DOI: 10.61782/fa.2023.0706

forum acusticum 2023

THE DANISH SENTENCE TEST (DAST) CORPUS OF AUDIO AND AUDIO-VISUAL RECORDINGS OF SENTENCES AND MONOLOGUES

Abigail Anne Kressner[1,2*] **Kirsten Maria Jensen Rico**[1,2] **Johannes Kizach**[1] **Brian Kai Loong Man**[1,3] **Anja Kofoed Pedersen**[4] **Lars Bramsløw**[3] **Brent Kirkwood**[5]

[1] Health Technology, Technical University of Denmark, Denmark

[2] Copenhagen Hearing and Balance Centre, Rigshospitalet, Denmark

[3] Demant, Smørum, Denmark

[4] WS Audiology, Lynge, Denmark

[5] GN Hearing, Ballerup, Denmark

ABSTRACT

A new, larger corpus of Danish sentences has been developed to create a new Danish Sentence Test (DAST). The corpus is made up of audio and audio-visual recordings of 1200 linguistically balanced sentences, all of which are spoken by both two male and two female professional talkers. The sentences were constructed using a template-based method that facilitated control over both word frequency and sentence structure. The resulting written sentences were evaluated linguistically in terms of phonetic distributions and naturalness. Due to a relatively low number of sentences being rated as problematic, all 1200 sentences were included in the recording stage. Besides the sentences for each talker, 30+ minutes of monologues were also recorded from each of the same four talkers. For one selected talker, the sentences were presented to 60 normal-hearing listeners at a fixed set of signal-noise ratios to estimate a psychometric function for each sentence. These psychometric functions, together with the naturalness results, phonetic distributions, and quality assessments, were employed to build equivalent lists of 20 sentences for speech-in-noise testing. Development of an adaptive version of the test to measure speech reception thresholds is under way, as is the validation of the equivalency of the resulting lists.

[*] ***Corresponding author***: *aakress@dtu.dk*

Keywords: *speech intelligibility, corpus, audio, audio-visual*

1. INTRODUCTION

Many research and development projects in the fields of speech and hearing sciences rely heavily on speech corpora. For example, a signal processing engineer working on the development of a new hearing aid algorithm will often use sentence testing with a group of listeners to optimize the parameters of their algorithm, and a machine learning engineer working on the development of a new noise reduction algorithm often uses sentences from a speech corpus to train their classifiers. In such use cases, it is valuable to have access to a large number of sentences in order to test different algorithm parameters, to test with or train on different noise types or amounts of reverberation, and to investigate test-retest reliability or generalizability. Although speech corpora large enough for such sentence testing exist in other languages, a large enough corpus in Danish does not.

The goal of this study was to develop a new, larger corpus of Danish sentences for the development of a sentence test. The corpus is made up of audio and audio-visual recordings of 1200 linguistically balanced sentences, contains recordings of the sentences from two male and two female talkers, and includes continuous monologue recordings by the same talkers of both spontaneous speech and an audiobook.

2. METHODS

The sentences were constructed using a template-based method that facilitated control over both word frequency and sentence structure. Specifically, each "template" defined a specific syntax that can be found naturally in Danish such that sentences built from the template would sound like naturally spoken Danish sentences. Furthermore, each template defined the distribution of a set of three lexical keywords within the sentence so that the keywords would be well-distributed, and to the extent that was possible, so that each sentence would end with a keyword (i.e., for facilitation of a sentence-final word identification and recall, SWIR, adaptation of the test later [1]). Ten of these templates were constructed so that a list of 20 sentences would contain two of each type. The keywords were taken from a list of the most frequently written words in Danish [2] (i.e., a list of the most frequency spoken words in Danish does not exist), and each keyword was repeated maximally three times within the corpus. Function words and adverbs were employed as needed to construct sentences that would be as natural as possible. The naturalness of each sentence was then rigorously evaluated using an online survey where participants rated the naturalness of the sentence in written form on a 7-point Likert scale, the results of which were presented in [3].

Two female and two male talkers were recorded speaking each of the sentences aloud in an acoustically treated, green screen studio. Each talker also recorded 30+ minutes of monologues, including 15-20 minutes of spontaneous speech and 15-20 minutes of reading aloud from a book, and for a subset of the talkers, approximately 120 minutes of dialogues with two of the other talkers. The audio-visual material was imported into Adobe Premier Pro, segmented into individual sentences, rendered to contain a medium-tone gray background, and then exported into separate audio and visual files (i.e., using the WAV and MP4 file formats, respectively). The audio clips were then post-processed in Matlab with a high-pass filter (i.e., zero-phase filtering with a 70 Hz cutoff), and a half-sided Hanning window with a duration of 500 ms was applied before the onset of and after the offset of each sentence. Finally, each of the audio and visual recordings of the sentences from all four of the talkers were assessed qualitatively in terms of phonetics (e.g., pronunciation and voice quality), the sound quality, and the visual component (e.g., facial expression and movement).

For one selected talker (i.e., the talker with the largest number of sentences approved for use in a speech-in-noise test), the sentences were presented to 60 normal-hearing listeners (aged between 18 and 75 yrs) at a fixed set of signal-noise ratios (SNRs) to estimate psychometric functions for each individual sentence. Each participant listened to 21 blocks of 20 sentences each, where the first block was for familiarization to the task and the latter 20 blocks were scored. The sentences were pseudo-randomized across participants so that each participant heard 420 unique sentences; each participant heard an even number of each of the SNRs (-10, -7.5, -5, -2.5, and 0 dB); and each combination of sentence and SNR was heard by exactly four participants. The sentences were presented diotically over headphones (Sennheiser HD650) together with stationary speech-shaped noise (SSN) shaped to the long-term average spectrum of the selected talker. The SSN was presented at 70 dB SPL (Z) in each ear, and the target level was adjusted depending on the SNR. Equalization filters were applied to ensure a flat frequency response. All words were scored, but only keywords were included in the analysis.

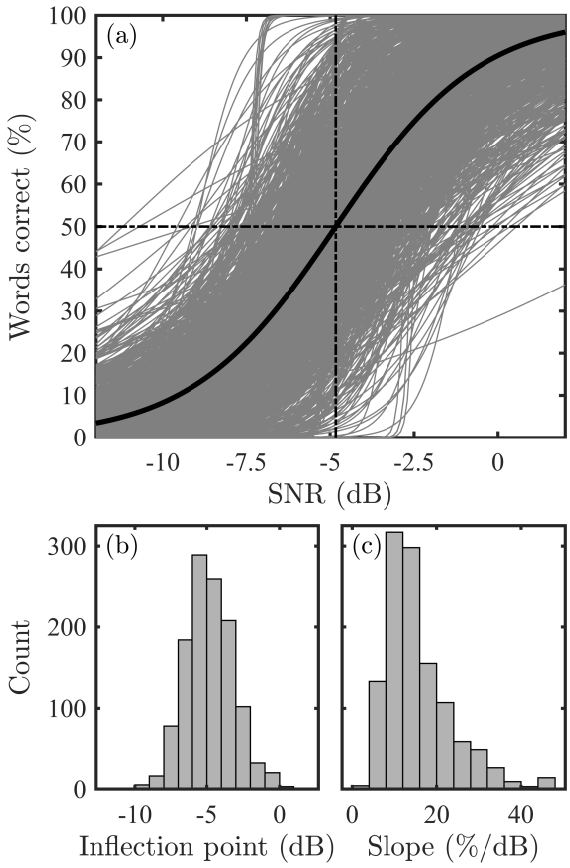

Figure 1. Estimated psychometric functions (a) for each individual sentence (gray) and for the overall mean (black). Histograms of the estimated (b) inflection points and (c) slopes.

3. RESULTS

For each sentence, psychometric functions were fitted with scores based on the number of keywords correctly identified from the 20 participants who listened to the sentence (i.e., four unique scores at each SNR). The fits were based on nonlinear regression with the following logistic function,

$$\Psi(\mathrm{SNR}) = \frac{100}{1 + \exp\left(-4\frac{k}{100}(\mathrm{SNR} - \mathrm{SRT}_{50})\right)} \quad (1)$$

where SRT_{50} is the estimated inflection point at the 50%-point of the curve (i.e., the speech reception threshold, SRT), and k is the estimated slope of the psychometric function at the inflection point.

The resulting psychometric function estimates for each sentence are plotted in Fig. 1(a), and histograms of the estimated inflection points and slopes are shown in Figs. 1(b) and 1(c), respectively. The mean inflection point across all sentences was -4.8 dB (standard deviation 1.7 dB), and the median slope was 13.8 %/dB. The estimated inflection points were roughly normally distributed (Q_1: -5.9; Q_2: -4.9; Q_3: -3.7), whereas the distribution of the slopes was skewed (Q_1: 10.3; Q_2: 13.8; Q_3: 19.6). These distributions align with existing literature (e.g., [4]).

4. CONCLUSION

A new, larger corpus of Danish sentences has been developed to facilitate speech-in-noise sentence testing. The corpus is made up of audio and audio-visual recordings of 1200 linguistically balanced sentences. In this study, a psychometric function has been estimated for all acoustically presented sentences from one of the talkers in the corpus. These psychometric functions will be employed to create lists of sentences that achieve equivalent SRT estimates through an adaptive speech-in-noise test procedure.

5. ACKNOWLEDGMENTS

Special thanks to Lise Bruun Hansen, Jens Bo Nielsen, Sofie Bundgaard, Amal Abdulqadir Ali, Pernille Holtegaard, Michael Nielsen, Laura Balling, Tobias Andersen, Tobias May, Filip Rønne, David Harbo Jordell, Jens Hjortkjær, and Torsten Dau for their valuable contributions to this project. The DAST Project has been supported by Demant, GN Hearing, and WS Audiology.

6. REFERENCES

[1] E. Ng, M. Rudner, T. Lunner, M. Pedersen, and J. Rönnberg: "Effects of noise and working memory capacity on memory processing of speech for hearing-aid users," *International Journal of Audiology*, 52, pp. 433-441, 2013.

[2] J. Asmussen: *De hyppigste ord i Dansk*. Det Danske Sprog- og Litteraturselskab. Retrieved December 2021 from https://korpus.dsl.dk/resources/details/freq-lemmas.html, 2017.

[3] A. Kressner, K. Rico, J. Kizach, B. Man, A. Pedersen, L. Bramsløw, and B. Kirkwood, "The Danish Sentence Test (DAST) corpus of audio and audio-visual recordings of sentences and monologues," in *Proc. Of the Speech in Noise Workshop*, (Split, Croatia), 2023.

[4] K. Miles, G. Keidser, K. Freeston, T. Beechey, V. Best, and J. Buchholz: "Development of the Everyday Conversational Sentences in Noise test," *Journal of the Acoustical Society of America*, 147 (3), pp. 1562-1576, 2020.