Measuring open-set, word recognition in school-aged children: Corpus of monosyllabic target words and speech maskers

Angela Yarnell Bonino and Ashley R. Malley
Department of Speech, Language, and Hearing Sciences, University of Colorado Boulder, Boulder, Colorado 80309, USA
angela.bonino@colorado.edu, ashley.malley@colorado.edu

Abstract: A corpus of stimuli has been collected to support the use of common materials across research laboratories to examine school-aged children’s word recognition in speech maskers. The corpus includes (1) 773 monosyllabic words that are known to be in the lexicon of 5- and 6-year-olds and (2) seven masker passages that are based on a first-grade child’s writing samples. Materials were recorded by a total of 13 talkers (8 women; 5 men). All talkers recorded two masker passages; 3 talkers (2 women; 1 man) also recorded the target words. The annotated corpus is freely available online for research purposes.

© 2019 Acoustical Society of America

Date Received: August 17, 2019 Date Accepted: September 30, 2019

1. Introduction

Children’s everyday listening environments are noisy, requiring them to hear speech embedded in competing background sounds. In order to capture school-aged children’s listening difficulty in these situations, clinicians and researchers frequently measure masked speech recognition. The commercially-available, pediatric clinical tests evaluate word or sentence recognition in either a speech-shaped noise or a multi-talker babble, a masker comprised of ≥ 4 talkers (BKB-SIN, 2005; Nilsson et al., 1994; Wilson et al., 2010; the exception is Jerger and Jerger, 1982). In contrast to this traditional clinical approach for measuring masked speech recognition, there is a growing research interest to measure speech recognition in a two-talker speech masker. To evaluate performance in a two-talker masker, researchers must develop their own in-house stimuli, as no standardized materials for school-aged children are currently available. This process can be time intensive and acoustical differences between stimuli may lead to different results across studies. The purpose of this paper is to provide a corpus of speech materials that can be used across laboratories to evaluate school-aged children’s open-set word recognition in speech maskers.

The motivation for measuring speech recognition in a two-talker masker is that, compared to speech-shaped noise, this condition results in a pronounced and prolonged course of auditory development (e.g., Bonino et al., 2013; Corbin et al., 2016; Hall et al., 2002). For example, Hall et al. (2002) measured speech recognition thresholds for a forced-choice spondee recognition task in one of two continuously played maskers: a two-talker or speech-shaped noise. Compared to the 3-dB child–adult difference reported for the speech-shaped noise masker, the 7-dB child–adult difference for the two-talker masker was greater. Furthermore, adult-like performance is achieved at a later age during childhood for a two-talker masker than for a speech-shaped noise masker (e.g., Bonino et al., 2013; Corbin et al., 2016; Hall et al., 2002). For example, adult-like open-set word recognition performance is achieved around 8 to 10 years for speech-shaped noise and 13 to 15 years for a two-talker masker (Corbin et al., 2016).

The observed differences between a two-talker masker and a speech-shaped noise masker is a result of the maskers generating different types of masking. Hearing in speech-shaped noise is primarily limited by energetic masking: the overlap of the signal and masker at the level of the peripheral auditory system (e.g., Fletcher, 1940). In contrast, a two-talker speech masker is expected to produce substantial informational masking, in addition to energetic masking. Informational masking is thought to be the result of limited and/or ineffective central auditory processes, including separating the signal from the masker and selectively attending to the signal while disregarding the masker.
(e.g., Durlach et al., 2003). The ability of a speech masker to produce informational masking is related to the number of talkers in the masker stream. Results from adult listeners suggest that informational masking is the greatest for speech maskers comprised of two talkers and decreases as more talkers are added, with maskers of ≥10 talkers producing little or no informational masking (Freyman et al., 2004).

This paper aims to provide standardized speech materials for measuring masked open-set word recognition in children as young as 5 years. We opted to focus on open-set word recognition because there is a critical need for the development of a children’s monosyllabic word corpus that includes a large number of words, detailed phonetic and lexicon information, and high-quality audio recordings. Building upon our previous laboratory work (e.g., Corbin et al., 2016), here we provide a new collection of (1) 773 monosyllabic target words recorded by 3 talkers and (2) single-talker masker passages recorded by 13 talkers. A description of the recording process, acoustical analysis of the talkers, analysis of phonotactic probability and lexical features of the target words, and access to the materials follows.

2. Corpus

2.1 Selection of target words

Drawing on the corpus used in previous experiments of school-aged children’s open-set word recognition (Bonino, 2012; Browning et al., 2019; Corbin et al., 2016), we selected 773 target words to be included in the corpus presented here. Three audiology graduate students reviewed the original list of 842 words and removed duplicate words (e.g., homophones and different tenses of the same base word). Based on their cultural background and experience, reviewers also removed words judged to be potentially inappropriate (e.g., “war” and “knife”) or unfamiliar to young children. In order to verify that words were within the lexicon of 5-year-old children, all words were required to be present in the on-line Child Calculator (Storkel and Hoover, 2010). The Child Calculator draws on child corpora of American English from oral language samples of kindergarteners (Kolson, 1960) and first graders (Moe et al., 1982). Selected monosyllabic target words were also required to be 2 to 5 phonemes in length. Of the 773 words selected, 12% were 2-phoneme words \((n = 91)\), 59% were 3-phoneme words \((n = 454)\), 26% were 4-phoneme words \((n = 203)\), and 3% were 5-phoneme words \((n = 25)\).

2.2 Creation of masker passages

Masker passages were created by compiling writing samples from a first-grade child who was a native speaker of American English and typically-developing based on parental report. Masker passages were created from narrative and expository writing samples, as well as brief daily journal entries. Writing samples were edited to correct spelling and punctuation errors, but word selection and syntax were only edited if readability was judged to be poor for the sentence. Writing samples were then assigned to one of seven masker passages. On average, masker passages were 532 words in length [standard deviation (SD) = 65.8; min = 481; max = 672 words]. Examination of Lexile text scores (MetaMetrics, 2019) confirmed that text complexity was similar across all masker passages (500−600 L). Transcripts of all masker passages are provided in the online resource. Individual talker masker recordings were not verified against the original text to identify possible alterations by the talker.

2.3 Talker population

A total of 8 women (24 to 29 years) and 5 men (22 to 28 years) served as talkers for this corpus. All individuals were monolingual and native speakers of American English. All individuals signed a waiver and release of their audio recordings.

2.4 Recording procedure

Individual talkers were recorded in a double-walled, acoustically-isolated booth (Industrial Acoustics Company, North Aurora, IL). A condenser microphone (Shure KSM42/SG, Shure Microphone, Niles, IL) was placed approximately 6 in. from the talker’s mouth on a microphone stand with a pop filter. The recordings were amplified (M-Audio M-Track 2 × 2, M-Audio, Cumberland, RI) and digitized at a resolution of 32 bits with a sampling rate of 44.1 kHz which were captured in Audacity (version 2.2.6). All talkers were instructed to speak naturally. An experimenter monitored the recording session and made changes to the recording hardware and provided feedback to the talker. For target word recordings, each target talker recorded all 773 words in isolation without the use of a carrier phrase. Target words were randomly assigned into blocks of 20 words; recording order of the blocks was then randomized across
talkers. A minimum of two recordings were made for each target word. For masker recordings, talkers read two passages, with a minimum of two recordings per passage.

In addition to recording the materials for the corpus, all talkers provided audio recordings to allow for acoustical analysis of their voice. Talkers were asked to sustain the vowels /a/ and /i/ for at least 5 s. Talkers also recorded the following sentence, “I am subject number [] and I am a student at the University of Colorado Boulder.”

2.5 Post-processing of stimuli

Target word recordings were spliced by hand in Audacity (version 2.2.0). Three experimenters (including A.R.M.) independently evaluated each spliced token. Tokens identified with speech disfluencies or amplitude inconsistencies (e.g., peak clipping) were rerecorded. Individual .wav files for each target word were created and separated by talker. For each talker, .wav files were resampled at a rate of 24.414 kHz and scaled to equivalent root-mean-square (RMS) level using MATLAB. The targeted RMS level was set to maximize the level for each individual target talker, resulting in different RMS levels across the three target talkers. Our online resource provides the “raw” and “resampled” audio files, the MATLAB script used to scale tokens, and metadata on the duration and level for individual tokens.

The masker samples were manually edited to eliminate silent pauses > 300 ms in Audacity. Resulting masker files were 106 to 206 s in duration (see Table 1). Using Praat (version 6.0.52; Boersma and Weenink, 2019), recordings were resampled at 24.414 kHz, RMS scaled to an average intensity level of 70 dB sound pressure level, and exported as mono .wav files. Two individual masker recordings—in both the raw and resampled formats—are provided for each talker in the corpus.

3. Acoustical analysis of target and masker talkers

Acoustical analyses were performed using the RMS scaled masker files from each talker in Praat (version 6.0.52; Boersma and Weenink, 2019). This information is provided to help guide the selection of target and masker talkers because the mix of their vocal characteristics appears to be important in determining the effectiveness of a two-talker masker (e.g., Brungart et al., 2001; Calandruccio et al., 2019). Table 1 reports the mean and SD for fundamental frequency ($F_0$). Mean $F_0$ ranged from 186.8 to 230.5 Hz and 95.8 to 153.7 Hz for female and male talkers, respectively. The long-term-average-speech spectra (LTASS) for individual talkers are provided in Fig. 1. Each curve represents an individual talker, with the solid lines indicating the target talkers. The LTASS were only computed for the “story 1” audio files. To examine another potential difference across the talkers, speech rate was calculated using an automated procedure in Praat that detects syllable nuclei (version 2; De Jong and

| Voice Characteristics | Masker Recordings |
|-----------------------|-------------------|
| **Fundamental Frequency (Hz)** | **Reading Passage Number** | **Total Duration (s)** | **Speech Rate (syllables/s)** |
| **Story 1** | **Story 2** | **Story 1** | **Story 2** | **Story 1** | **Story 2** |
| **Talker ID** | | | | |
| Female_1 | 205.1 (27.8) | 205.6 (27.6) | 4 | 1 | 134.4 | 141.1 | 4.0 | 4.0 |
| Female_2* | 192.5 (23.5) | 193.9 (22.7) | 3 | 2 | 146.1 | 123.5 | 4.1 | 4.0 |
| Female_3* | 211.1 (19.6) | 223.6 (22.6) | 5 | 5 | 206.9 | 145.7 | 3.8 | 3.9 |
| Female_4 | 226.8 (34.2) | 217.8 (25.5) | 1 | 6 | 126.7 | 128.4 | 4.2 | 4.1 |
| Female_5 | 186.8 (31.0) | 198.7 (39.4) | 2 | 7 | 129.5 | 106.3 | 4.3 | 5.1 |
| Female_6 | 213.4 (18.4) | 203.9 (18.7) | 6 | 3 | 117.6 | 135.0 | 4.4 | 4.1 |
| Female_7 | 225.4 (38.1) | 225.1 (42.5) | 7 | 4 | 129.8 | 139.6 | 4.4 | 4.2 |
| Female_8 | 200.9 (17.5) | 230.5 (42.7) | 5 | 4 | 170.8 | 134.5 | 4.2 | 4.1 |
| Male_1* | 104.2 (18.3) | 108.3 (20.8) | 5 | 6 | 181.0 | 150.8 | 4.1 | 3.8 |
| Male_2 | 96.4 (12.1) | 106.7 (18.5) | 7 | 1 | 124.7 | 117.1 | 4.3 | 4.3 |
| Male_3 | 153.7 (21.7) | 153.3 (23.6) | 1 | 5 | 116.7 | 144.4 | 4.6 | 4.8 |
| Male_4 | 97.5 (10.8) | 95.8 (15.1) | 1 | 2 | 159.8 | 143.3 | 3.9 | 3.8 |
| Male_5 | 114.8 (21.6) | 115.2 (21.2) | 7 | 3 | 109.7 | 117.8 | 4.5 | 4.4 |

Table 1. Fundamental frequency ($F_0$), duration, and speech rate are reported for each masker recording by talker ID number. SD values are reported in parentheses. The three talkers that recorded target words are noted with an asterisk.
As shown in Table 1, speech rate (number of syllables/total duration) ranged from 3.8 to 5.1 syllables/s across the masker files. Our online resource provides detailed documentation of the above analyses as well as supplemental analyses, including estimates of the variation in frequency (jitter) and amplitude (shimmer) for each talker based on their sustained vowel recordings.

4. Analysis of phonotactic probability and lexical features of target words

Using the Child Calculator (Storkel and Hoover, 2010), estimates of phonotactic probability (positional segment average and biphone average), neighborhood density, and word frequency were calculated for each target word because these parameters can affect listener performance (e.g., Jesse and Helfer, 2019; Ren et al., 2015). As described in more detail by Storkel and Hoover (2010), the positional segment average is calculated by adding the positional segment frequency for each sound in the target word and then dividing it by the number of sounds in the target word. This value represents the likelihood of the target word having its particular order of phonemes based on the child corpora (Kolson, 1960; Moe et al., 1982) used by the Child Calculator. The biphone average is computed in a similar manner, but represents the likelihood of pairs of adjacent sounds occurring, rather than individual phonemes. The biphone average is computed by dividing the sum of the biphone frequencies of each sound pair in the target word by the number of sound pairs in the word. Across all target words in our corpus, the mean positional segment average value was 0.053 (SD = 0.017; max = 0.099; min = 0.0009) and the mean biphone average value was 0.005 (SD = 0.003; max = 0.015; min = 0.0002). Neighborhood density is a count of the number of words in the child corpora used by the Child Calculator that differ from the target word by one sound via substitution, addition, or deletion. On average, target words in our corpus have 11.608 neighbors (SD = 6.928; max = 34; min = 0). Word frequency (in log10) was calculated based on the number of occurrences that the target word, based on its phonetic transcription, occurred in the child corpora used by the Child Calculator. Mean word frequency was 3.064 (SD = 0.901; max = 5.670; min = 1.0) for our corpus.

As is well established in the literature (e.g., Storkel, 2004), the four parameters described above are influenced by word length. In order to better visualize the effect of word length, Fig. 2 provides individual and average data for each parameter based on word length. A single parameter is shown per panel; left to right is: positional segment average, biphone average, number of neighbors, and word frequency (in log10). Boxes represent the 25th to 75th percentile and whiskers represent the 10th to 90th percentile. Open squares and vertical lines indicate the mean and median, respectively. Filled diamonds depict the values for individual target words. Using JMP Pro (version 13.1.0), separate One-Way Analysis of Variance tests were conducted for each individual parameter to determine the effect of word length. For our corpus, a significant effect of word length was confirmed for each parameter: positional segment [F(3, 769) = 58.46, p < 0.001]; biphone average [F(3, 769) = 96.456, p < 0.001]; number of neighbors [F(3, 769) = 165.529, p < 0.001]; and word frequency [F(3, 769) = 13.851, p < 0.001]. For each parameter, post hoc analyses were conducted with Tukey-Kramer HSD to determine...
differences between word length ($\alpha = 0.05$). Significant post hoc differences are indicated in Fig. 2 by the letter to the right of each upper whisker. Word length groups not connected by the same letter are significantly different than one another.

Results from the above analysis highlight that care should be used during target word selection if the goal is to create balanced word lists from the larger corpus. For specific target words in the corpus and their associated values, we refer the reader to a detailed .xlsx file provided in our online resource. The phonetic transcription for each target word is provided in Klattese, a computer-readable version of the International Phonetic Alphabet (Vitevitch and Luce, 2004). For each target word, segment and mean values can be found for phonetic probability, as well as estimates of word frequency and neighborhood density based on the child corpora used by Storkel and Hoover (2010). $Z$-scores—referenced to values from the child corpora (reported in Appendix A in Storkel and Hoover, 2010)—were also calculated for positional segment, biphone average, and number of neighbors. $Z$-scores are provided because they are preferred to raw scores for some statistical analyses when multiple word lengths are used (Storkel, 2004).

5. Corpus description

The corpus described here is being made freely available for download under a Creative Commons CC-BY-NC 4.0 International License (https://osf.io/43xfh/). The downloads consist of individual audio files of the 773 monosyllabic target words that were recorded by 3 talkers (2 women; 1 man). One-talker masker recordings are also available from 13 talkers (8 women; 5 men). All audio files are available in a .wav format in either their raw format (not RMS equated; 44.1 kHz sampling rate; 32 bits) or in their resampled format (RMS scaled; 24.414 kHz; 32 bits), as described above. Acoustical analyses and the associated audio files are also provided for each talker. The downloads are accompanied by documents describing associated metadata, including masker transcriptions and a spreadsheet with individual target words and their associated data from the Child Calculator (Storkel and Hoover, 2010). The corpus is available at https://osf.io/43xfh/ (Bonino and Malley, 2019).

6. Summary

A large corpus of monosyllabic target words and single-talker masker streams is available for research use (Bonino and Malley, 2019). Based on the lexical features of this corpus, the target words are expected to be appropriate for children as young as 5 years who have typical speech and language development. By providing values for phonotactic probability and lexical features for each target word, researchers will be able to control for these parameters when selecting target words. Phonetic transcripts for each target word are provided which will allow researchers to implement a phoneme scoring algorithm or to select target words that contain particular sounds. The provision of detailed acoustical analyses for talkers enables researchers to achieve the desired vocal characteristics of target and masker talkers. The benefit of the corpus containing single-talker masker streams from 13 individual talkers is that it gives researchers the ability to create a variety of masker conditions.

Acknowledgments

This work was supported by a Graduate Fellowship (A.R.M.) and an Undergraduate Research Opportunity Program Award at the University of Colorado Boulder. We appreciate help with recording and editing tokens from Christopher Hanson, Madison...
References and links

For each parameter, post hoc testing was conducted with Tukey-Kramer HSD to determine differences between word length. For the mean positional segment scores, post hoc testing revealed that 2-phoneme words had a lower positional segment score than all other word lengths ($p < 0.001$). Positional segment scores for 4-phoneme words were also significantly higher than 3-phoneme words ($p = 0.038$). Five-phoneme words were not significantly different from 3-phoneme words ($p = 0.972$) nor 4-phoneme words ($p = 0.917$). For biphone average scores, all word lengths were significantly different from one another ($p < 0.0001$) except for 4- and 5-phoneme words ($p = 0.996$). For neighborhood density, all word lengths were significantly different from one another ($p = 0.01$ for 4- vs 5-phonemes; $p < 0.0001$ for all other comparisons). For word frequency values, 2-phoneme words occurred significantly more frequently than all other word lengths ($p < 0.0001$). No other significant differences in word frequency were seen for the other phoneme comparisons (all $p$-values $> 0.3$).

BKB-SIN: Bamford-Kowal-Bench Speech in Noise Test. (2005). Etymotic Research, Elk Grove, IL.

Boersma, P., and Weenink, D. (2019). “Praat: Doing phonetics by computer (version 6.0.52) [computer program].” http://www.praat.org/ (Last viewed July 9, 2019).

Bonino, A. Y. (2012). “The effect of signal-temporal uncertainty during childhood,” The University of North Carolina at Chapel Hill, ProQuest Ann Arbor, MI, Dissertations Publishing.

Bonino, A. Y., Leibold, L. J., and Buss, E. (2013). “Release from perceptual masking for children and adults: Benefit of a carrier phrase,” Ear Hear. 34, 3–14.

Bonino, A. Y., and Malley, A. (2019). “Corpus for measuring masked open-set word recognition in school-aged children,” retrieved from osf.io/43xfh. DOI: 10.17605/OSF.IO/43XFH (Last viewed September 30, 2019).

Brown, J. M., Buss, E., Flaherty, M., Vallier, T., and Leibold, L. J. (2019). “Effects of adaptive hearing aid directionality and noise reduction on masked speech recognition for children who are hard of hearing,” Am. J. Audiol. 28, 101–113.

Brungart, D. S., Simpson, B. D., Ericson, M. A., and Scott, K. R. (2001). “Informational and energetic masking effects in the perception of multiple simultaneous talkers,” J. Acoust. Soc. Am. 110, 2527–2538.

Calandruccio, L., Wasuk, P. A., Buss, E., Leibold, L. J., Kong, J., Holmes, A., and Oleson, J. (2019). “The effect of target/masker fundamental frequency contour similarity on masked-speech recognition,” J. Acoust. Soc. Am. 146, 1065–1076.

Corbin, N. E., Bonino, A. Y., Buss, E., and Leibold, L. J. (2016). “Development of open-set word recognition in children: Speech-shaped noise and two-talker speech maskers,” Ear Hear. 37, 55–63.

De Jong, N. H., and Wempo, T. (2009). “Praat script to detect syllable nuclei and measure speech rate automatically,” Behav. Res. Methods 41, 385–390.

Durlach, N. I., Mason, C. R., Kidd, G., Arbogast, T. L., Colburn, H. S., and Shinn-Cunningham, B. G. (2003). “Note on informational masking,” J. Acoust. Soc. Am. 113, 2984–2987.

Fletcher, H. (1940). “Auditory patterns,” Rev. Modern Phys. 12, 47–65.

Freyman, R. L., Balakrishnan, U., and Helfer, K. S. (2004). “Effect of number of masking talkers and auditory priming on informational masking in speech recognition,” J. Acoust. Soc. Am. 115, 2246–2256.

Hall, J. W., Grose, J. H., Buss, E., and Dev, M. B. (2002). “Spondee recognition in a two-talker masker and a speech-shaped noise masker in adults and children,” Ear Hear. 23, 159–165.

Jerger, S., and Jerger, J. (1982). “Pediatric Speech Intelligibility Test: Performance-intensity characteristics,” Ear Hear. 3, 325–334.

Jesse, A., and Helfer, K. S. (2019). “Lexical influences on errors in masked speech perception in young, middle-aged, and older adults,” J. Speech, Lang., Hear. Res. 62, 1152–1166.

Kolson, C. J. (1960). “The vocabulary of kindergarten children,” unpublished doctoral dissertation, University of Pittsburgh.

MetaMetrics (2019). “Free Lexile Analyzer” [computer software]. Retrieved from https://la-tools.lexile.com/free-analyze/ (Last viewed July 31, 2019).

Moe, A. J., Hopkins, K. J., and Rush, R. T. (1982). The Vocabulary of First Grade Children (Thomas, Springfield, IL).

Nilsson, M., Soli, S. D., and Sullivan, J. A. (1994). “Development of the Hearing in Noise Test for the measurement of speech reception thresholds in quiet and noise,” J. Acoust. Soc. Am. 95, 1085–1099.

Ren, C., Liu, S., Liu, H., Kong, Y., Liu, X., and Li, S. (2015). “Lexical and age effects on word recognition in noise in normal-hearing children,” Int. J. Pediatric Otorhinolaryngol. 79, 2023–2027.

Storkel, H. L. (2004). “Methods for minimizing the confounding effects of word length in the analysis of phonotactic probability and neighborhood density,” J. Speech, Lang., Hear. Res. 47, 1454–1468.

Storkel, H. L., and Hoover, J. R. (2010). “An online calculator to compute phonotactic probability and neighborhood density on the basis of child corpora of spoken American English,” Behav. Res. Methods 42, 497–506.

Vitevitch, M. S., and Luce, P. A. (2004). “A Web-based interface to calculate phonotactic probability for words and nonwords in English,” Behav. Res. Methods 36, 481–487.

Wilson, R. H., Farmer, N. M., Gandhi, A., Shelburne, E., and Weaver, J. (2010). “Normative data for the Words-in-Noise Test for 6-to 12-year-old children,” J. Speech, Lang., Hear. Res. 53, 1111–1121.