An improved competing voices test for test of attention

LARS BRAMSLØW*, MARIANNA VATTI, RIKKE ROSSING,
AND NIELS HENRIK PONTOPPIDAN

Eriksholm Research Centre, Snekkersten, Denmark

People with hearing impairment find competing voices scenarios to be challenging in terms of their ability to switch attention and adapt to the situation. With the Competing Voices Test (CVT), we can explore how they can adapt and change their attention between voices. The CVT provides three male and three female speakers, played in pairs. The task of the listener is to repeat the target sentence. Three methods of cueing the listener to the target sentence were tested: a male/female cue (for male-female sentence pairs), an audio voice cue and a text cue using one word from the target sentence. The cue was presented either before or after the sentence pair playback. The CVT was evaluated on 14 moderate-severely hearing impaired listeners with four spatial conditions: summed (diotic), separate (dichotic) plus two types of ideal masks for separating the two speakers from the sum. The results show that the test is sensitive to the spatial conditions, as intended. The text cue is the most sensitive to spatial condition. The text cue has the further advantage that it can be used for, e.g., male-male speaker pairs as well. Furthermore, the applied ideal masks show test scores very close to the ideal separate spatial condition.

INTRODUCTION

Competing voices are part of the everyday challenges for a hearing aid user. This might occur for instance while attending to two voices in a restaurant or while watching TV and attending to a voice in the room at the same time. In order to test the performance of hearing aids in this user scenario, a new type of speech test has been developed.

Compared to traditional speech tests, a competing voices test would have two or more targets that are equally important to follow, and in the simplified case no masker. Tests of this type have been reported in the literature (Mackersie *et al.*, 2001; Helfer *et al.*, 2010), but no particular test has been put to common use. Furthermore, they are not available in Danish.

The purpose of the present project was to develop a competing voices test (CVT) in Danish. The proposed CVT has evolved in a number of iterations and applications using other cue timings and different speech material, this is documented in a series of posters (Bramsløw *et al.*, 2014, 2015a, 2015b, 2016b). The present paper presents the newest and improved version of the CVT.

*Corresponding author: labw@eriksholm.com

Proceedings of the International Symposium on Auditory and Audiological Research (Proc. ISAAR), Vol. 6: Adaptive Processes in Hearing, August 2017, Nyborg, Denmark. Edited by S. Santurette, T. Dau, J. C.-Dalsgaard, L. Tranebjærg, T. Andersen, and T. Poulsen. The Danavox Jubilee Foundation, 2017.
 ISBN: 978-87-990013-6-1.

The aim of the present study was thus to refine and validate the competing voice test. The CVT should have the following properties:

- Be sensitive to signal processing contrasts, in this case spatial contrasts;
- Be applicable for older hearing-impaired listeners without floor and ceiling effects in the outcome measure;
- Be suitable for quick testing of multiple conditions in the laboratory.

METHOD

Speech material

In order to minimize development time, the Danish Hearing in Noise Test (HINT) was chosen as the speech corpus for the CVT, being an established and well documented natural sentences speech material (Nielsen and Dau, 2009; 2011). The Danish HINT has five words per sentence and is available with one male speaker.

Two males and three females were recruited as additional speakers. The recording was conducted as follows: The speaker was located in an audiometry booth with a microphone and a PC installed with our Danish HINT test software. The speaker would use the software to play one sentence and then repeat the sentence using the same intonation as the original speaker. This was done to ensure recordings of the same vocal quality. Each list was recorded in one take, but recorded twice to have two versions. All sentences were cut out into separate wave files, and the better of the two sentences was chosen. Each sentence was now time aligned to the original male recording by estimating the cross correlation against original recording and time shifting the new recording accordingly. Then, each sentence was level adjusted to have the same RMS value as the same original male sentence.

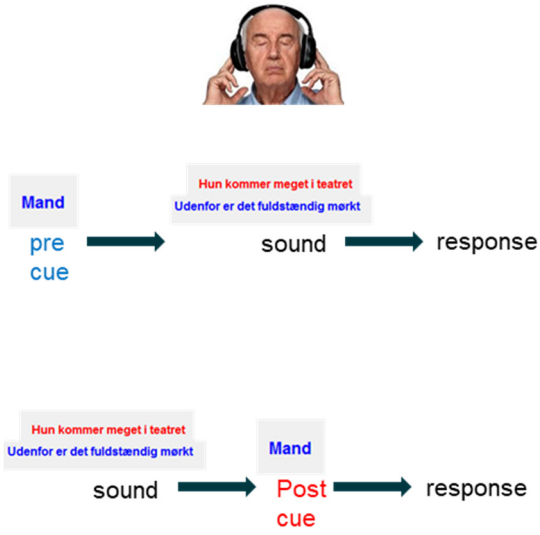

Fig. 1: Example of the Competing Voices Test with two sentences played simultaneously in a pair and the cue (male/female speaker) as showed on a screen.

Test procedure

Each CVT trial presented two sentences in pairs by selecting two different lists and randomizing the sentence order. Within each trial, all the included speaker pairs were presented in random order. In the separated (dichotic) cases, the target speaker was furthermore randomized to the left or right ear in order to make the test as unpredictable as possible for the listener. The task of the listener was to repeat the target sentence as cued by a sign on a PC monitor.

The cue could be presented before playback ('pre') or after playback ('post'). The pre cue corresponds to a classical target-masker scenario, whereas the post cue requires equal attention to both speakers, which we refer to as the 'competing voices scenario'. This is illustrated in Fig. 1.

Three cue types were tested: Audio, Text and Talker. The Audio cue is the word 'Tomato' spoken by the target speaker, thus the listener must recognise that voice in the mixture and repeat that target sentence. The Text cue is showing the first or last word from the target sentence on a screen in front of the listener: With pre cue, it will be the first word and with post cue, the last word. The words score can then be 0-4. Finally, the Talker cue uses a male-female mixture and the screen is indicating male or female to identify the target sentence.

Fourteen hearing-impaired listeners with moderate sloping hearing loss participated in the test; These are labelled Test Persons (TP) in the following.

Spatial contrasts

The sensitivity of the CVT was assessed by testing four spatial conditions: Sum (diotic), separate (dichotic), ideal binary mask (IBM) and ideal ratio mask (IRM). The ideal time-frequency masks were calculated by comparing the energy of the two clean signals in 125-Hz by 4-ms bins – as either binary masks (gain 0 or 1) or ratio masks (gain 0-1) (Naithani *et al.*, 2017). These masks were applied to the Sum signal to make an artificial separation, which was then presented dichotically. The ideal mask conditions were included to have a larger diversity of spatial conditions.

Test design

The overall test design thus consisted of the following experimental factors and levels:

- Spatial Processing: Sum, Separate, IRM, IBM
- Cuetype: Audio, Text, Talker
- Cuetime: Pre, Post
- Gender mix: Male-Female (MF), Male-Male (MM), Female-Female (FF).
- 14 test persons (TP).

The first three conditions were rotated across test persons in a balanced Latin square order, while the gender mix was varied randomly, within a given 20-pair trial. The lists order across trials was randomized such that no lists were repeated in successive

trials. Finally, the sentence order within trials was randomized such that all sentences were used equally and that the initial or last words were different in the 'Text' cuetype.

RESULTS

The outcome measure from each sentence pair was a percent correct word score, based on five words (Audio cue, Talker cue) or four words (Text cue). It was then rau-transformed to provide better 'normal' distribution of the data (Studebaker, 1985). The rau scores are practically equal to %-scores in the 10-90 range and extended beyond those limits to cover the range −18 to +118.

All data were analysed using a mixed-model analysis of variance (ANOVA) with TP as a random factor and gender mix nested under cuetype (the Talker cuetype can only use the MF combinations). The ANOVA table is shown in Table 1 below.

All main effects are significant and so are the two-way interactions spatial*cuetype, spatial*gender and the three-way interaction spatial*cuetype*gender. Regarding the random factor TP effects, the TP*cuetime interaction is significant. It is also interesting to note that there are no significant interactions between cuetime and the other fixed conditions; This means that the choice between 'Pre' and 'Post' cue may be used to set the overall performance in a future application of the CVT, if both cue timing options are considered valid use cases in the given application.

	Effect (Fixed/ Random)	SS	df	MS	Syn df	Syn MS	F	p
Intercept	*Fixed*	*3383251*	*1*	*3383251*	*12.84*	*6845.43*	*494.24*	*0.00*
spatial	*Fixed*	*40281*	*3*	*13427*	*46.52*	*360.58*	*37.24*	*0.00*
cuetype	*Fixed*	*102637*	*2*	*51318*	*29.31*	*330.72*	*155.17*	*0.00*
cuetime	*Fixed*	*57586*	*1*	*57586*	*16.83*	*543.51*	*105.95*	*0.00*
gender(cuetype)	*Fixed*	*22044*	*2*	*11022*	*659.00*	*288.49*	*38.21*	*0.00*
spatial*cuetype	*Fixed*	*15756*	*6*	*2626*	*659.00*	*288.49*	*9.10*	*0.00*
spatial*cuetime	Fixed	635	3	212	659.00	288.49	0.73	0.53
cuetype*cuetime	Fixed	1574	2	787	659.00	288.49	2.73	0.07
spatial*gender	*Fixed*	*7420*	*6*	*1237*	*659.00*	*288.49*	*4.29*	*0.00*
spatial*cuetype* gender	*Fixed*	*7018*	*6*	*1170*	*659.00*	*288.49*	*4.05*	*0.00*
TP	*Random*	*77667*	*13*	*5974*	*18.10*	*643.71*	*9.28*	*0.00*
TP*spatial	Random	14395	39	369	659.00	288.49	1.28	0.12
TP*cuetype	Random	8677	26	334	659.00	288.49	1.16	0.27
TP*cuetime	*Random*	*8048*	*13*	*619*	*659.00*	*288.49*	*2.15*	*0.01*
Error		190113	659	288				

Table 1: Summary of Analysis of Variance (ANOVA). Significant effects ($p < 0.05$) are shown in italics.

Figure 2 shows the combined effect of spatial and cuetype. The largest sensitivity to the spatial contrast is shown for the Text cuetype, with sum score at 58 rau and the three separated conditions around 85 rau, i.e., an effect of approx. 27 rau: the Tukey HSD post-hoc test is significant at $p < 0.00002$. A smaller, but significant, contrast of 15 rau is shown for the Talker cue between Sum and Separate (Tukey HSD: $p < 0.03$). The Audio cuetype has no significant differences across the spatial conditions.

The main effect of cuetiming is also significant with mean scores at 78 rau for Pre and 60 rau for Post (not shown). Interestingly, the cue timing does not interact with any other factors than TP: Thus, cuetiming (Pre vs Post) could be used to shift the overall performance down in a given test, by altering the test paradigm from target-masker to competing voices dual attention. The only interaction with cuetiming is the TP interaction (not shown), indicating that different persons have different gains by going from Post to Pre, which can be explained by the added cognitive load for the Post timing.

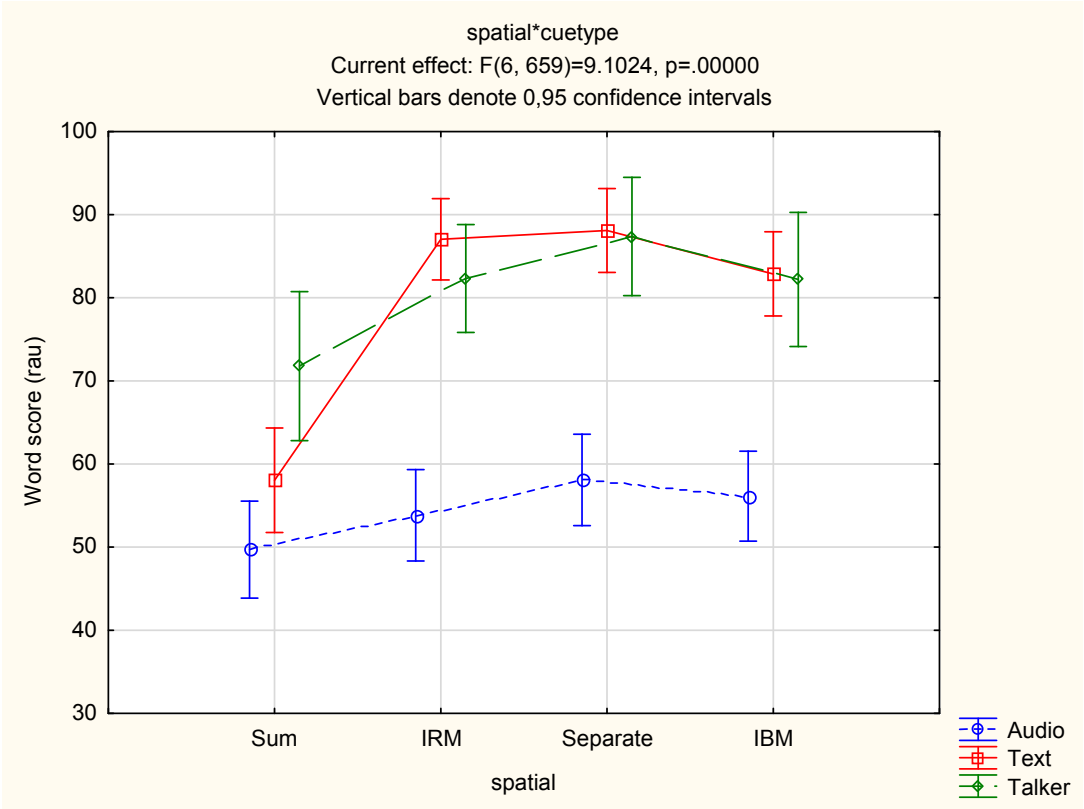

Fig. 2: The combined effect of Spatial and Cue type. The 'Text' cuetype shows the largest contrast between the spatial conditions.

Concerning the spatial modes, we find a large effect of 27 rau between Sum and the three other modes ($p < 0.00002$), and the ideal masks (IBM and IRM) are not significantly different from the perfect separation in Separate. The difference between Sum and Separate is 30 rau, which is a higher contrast than 22 rau obtained in a previous version of the CVT (Bramsløw *et al.*, 2016b).

Regarding gender mix, the test should ideally be insensitive to the gender mixes in order to have a free choice when designing new tests. These results are shown in Fig. 3. The Text cue shows no significant effect of gender mix, while the Audio cue shows a large, significant effect size going from 73 rau to 45 rau (Tukey HSD, $p < 0.00003$). The MM and FF (same gender) pairs have low scores, indicating that the two voices are easily confused when they are same gender, causing a high risk of missing what the target is. The Talker cue is robust as the Text cue, but logically only available for the male-female speaker pairs.

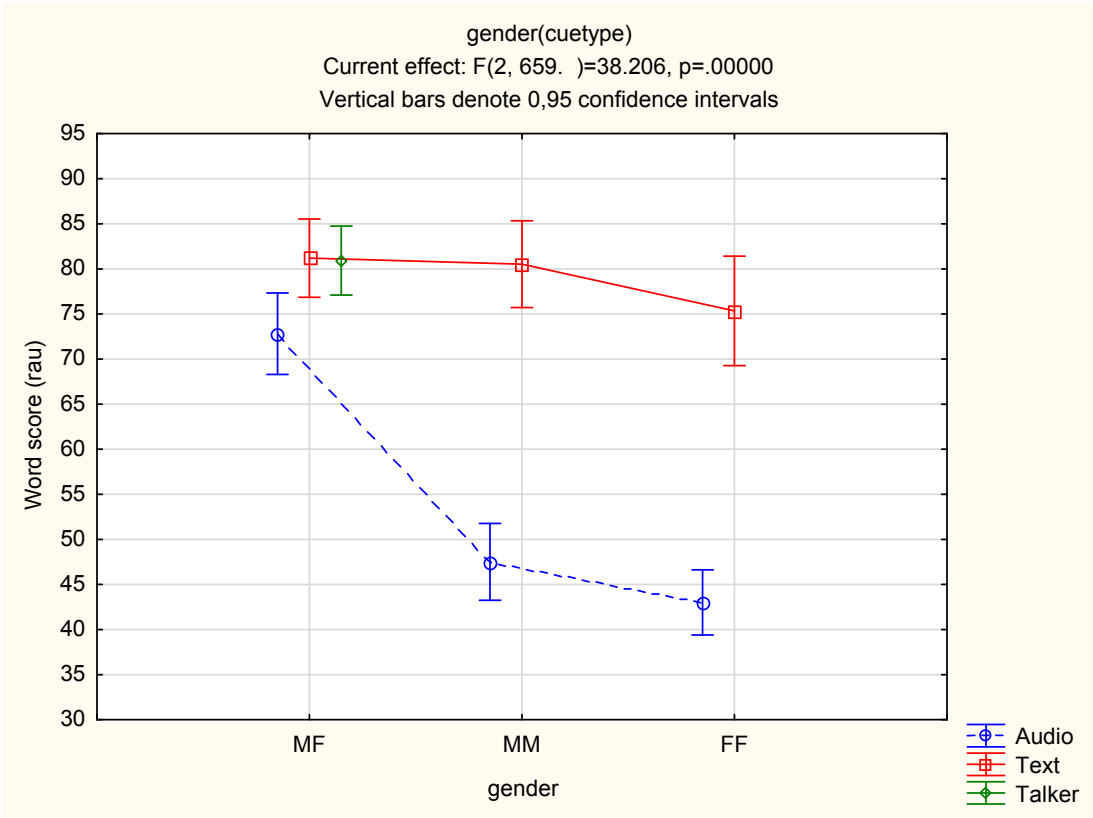

Fig. 3: The combined effect of gender mix and cuetype.

CONCLUSIONS

The CVT is now validated and may be used to evaluate signal processing algorithms such as noise reduction or speech separation algorithms. For the hearing-impaired listeners here we get scores between approx. 40 rau and 90 rau, which is in the middle between floor and ceiling. This should be compared to normal-hearing listeners, who score close to 100, i.e., close to ceiling (Bramsløw *et al.*, 2014). In general, scores do not go far below 50 rau, which may be due to listeners choosing, e.g., one ear consistently, regardless of the cue, which results in a 50% chance level if the intelligibility of a previously chosen target voice is close to 100%.

Among the three cue types Audio, Text and Talker, the Text cue was the most sensitive, providing a 30 rau contrast between Sum and Separate, compared to previously 22 rau (Bramsløw *et al.*, 2016b). The Text cue is recommended for future applications. Regarding the cue timing, the choice between Pre and Post does not affect the sensitivity to the other experimental factors, so it may in future tests be chosen to keep the scores away from floor and ceiling.

Regarding the test of ideal masks with the given time-frequency resolution, the two ideal masks, IRM and IBM are as good as the separated signals. Thus, the applied time-frequency masks are appropriate for testing of different mask-based speech separation algorithms.

When using the CVT, reuse of the ten HINT lists is unavoidable, as each trial uses two lists. Therefore, learning will take place (Bramsløw *et al.*, 2016a), and this needs to be addressed by proper balancing of the test conditions across listeners.

ACKNOWLEDGEMENTS

Thanks to Boris Søndersted, who edited all recordings into single sentence wave files.

REFERENCES

Bramsløw, L., Vatti, M., Hietkamp, R., and Pontoppidan, N.H. (**2014**). "Design of a competing voices test," International Hearing Aid Conference, Lake Tahoe, CA, USA. Retrieved from http://www.eriksholm.com/about-us/news/IHCON_2014

Bramsløw, L., Vatti, M., Hietkamp, R.K., and Pontoppidan, N.H. (**2015a**). "Binaural speech recognition for normal-hearing and hearing- impaired listeners in a competing voice test," Speech in Noise Workshop, Copenhagen, Retrieved from http://www.eriksholm.com/about-us/news/2015/SPIN_2015

Bramsløw, L., Vatti, M., Hietkamp, R., and Pontoppidan, N.H. (**2015b**). "Best application of head related transfer functions for competing voices speech recognition in hearing-impaired," International Symposium on Auditory and Audiological Research, Nyborg, Denmark.

Bramsløw, L., Simonsen, L.B., Hichou, M. El, Hashem, R., and Hietkamp, R.K. (**2016a**). "Learning effects as result of multiple exposures to Danish HINT," International Hearing Aid Conference, Lake Tahoe, CA, USA. Retrieved from http://www.eriksholm.com/about-us/news/IHCON_2016.aspx

Bramsløw, L., Vatti, M., Hietkamp, R.K., and Pontoppidan, N.H. (**2016b**). "A new competing voices test paradigm to test spatial effects and algorithms in hearing aids," International Hearing Aid Conference, Lake Tahoe, CA, USA. Retrieved from http://www.eriksholm.com/about-us/news/IHCON_2016.aspx

Helfer, K.S., Chevalier, J., and Freyman, R.L. (**2010**). "Aging, spatial cues, and single- versus dual-task performance in competing speech perception," J. Acoust. Soc. Am., **128**, 3625–3633. doi: 10.1121/1.3502462

Mackersie, C.L., Prida, T.L., and Stiles, D. (**2001**). "The role of sequential stream segregation and frequency selectivity in the perception of simultaneous sentences by listeners with sensorineural hearing loss," J. Speech Lang. Hear. Res., **44**, 19-28. doi: 10.1044/1092-4388(2001/002)

Naithani, G., Barker, T., Parascandolo, G., Bramsløw, L., Pontoppidan, N.H., and Virtanen, T. (**2017**). "Low-latency sound source separation using convolutional recurrent deep neural networks," IEEE Work. Appl. Signal Process. to Audio Acoust., New Paltz, NY.

Nielsen, J.B., and Dau, T. (**2009**). "Development of a Danish speech intelligibility test," Int. J. Audiol., **48**, 729-741. doi: 10.1080/14992020903019312

Nielsen, J.B., and Dau, T. (**2011**). "The Danish hearing in noise test," Int. J. Audiol., **50**, 202-208. doi: 10.3109/14992027.2010.524254

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. doi: 10.1121/1.408469

Studebaker, G.A. (**1985**). "A 'rationalized' arcsine transform," J. Speech Lang. Hear. Res., **28**, 455. doi: 10.1044/jshr.2803.455