng4	装了 (gets) zhuang1 le0	六个 (six) liu4 ge0	杏色的 (beige) xing4 se4 de0	草莓 (strawberry) cao3 mei2
小杜 (Du) xiao3 du4	看了 (looks at) kan4 le0	七个 (seven) qi1 ge0	黄色的 (yellow) huang2 se4 de0	桂圆 (longan) gui4 yuan2
小蔡 (Cai) xiao3 cai4	扔了 (throws) reng1 le0	八个 (eight) ba1 ge0	绿色的 (green) lv4 se4 de0	柚子 (grapefruit) you4 zi0
小谢 (Xie) xiao3 xie4	选了 (selects) xuan3 le0	九个 (nine) jiu3 ge0	橙色的 (orange) cheng2 se4 de0	菠萝 (pineapple) bo1 luo2
小田 (Tian) xiao3 tian2	抱了 (holds) bao4 le0	数个 (several) shu4 ge0	白色的 (white) bai2 se4 de0	龙眼 (longyan) long2 yan3

Definitions of auxiliary words added to the Name, Verb, Number, and Color categories: 小 = “little”; 了, 的 = Nonsense syllable typically added in conversational Mandarin; 个 = unit associated with number; 色 = color

Table 2
Distribution (in percent) of consonants, vowels and lexical tones from Tang (1995) and the CMS.

Consonant	Tang		CMS		Vowel	Tang		CMS		Tone	Tang		CMS	
None	12.4	13.3			u	8.4	8.3			T4	33.6	30.0		
j	7.9	8.3			i	8.2	10.0			T2	24.5	23.3		
l	7.0	8.3			an	5.5	5.0			T1	23.6	25.0		
x	6.0	5.0			ian	5.1	3.3			T3	17.3	20.0		
zh	5.9	5.0			ao	4.7	6.7			T0	1.0	1.7		
h	5.4	6.7			ang	4.0	1.7							
d	5.1	3.3			a	3.9	5.0							
q	5.1	5.0			ing	3.9	3.3							
b	5.1	5.0			iao	3.7	1.7							
g	4.3	3.3			eng	3.5	5.0							
t	4.2	3.3			ui	3.3	3.3							
sh	4.2	1.7			e	3.3	1.7							
ch	4.0	5.0			ai	3.2	3.3							
m	4.0	5.0			ong	3.1	5.0							
p	3.6	3.3			uo	3.1	1.7							
f	3.1	1.7			v	3.0	3.3							
k	2.5	3.3			ie	2.7	3.3							
s	2.4	3.3			ou	2.7	5.0							
z	2.4	3.3			en	2.3	1.7							
n	2.2	1.7			in	2.3	3.3							
c	1.9	3.3			uan	2.1	0.0							
r	1.3	1.7			iang	2.0	3.3							
					iu	2.0	3.3							
					ei	1.8	1.7							
					un	1.7	0.0							
					o	1.6	1.7							
					ia	1.5	0.0							

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Consonant	Tang	CMS	Vowel	Tang	CMS	Tone	Tang	CMS
	CMS			CMS			CMS	
			uang	1.4	1.7			
			van	1.3	3.3			
			vn	1.1	0.0			
			ve	1.0	0.0			
			ua	0.9	1.7			
			uai	0.6	1.7			
			iong	0.6	0.0			
			ueng	0.1	0.0			