Towards a Danish speech material for speech-on-speech masking investigations

JENS BO NIELSEN[1], TOBIAS NEHER[2], AND TORSTEN DAU[1],

[1] *Centre for Applied Hearing Research, Technical University of Denmark, DK-2800 Lyngby, Denmark*

[2] *Eriksholm Research Centre, Oticon A/S, DK-3070 Snekkersten, Denmark*

The objective of the present project is to create a Danish speech material designed specifically for investigations of speech-on-speech masking. Speech-on-speech masking is characterized by "informational" masking effects, often defined as the masking that speech creates in addition to the energetic masking by pure noise. In the present project, 804 unique sentences have been created, divided into three subsets, and uttered by three professional female talkers. Listening tests are in progress in order to compile test lists, each consisting of 20 sentences of approximately equal intelligibility. The final corpus is expected to consist of three sets of 10 test lists that are equally usable as target or masker signals. A pilot validation of the corpus has shown a high average speech recognition threshold (SRT) and a relatively small test-retest standard deviation compared to other materials.

INTRODUCTION

Speech-on-speech masking is often divided into an "energetic" and an "informational" effect. The energetic effect corresponds to that of a noise signal with the same spectro-temporal characteristics as the masking speech, while informational masking is the additional effect caused by the confusion and the uncertainty that arises when the masker is speech (Helfer and Freyman, 2009). Investigations of speech-on-speech masking are expected to be useful for developing new strategies for improving speech intelligibility in environments with competing speech.

The new Danish speech-on-speech corpus is designed to fulfil four essential requirements that will facilitate effective speech-on-speech masking investigations:

i. The measured speech recognition thresholds (SRTs) should be representative of the signal-to-noise ratios (SNRs) experienced in everyday speech communication. A value of approximately 0 dB is desirable. This is considerably higher than that achievable with existing speech materials.
ii. Each sentence begins with a call sign that allows cuing of the target sentence. When the target and masker sentences are similar, the call sign is essential for separating them.

Proceedings of ISAAR 2011: Speech perception and auditory disorders. 3rd International Symposium on Auditory and Audiological Research. August 2011, Nyborg, Denmark. Edited by T. Dau, M. L. Jepsen, J. Cristensen-Dalsgaard, and T. Poulsen. ISBN 87-990013-3-0. EAN 9788799001330. The Danavox Jubilee Foundation, 2012.

iii. The material is open-set, i.e. the response options are effectively unlimited. Everyday speech recognition involves processes – e.g. retrieving the target word from lexical memory – that are better simulated by open-set than closed-set tests (Clopper *et al*., 2006). An open-set test also leads to a higher SRT than closed-set tests.
iv. The speech rate is similar to that of natural speech to ensure ecological validity in this respect.

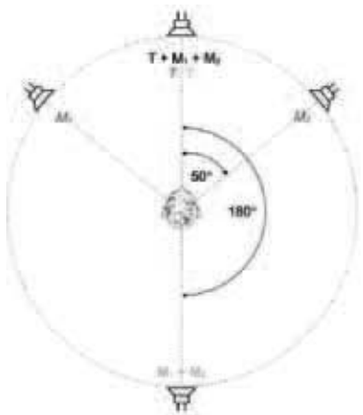

Fig. 1: Examples of different speech-on-speech masking set-ups with one target and two simultaneous maskers (T = target, M_1 = masker 1, M_2 = masker 2). The target is presented from the front while the maskers are presented from the same or different directions. The difficulty of the listening task depends strongly on the location of the maskers. From Neher *et al*. (2009).

SENTENCE FORMAT

The format of the sentences in the corpus is similar to that of the TVM-corpus developed by Helfer and Freyman (2009). The sentences are based on the (Danish) carrier:

***Navn* tænkte på ..... og ..... i går,**

where *Navn* is a name, and each blank represents a unique noun. This format has several advantages compared to natural sentences:

i. In listening tests, the scoring of the sentences is less prone to uncertainties on behalf of the test leader.
ii. Context effects are low and relatively consistent as long as frequent word pairs (such as knife and fork) are avoided. Removing context effects is an effective method for raising the SRT.

iii. It is relatively easy to produce a large number of sentences.
iv. The (usually cumbersome) process of assessing the naturalness of the sentences is simplified. When selecting the target words, it is deemed sufficient to avoid conflicts with the carrier sentence.
v. The two target words can be selected to produce sentences of similar duration. This ensures that masker and target sentences overlap in time.

CHOOSING THE CALL SIGNS

In order to reduce the risk that one of the call signs corresponds to the listener's name in an actual test, three Danish female names having a low occurrence in the Danish population, while at the same time not being considered strange, were selected:

i. Dagmar (1429 occurrences, Jan. 2011)
ii. Asta (4352)
iii. Tine (7852)

The durations of the spoken names are similar (300-350 ms). The names were judged to be equally identifiable when presented simultaneously.

CREATING THE SENTENCES

The target words were extracted from databases of Danish nouns according to the following requirements:

i. They contain one or two syllables. Three syllables were only allowed if the word was shorter than eight letters.
ii. They contain at least one 'strong' consonant, e.g. /k/, /t/, /s/, /f/, /p/.
iii. They are not negative, emotional, technical, abstract or slang-like.
iv. They are neutral and concrete.
v. They do not conflict with the carrier sentence.

The target words were combined in pairs by taking their overall duration and their number of lexical neighbors into account. (Two words are lexical neighbors if they only deviate by one phoneme). The number of lexical neighbors was minimized by including at least one noun without any neighbors in each pair of target words.

In total, 804 unique noun pairs were generated and inserted into the carrier sentence. These sentences were then divided into three sets of 268 sentences each. Finally, each set was uttered by one of three professional female talkers (AG, KL and MH) and recorded in a professional recording studio.

ASSESSING THE SENTENCE INTELLIGIBILITY

The sentence intelligibility assessments are based on the assumption that the sentence intelligibility is inversely related to the signal-to-noise ratio (SNR) at which a listener can correctly identify the noun pair in the carrier sentence. Each set of 268 sentences (AG, KL and MH) was presented in a custom-made, speech-shaped background noise to seven young normal-hearing (NH) listeners. The noise was presented at a constant level of 65 dB SPL. The initial SNR was −8 dB, increasing to 0 dB during the experiment. The sentences were presented in repeated runs, raising the sentence level by 2 dB for each run and only including the sentences that had not been correctly identified in the previous run. The correctly identified sentences were assigned a test result value corresponding to the SNR at which they were identified. Sentences that were not identified at 0 dB SNR were not tested further, but assigned a common value of 2 dB.

The mean SNR across the seven listeners was calculated for each sentence. The results for two talkers (AG and MH) are shown in Fig. 2 (sorted separately for each talker).

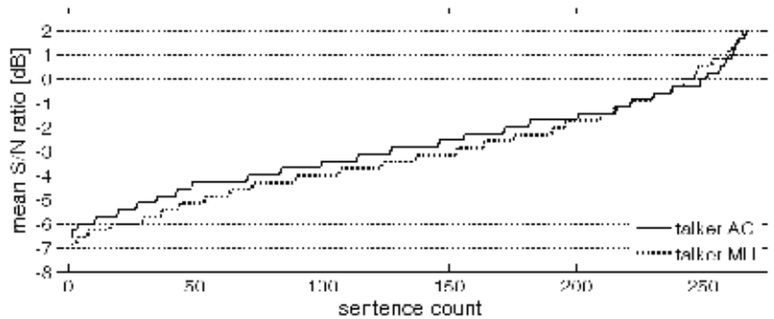

Fig. 2: The average SNR at which both nouns in the 268 sentences uttered by talkers AG and MH were correctly identified. The talkers uttered different sentences and they are thus sorted separately. A high SNR indicates that the sentence intelligibility is low. The SNR distribution for talker KL is very similar to that of AG and omitted here for clarity.

Sentences with the lowest intelligibility (a mean SNR above −1 dB) and the highest intelligibility (a mean SNR below −5.5 dB for AG and KL and −6 dB for MH) were discarded. The remaining 3 × 200 sentences were considered to be of approximately equal intelligibility and were randomly distributed into 10 lists of 20 sentences for each talker. These preliminary sentence lists formed the basis for a pilot validation of the corpus. (The intelligibility assessments were not finalized at the time of writing; the final assessments will be based on 16 NH listeners. Once these data

have been collected, the sentences for the final corpus will be selected and compiled into test lists according to well-defined criteria.)

PILOT VALIDATION WITH PRELIMINARY LISTS

Validation of the test material involves determining the normative SRT, i.e. the mean SRT for the corpus based on measurements with a group of NH listeners. In the pilot validation, the SRTs were determined using an adaptive procedure according to the guidelines in the original Hearing in Noise Test (HINT, Nilsson *et al*., 1994). Seven new NH listeners participated. Each participant listened to either talker AG (two listeners), talker MH (two listeners), or talker KL (three listeners) as the target. The set-up was free-field with the target from the front direction and the two maskers presented from ±50° as shown in Fig. 1. The two non-target talkers were used as maskers. Each masker was presented at 60 dB SPL, while the target level varied according to the listener response. The SNR was calculated relative to the individual masker level. An SNR of 0 dB thus corresponds to a target level of 60 dB SPL. The mean SRT for each talker and the corresponding within-subject standard deviations are listed in Table 1. The results indicate a significant difference in talker intelligibility. However, they also suggest that the new speech corpus results in high SRTs and that the within-subject standard deviation is fairly small.

	SRT	SD
Target AG	0.0 dB	1.2 dB
Target MH	−1.4 dB	1.2 dB
Target KL	0.7 dB	1.1 dB

Table 1: Mean SRT and mean within-subject standard deviation (SD) based on 20 SRT measurements for talkers AG and MH and 30 measurements for talker KL.

Psychometric functions were estimated for the individual listener data (Fig. 3). Each function corresponds to a given listener's responses to the 10 test lists of a given talker. The intelligibility was calculated as the percentage of correctly identified sentences (both nouns identified) for the last 16 levels of the adaptive procedure, i.e. a total of 160 responses per listener.

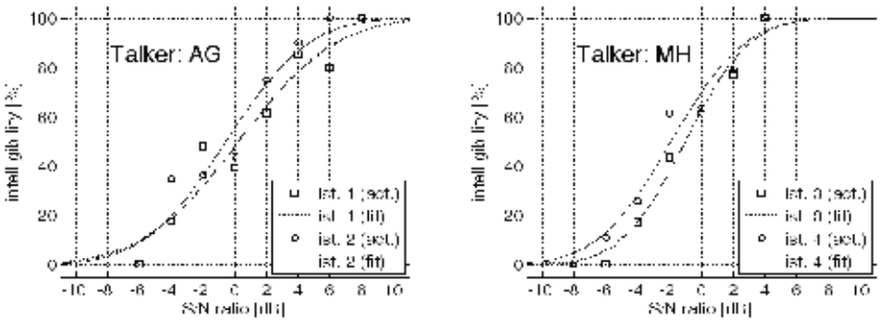

Fig. 3: Individual psychometric functions for four of the NH listeners in the pilot validation experiment. Each listener listened to 10 lists from one target talker. The curves are best fit cumulative normal distributions to the actual measurements. The steepest slope of the curves varies from 8.3%/dB (listener 1) to 12.5%/dB (listener 3).

PERSPECTIVES

The pilot validation indicates that the new speech-on-speech material is characterized by:

- a much higher SRT compared to other speech materials
- a relatively low within-subject standard deviation
- relatively steep psychometric functions

We thus expect that the material will be an effective tool for investigations of speech-on-speech masking, e.g., for measurements of speech recognition differences between different hearing-aid fittings in spatially complex, multi-talker listening conditions. The material will be made publicly available.

ACKNOWLEDGEMENTS

We would like to thank Lise Bruun Hansen, Oticon A/S, Morten Løve Jepsen, CAHR, and Thomas Ulrich Christiansen, CAHR, for their valuable contributions to this project.

REFERENCES

Clopper, C.G., Pisoni, D.B., and Tierney, A.T. (**2006**). “Effects of open-set and closed-set task demands on spoken word recognition,” J. Am. Acad. Audiol. **17**, 331-349.

Helfer, K.S. and Freyman, R.L. (**2009**). “Lexical and indexical cues in masking by competing speech,” J. Acoust. Soc. Am. **125**, 447-456.

Neher, T., Behrens, T., Carlile, S., Jin, C., Kragelund, L., Petersen, A.S., and van Schaik, A. (**2009**). “Benefit from spatial separation of multiple talkers in bilateral hearing-aid users: Effects of hearing loss, age, and cognition,” Int. J. Audiol. **48**, 758-774.

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 in noise,” J. Acoust. Soc. Am. **95**, 1085-1099.