Evaluation of a Danish speech corpus for assessment of spatial unmasking

THOMAS BEHRENS, TOBIAS NEHER, AND RENÉ BURMAND JOHANNESSON

Oticon A/S, Research Centre Eriksholm, Kongevejen 243, 3070 Snekkersten, Denmark

This paper reports on an evaluation of a new Danish speech corpus for assessment of spatial unmasking. The structure of the sentences used (Example "Michael had seven yellow boxes"), lends itself to be used in a multitalker speech intelligibility task with selective attention, by using the leading name as a call sign.

In experiment 1 the speech material was evaluated for speech recognition homogeneity of the compiled test lists. The evaluation was carried out using 9 normal hearing native Danish speaking subjects, who listened to and repeated sentences from a female target talker presented against a background of two concurrent female talkers. All sound was presented from a single loudspeaker. Data analysis, carried out on group data, revealed systematic differences in speech intelligibility, which could be related to target call sign and target talker. Based on the analysis, the selection of the speech material to use in future studies was limited to reduce spread in speech intelligibility. Training effects were observed, but they were of small magnitude.

In experiment 2 the speech material selected in experiment 1 was used for assessing spatial unmasking in a group of 9 normal hearing native Danish speaking subjects. Target sentences were always played from directly in front of the subject. Spatial unmasking was assessed in two unmasked conditions, offering different sets of spatial cues. The first condition elicited unmasking along the median plane by playing maskers from one speaker directly behind the subject. The benefit obtained by subjects was about 10 dB on average, but varied considerably. In the second unmasked condition, maskers were played from two speakers placed +/- 50 degrees to the left and right of the target. The benefit obtained by subjects was much more consistent than in the first condition, with an average effect of about 14 dB.

INTRODUCTION

In Danish speech intelligibility research, the Dantale II corpus (Wagener *et al.*, 2003) is a well-established tool. However, this corpus is not suitable for multi-talker spatial unmasking (SU) assessments. One English corpus used frequently in such assessments is the Coordinate Response Measure (CRM; Bolia *et al.*, 2000).

The research reported here was carried out within the context of enabling Danish multi-talker SU assessments. In a first experiment, a newly recorded speech corpus that was inspired by the Dantale II and CRM corpora was tested for its homogeneity in terms of speech intelligibility. In a second experiment, the corpus was used to obtain normal-hearing reference data for an SU task to be completed later by hearing-

Auditory signal processing in hearing-impaired listeners. 1st International Symposium on Auditory and Audiological Research (ISAAR 2007). T. Dau, J. M. Buchholz, J. M. Harte, T. U. Christiansen (Eds.).
ISBN: 87-990013-1-4. Print: Centertryk A/S.

impaired subjects (see companion paper by Neher *et al.*, 2007).

EXPERIMENTAL CONDITIONS

Experiment 1

Nine normal hearing adults aged 25 – 46 served as subjects.

Speech material

The Dantale sentences have the following structure “Name Verb Numeral Adjective Noun”, with a typical example being “Michael had seven yellow boxes”. The context is low, which enables scoring at the word level. For each word category, there are 10 alternatives, which can in principle be combined randomly. The combinations chosen for the Dantale lists ensure a phonetic balance.

In the present context where the corpus is to be used in a multitalker speech intelligibility task, it is necessary to have one part of the sentence identify the target sentence and another part to be used for evaluating what has been heard by the subject. With the present sentences it was decided to use the leading name as the call sign and the remaining four words for evaluation.

Recordings were made of 5 trained female talkers. Talkers were instructed to read 150 sentences from the Dantale sentence lists with normal vocal effort. These correspond to those reported on by Wagener *et al.* (2003), with the exception of list 13, which has consistently yielded lower performance in previous testing at the Eriksholm Research Centre and therefore has been excluded.

The obtained recordings were segmented into .wav files, each containing one sentence. The length and level of each file was then analysed. Following analysis, the length was adjusted +/- 5% in order to minimize the variance on the sentence duration, whilst preserving pitch. Also, the level of the individual sentences was equalised.

The Dantale sentences are structured in lists with different call signs, each containing 10 sentences. All call signs are used exactly 15 times in the recorded selection and therefore sentences have been reorganized into 10 lists, each containing 15 sentences with the same call sign.

Experimental procedure

Subjects were presented with 750 possible combinations of sentences and talkers as the target. Two maskers were randomly selected. All concurrent sentence presentation was co-located, at a target-to-masker ratio (TMR) of 0 dB. Presentation level was 65 dB(A) SPL. Stimuli were presented from a speaker placed directly in front of the subject at a distance of 1.6 m. Testing took place in an anechoic chamber. To enable analysis of masking properties of the speech corpus, the responses were coded at the word level as target, masker or extraneous. The experiment was divided into 7 sessions, each session containing 105 trials, with the exception of the last session which contained 120 trials.

Experiment 2

Nine normal hearing adults aged 25 – 46 served as subjects. Each subject took part in a training and a test session. Speech material selection from experiment 1 was used as stimuli.

Physical test setup

All training and testing was carried out under anechoic conditions. Four loudspeakers were positioned in the horizontal plane at 0°, ±50° and 180° (cf. figure. 1). The distance to the listening position was 1.6 m. Below the frontal loudspeaker an LCD screen was hung that was used for displaying instructions. The subject was seated in a custom-made chair equipped with a head rest that was small enough, so that sound reaching a subject's ears from behind was not obstructed. The chair was adjusted as necessary to ensure that the subject's head was located precisely in the middle of the test set-up and that the subject was seated comfortably. The subject was instructed to move as little as possible whenever measurements were made. This could also be checked by means of a camera and a monitor placed in the control room.

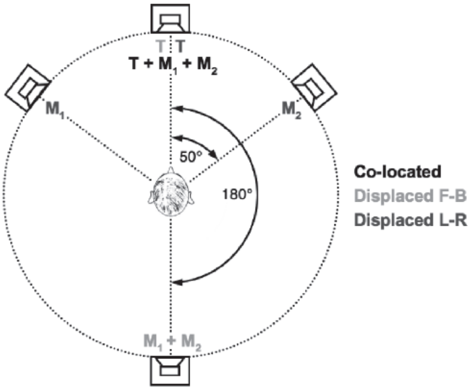

Fig. 1: Physical test set-up and spatial test condition

Spatial test conditions

To quantify SU, speech intelligibility was measured for three spatial test conditions: (1) co-located, (2) displaced F-B, and (3) displaced L-R (cf. Fig. 1). In the co-located condition, three speech signals were presented simultaneously from the frontal loudspeaker. One of the signals served as the target (T), while the other two signals served as maskers (M1 and M2). In the displaced F-B condition, the target stayed in front and the two maskers were presented from behind. In this condition, only high-frequency monaural spatial cues (e.g. Middlebrooks & Green, 1991) were therefore potentially available to the subjects to spatially separate the target from the maskers. In the displaced L-R condition, the target came still from in front, whilst the maskers were presented from the left and right loudspeaker, respectively. In this condition, both interaural and monaural spatial cues were therefore potentially available to the subjects. To allow the subject to identify the target signal in the co-located condition, the first word of the target sentence (the “call sign”) was displayed on the LCD screen. The subject’s task was to repeat all five words of the target sentence (the call sign scores were not included when estimating speech intelligibility).

To ensure that the subject was ready for the next sentence being presented, an auditory pre-warning in the form of the word “ready” was played from the frontal speaker immediately before the actual sentence.

Estimation of spatial unmasking

In order to estimate SU, 50%-correct speech intelligibility thresholds were measured for the three spatial conditions. SU was then calculated by taking the difference between the TMR corresponding to the 50%-correct threshold estimate obtained for the co-located condition and the TMR corresponding to the 50%-correct threshold estimate obtained for either the displaced F-B ($\mathrm{SU}_{\mathrm{F\text{-}B}}$) or the displaced L-R ($\mathrm{SU}_{\mathrm{L\text{-}R}}$) condition:

$$SU_{F\text{-}B} = TMR_{Co\text{-}Ioc}\text{-}TMR_{Displ\_\,F\text{-}B}, \text{ [dB]} \qquad \text{(Eq. 1)}$$

$$SU_{L\text{-}R} = TMR_{Co\text{-}Ioc}\text{-}TMR_{Displ\_\,F\text{-}B}, \text{ [dB]} \qquad \text{(Eq. 2)}$$

The individual 50%-correct threshold estimates were extracted from psychometric functions that had been derived using the method of constant stimuli. Before the start of the actual SU measurement, all subjects completed a brief “task brush-up” that included a few “easy” TMRs, so that they could get used to the different conditions again. Next, 30 “pre-trials” were run per condition to get an indication of where each subject’s threshold was likely to lie. The resultant data were then used to derive psychometric functions with the help of a maximum-likelihood estimation procedure. From these functions, a few suitably placed TMRs were extracted and another block of 30 trials was run per condition. Only the data from the pre-trials and subsequent trials were used to estimate the final thresholds.

Training program

The design of the training program was based on a gradual build-up of the task complexity, concluding with the actual SU task. It consisted of four steps, which are summarised in Table 1. At the start of each step, the subject was provided with verbal and written instructions regarding the details of the subsequent stimuli. After each stimulus presentation, the experimenter provided verbal feedback to let the subject know if a given response was correct or not. If the subject had made a mistake, the experimenter gave instructions in terms of what aspect of the stimulus to pay attention to and played the same stimulus up to two times more, so that the subject could correct his response. Such feedback provision is known to play an important role in the successful outcome of training programs, including audiological ones (e.g. Sweetow & Palmer, 2005). The duration of the training program was approximately 30 minutes.

Step	Signal(s)	What's new?	No. of trials
1	T + M1	T and M1 introduced; Task: Repeat T, locate M1	12
2	T + M1 + M2	M2 introduced; Task: Repeat T, locate M1 + M2	12
3	T + M1 + M2	Target call sign changes; Task: Repeat T	15
4	T + M1 + M2	TMRs change	45

Table 1: Training program details. For each step, the signals presented, the new elements introduced, and the number of trials run are indicated.

Experimental time course

The study comprised two separate visits per subject. At the first visit, all subjects were trained according to the program outlined in Table 1. At the second visit, SU performance was measured as described above. The decision to carry out the training program and first SU measurement on two separate days was motivated by research findings related to perceptual learning (Ortiz and Wright, 2005).

RESULTS

Experiment 1

The data extracted for further analysis by means of a repeated measures ANOVA were the average scores of each subject for each of the 50 lists presented (10 call signs * 5 talkers). It is shown in Table 2.

	Max. (%)	Min. (%)	F-value	p-value
CALL SIGN	53.9	66.2	6.66	<0.01*
TALKER	57.5	64.0	1.42	0.25
CALL SIGN*TALKER	46.3	72.2	2.14	<0.01*

Table 2: Results of repeated measures ANOVA on group list score data along with maximum and minimum observed. Statistically significant effects at the p< 0.05 level are indicated by an asterisk.

It can be seen from the table that there are statistically significant effects of call sign and call sign*talker. When observing list scores, it is seen that the spread is very large, ranging from 46 % to 72 %. It can also be observed from the table that the variance associated with call sign is larger than that associated with talker. To minimize variation in the selection used for further testing the call signs which increased variance the most were excluded. Then, the three talkers yielding minimum variance within the remaining lists were selected.

The call signs excluded were not only those that objectively gave the lowest performance, but also those ones the test subjects reported having the most problems hearing, when presented co-located in competition with two randomly selected maskers.

The results of a new repeated measures ANOVA on the final selection of sentence lists is shown in Table 3. It showed the following.

	Max. (%)	Min. (%)	F-value	p-value
CALL SIGN	61.2	67.2	1.48	0.22
TALKER	62.9	66.8	0.61	0.56
CALL SIGN*TALKER	58.9	70.6	1.56	0.13

Table 3: Results of repeated measures ANOVA on group list score data from the selected lists along with maximum and minimum observed. Statistically significant effects at the $p< 0.05$ level are indicated by an asterisk.

It can be observed that neither the call sign effect, nor the call sign*talker effect is statistically significant. This is in line with a post-hoc analysis (Bonferroni), which revealed no statistically significant differences between the list scores. It can further be seen from the data on the call sign*talker effect that the variation in list scores has now been reduced to a 12% interval. This is in line with the variation reported for QuickSIN (McArdle and Wilson, 2006) and Dantale (Wagener *et al.*, 2003).

Training Effects

These were found to be low compared to other speech intelligibility tasks. The training effects found fell into two categories. One group showed training effects on the order of 0.04% per 10 sentences presented. For the other group the corresponding number was 0.25% per 10 sentences presented.

Masking Effects

The purpose of this analysis is to obtain data on the type of masking occurring in the present multitalker experiment. Results are given in Fig. 2 below. At a 0 dB target-to-masker ratio the responses given by subjects are on average 60.6% from the target talker, 15.5% from the masking talkers and the remaining 23.9 % were responses extraneous to what was presented.

Figure 2 also compares data from the current study with similar data from corpora developed for similar purposes, such as the Coordinate Response Measure, CRM (Brungart *et al.*, 2001), and the TVM Sentences (Helfer & Freyman, 2006). The TMR for the CRM data was 0 dB, whilst for the TVM sentences a SNR of 2 dB was used. Another difference between the corpora is that the CRM is a closed set, where subjects are given all response possibilities on a touch screen, whereas in the TVM and the current study subjects are asked to repeat what was heard. Finally, the CRM material has a higher pace than the Dantale material. This may impose difficulties in older listeners

The comparison reveals that the amount of responses in the current study that were masker words, are comparable to what is found with the CRM corpus and somewhat more than what has been found with the TVM sentences.

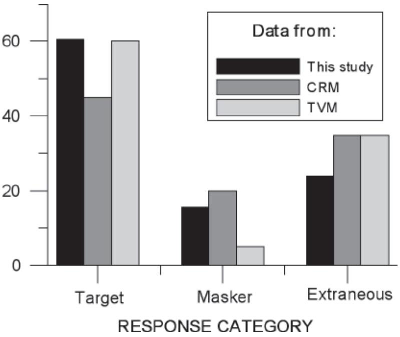

Fig.2: Average response pattern by subjects being subjected to the corpus evaluated in this study along with data from the CRM and TVM corpora.

Experiment 2

The results of experiment 2 in terms of thresholds in the three tested spatial conditions along with the accompanying spatial unmasking is given in the left and right panels of Fig. 3.

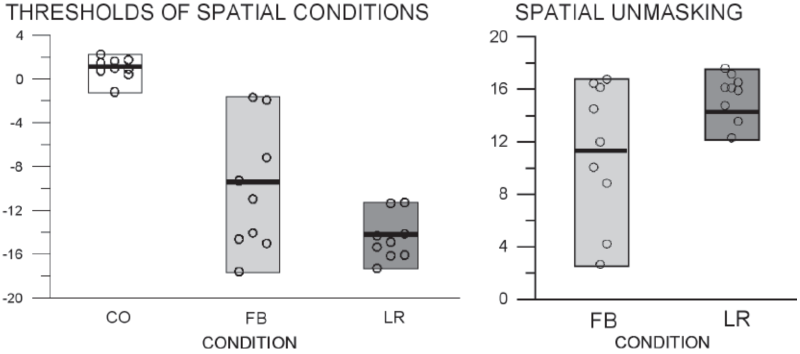

Fig. 3: Thresholds of the three different conditions (left) used for estimating spatial unmasking as shown on the right.

It can be observed in the right panel of figure 4 that there is a large variation in performance within and across spatial conditions. In the co-located condition, performance is generally quite homogenous; the average speech reception threshold is 1.1 dB, with maximum and minimum being 2.2 and -1.2 dB. In the displaced F-B condition performance varies substantially, maximum and minimum values range from -1.7 dB to -17.6 dB; the average is -9.4 dB. In the displaced L-R condition performance is more homogenous; the average is -14.2 dB with maximum and minimum being -11.3 dB and -17.3 dB.

The spread observed in the displaced F-B and displaced L-R thresholds are reflected in the results seen for spatial unmasking in the right panel of Fig. 3 The amount of spatial unmasking observed for the displaced F-B condition varies from 2.6 dB to 16.7 dB, with an average of 11.2 dB. In the displaced L-R condition performance varies from 12.2 dB to 17.5 dB with an average of 15.5 dB.

Whilst there do not seem to be other published data for the displaced F-B condition, the displaced L-R data can be compared to those reported by Marrone *et al.* (2007). In that study, spatial unmasking was assessed using the CRM corpus using young normal hearing adults. The amount of spatial unmasking reported by Marrone *et al.* (2007) was on average 12.6 dB with a range of 8-15 dB for an unmasked condition with maskers placed at +/- 90 degrees. The experiment was conducted in a typical IAC booth. Thus, the experimental conditions used in this study seem to yield unmasking of a larger magnitude than what has previously been reported.

CONCLUSION

A new Danish speech corpus for use in multitalker speech intelligibility research has been recorded and evaluated. In a first experiment systematic differences in speech intelligibility as a function of call sign as well as call sign and talker interactions were found. Therefore, the amount of the material to be used in future research was limited to what yielded homogeneous performance in normal hearing subjects. The amount of variation between lists in the final selection was similar to what has been found for other corpora used for speech intelligibility research. Training effects were analysed and found to be low. Masking effects were also analysed and it was observed that the corpus seems to produce considerable amounts of informational masking, as indicated by the relatively large proportion of masker words repeated by the subjects.

In a second experiment, the final selection of sentence material that came out of the first experiment was used. Spatial unmasking for the displaced L-R condition ranged from about 12 dB to about 18 dB. This is more than what has been obtained in other studies. Whether this is due to the corpus, the training or the anechoic conditions is yet unknown. Spatial unmasking for the displaced F-B condition varied substantially more than the displaced L-R condition. The range found was 2 dB to 16 dB. Further research is necessary to show the exact reasons.

REFERENCES

Bolia, R., Ericson, M., and Simpson, B. (**2000**). “ A speech corpus for multitalker communications research,” J. Acoust. Soc. Am, **107**, No. 2

Brungart, D. S., Simpson, B. D., Ericson, M. A., and Kimberly, R. S. (**2001**). “Informational and energetic masking effects in the perception of multiple simultaneous talkers,” J. Acoust. Soc. Am., **110**, No. 5

Helfer, K. S., and Freyman, R. L. (**2006**). “How do task instructions influence speech-on-speech masking?,” Poster presented at 151st Meeting of the Acoust. Soc. Am.

Neher, T., Behrens, T., Kragelund, L., and Petersen, A.S. (**2007**). “Spatial Unmasking in Aided Hearing-Impaired Listeners and the Need for Training.” Proceedings of the International Symposium on Audiological and Auditory Research, Helsingør, Denmark, Aug. 29-31.

Marrone, N. L., Mason, C. R., and Kidd, G.Jr. (**2007**). "Spatial Release from Masking; Hearing Loss, Age and Reverberation Effects," Poster presented at Am. Aud. Soc. Meeting, Phoenix, NM

McArdle, R. A., and Wilson, R. H., (**2006**). "Homogeneity of the 18 QuickSIN Lists," J. Am. Acad. Audiol., **17**, No. 3

Middlebrooks, J.C., and Green, D.M. (**1991**). "Sound localization by human listeners," Ann. Rev. Psychol., **42**, 135-159.

Ortiz, J. A., and Wright, B.A. (**2005**). "Effects of different amounts of brief training and rest on the generalization of learning from interaural-level-difference to interaural-time-difference discrimination," J. Acoust. Soc. Am., **117**, 2561.

Sweetow, R. and Palmer, C. V. (**2005**). "Efficacy of individual auditory training in adults: A systematic review of the evidence," J. Am. Acad. Audiol., **16**, 494-504.

"Design, Optimization and Evaluation of a Danish Sentence Test in Noise," Int. J. Audiol., **42**, No.1.