Comparing a BMVDR with an IBM regarding SRT and subjective listening effort

Ewald Strasser[1], Jan Rennies[2], Thomas Brand[1]

[1] *Carl-von-Ossietzky University, Oldenburg, Germany E-Mail: Ewald.Strasser@uni-oldenburg.de*

[2] *Fraunhofer Institute for Digital Media Technology IDMT, Hearing, Speech and Audio Technology, Oldenburg, Germany*

Introduction

The aim of our study was to investigate the effects of a binaural minimum-variance distortionless response beamformer (BMVDR) [1] [2] and an ideal binary mask (IBM) [3] [4] in speech-on-speech masking conditions on perceived listening effort and speech intelligibility in comparison to unprocessed stimuli. There are some studies that investigate the effects of these algorithms (e.g., [5], [6], [7], [3], [4], [8], [9], [10]) but neither of them compared them in the same study nor were intelligibility and listening effort measured in the same study. One specific goal was to measure how the BMVDR algorithm, which is online-capable and could be implemented in an actual hearing device, compared to an IBM algorithm which requires oracle knowledge of the stimuli.

For our evaluation we used two different spatial listening scenarios and three different signal processing strategies. The first condition consisted of two interfering speakers that were situated at +90° and -90° and the target at 0°. In the second condition all speakers were located at 0°. The algorithms IBM and BMVDR-N as well as an unprocessed condition were used under these spatial constellations. We measured the speech recognition thresholds (SRTs) and categorically scaled subjective listening effort [11] in the same session.

Methods

Subjects

Twenty subjects aged between 19 and 29 years participated in this study. All subjects were German native speakers and had self-proclaimed normal hearing. Subjects were paid for their participation and gave informed consent before testing.

Stimuli

The stimuli consisted of a target talker and two interfering talkers. All talkers were male. The sentences of the target talker were taken from the German Oldenburg matrix sentences test (OlSa) [12]. These sentences have the fixed structure of *name word, verb, numeral, adjective, object*. For each word, ten alternatives are available which can be randomly combined to produce syntactically correct but semantically unpredictable sentences which cannot be memorized by the subjects. The interfering talkers also uttered sentences from the OlSa test, but were recorded with different talkers [13]. The target talker had an average speaking rate of 3.8 syllables per second and the interfering talkers had an average speaking rate of 5.8 syllables per second. When subjects tried to focus on the target talker, at least without any strong unmasking cues such as, e.g., spatial separation [14] this stimuli should be perceived as rather challenging.

The interfering speech of each of the two maskers was generated by concatenating three randomly assigned sentences, and then selecting a random starting point within these sentences. The earliest possible start was at the beginning of the 3-sentence string and the latest possible start was the point at which the rest of the three-sentence string was equal to the length of the current target sentence.

To render the desired spatial location of the different talkers we used head-related room impulse responses (HRIRs) of [15]. The target was always presented from the front (0°). The interfering talkers were either presented at azimuthal angles of +90° and –90° or also from the front (co-located, 0°). To avoid an effect of (long-term) better-ear listening in the spatially separated condition, the talkers were equalized in long-term RMS level. The presentation level of the combined target-interferer signal was fixed at 65 dB SPL. The SNR for all signals was defined as the ratio between the target and the sum of both interferers after the convolution with the HRIRs, but before any signal enhancement.

The setup was calibrated to 100 dB sound pressure level (SPL) using a Brüel and Kjær (B&K) 4153 artificial ear, a B&K 4947 ½ inch microphone, a B&K ZC-0032 preamplifier, and a B&K 2250 sound level meter. Stimuli were presented to the subjects via Vic Firth SIH2headphones.

Algorithms

For the BMVDR and the IBM the processing was applied to these signals. Depending on how much of the interfering talkers energy was removed by the processing, the resulting signals could have lower level.

The BMVDR used a binaural beamformer and mixed a portion of the original signal back in to preserve some binaural cues. The proportion of original signal to beamformed signal is determined by the parameter η. For the present study, we used an η of .36 which is often cited (e.g., [8; 5; 10; 9]) as a good trade-off between preserving binaural cues while still getting the energetic advantages of the beamformer.

For the IBM, both, the target and the masker signal were separated into 128 frequency channels (80 Hz to 8 kHz) and 20 ms time windows with an overlap of 10 ms. In the resulting time-frequency (T-F) representation, glimpses were categorized in tiles in which the energy of the target was higher than the summed energy of the masking talkers or not. Only the glimpses that contained more energy from the target compared to the interferers (>0dB) were kept. All other T-F tiles were set to zero before reconstructing the time-domain signal. This type of processing assumes that T-F tiles, in which the target energy is lower than the maker energy, are not accessible to the subject due to energetic masking. Hence, deleting these tiles does not remove useful target information. It does, however, remove a considerable portion of the masker, and practically renders the masking talkers

unintelligible. This type of processing was shown to considerably reduce the impact of information masking because explicit confusions of target and masker words are no longer possible [6].

Procedures

Subjects performed an SRT test and a listening effort test for each of the six combinations of spatial constellation and signal processing. Whether the intelligibility test or the listening effort test was done first was predetermined with a permutation plan. The order of the six different conditions was randomized. Each of the two tests was preceded by two training conditions that made sure that subjects understood the test and to familiarize them with the OlSa [12] procedure and speech material.

The SRT measurement was conducted using an open-source Matlab framework [16]. For each condition, one test list of 20 OlSa sentences was used. Subjects used a GUI displaying the entire matrix of 50 words (5 words x 10 alternatives per word) as pushbuttons to indicate the words they had heard. Depending on the number of correctly recognized words, the SNR in the next trial was adjusted according to an adaptive procedure [17] to converge to the SRT.

Subjective listening effort was measured with the Adaptive Categorical Listening Effort Scaling (ACALES) [11]. The ACALES method utilizes the effort scale categorical units (ESCUS) with 13 categories, which ranges from 'extreme effort' (13) to 'no effort' (1), with an extra category 'only noise' for trials in which subjects cannot detect the target talker at all. The task of the subjects was to rate how effortful it was to listen to the previously presented sentence of the target talker.

Test facilities

Both above-described procedures were repeated at two different days, once in a listening booth and once in the lobby of the listening booth. To get an idea if such studies could reliably be done in a less controlled setting, we compared the data gathered in the lobby and the listening booth. The effect of the two locations was considered in our statistical analysis but it was found to be negligible. This aspect of the study is also beyond the scope of the present article and will not be discussed. Whether the booth or the ML came first was predetermined with the permutation plan. The second test was done at least one day and at most a week after the first test.

Results

For the analysis we fitted two different linear mixed models with SRT and listening effort as the dependent variables. The analysis was done with R [18], the package lme4 [19] and modelbased [20].

SRT

An overview over the SRT data can be found in Figure 1.

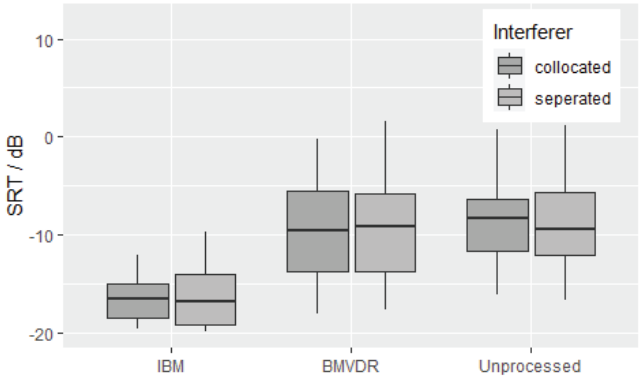

Figure 1: SRT in dB for all conditions. Algorithms are grouped along the x-axis, spatial conditions are colour-coded.

We fitted a linear mixed model (estimated using REML and nloptwrap optimizer) to predict the SRT with spatial location (collocated, separated), algorithmic processing (IBM, BMVDR, Unprocessed) and test location (booth, lobby). The model also included interactions between the algorithms and spatial conditions and a random intercept for each participant. The model explained about 76% of the observed variance. The model's intercept was defined as the IBM, collocated and tested at the booth. The intercept was estimated to be at -15.94 dB (95% CI [-16.96, -14.93], t(219)=-30.87, p<.001). Within this model we found the following effects.

The effect of spatial separation was statistically significant and negative (beta=-3.99, 95% CI [-5.32, -2.66], t(219) = -5.90, p<.001).

The effect of BMVDR was statistically significant and positive (beta=10.65, 95% CI [9.31, 11.98], t(219)=15.74, p<.001).

The effect of the unprocessed condition was statistically significant and positive (beta=10.18, 95% CI [8.85, 11.51], t(219)=15.05, p<.001).

The effect of test location was statistically non-significant and positive (beta = -0.08, 95% CI [-0.85, .69], t(219)=-0.21, p=0.836; Std. beta=-0.01).

The interaction effect of BMVDR and spatial separated speech was statistically significant and negative (beta=-4.13, 95% CI [-6.01, -2.24], t(219)=-4.31, p<.001).

The interaction effect of the unprocessed speech and spatial separation was statistically significant and negative (beta=-2.07, 95% CI [-3.95, -0.18], t(219)=-2.16, p=0.032).

Of particular interest for the present study was the question how the algorithms performed compared to each other. Therefore, we did a marginal contrast analysis to see how the algorithms compared within the same spatial condition. We found that when all speakers were collocated the IBM outperformed the BMVDR (-10.7 dB difference, t(204)=-15.78, p<.001) and the unprocessed condition (-10.2 dB difference, t(204)=-15.08, p<.001). We found no significant difference between the BMVDR and the unprocessed condition at the collocated condition (0.5 dB difference, t(204)=0.69, p<.982).

When target and interferers were separated the IBM also outperformed the BMVDR (-6.5 dB difference, t(204)=-9.66, p<.001) and the unprocessed condition (-8.1 dB difference, t(204)=-12.02, p<.001). We found no significant difference between the BMVDR-N and the unprocessed condition at the

spatially separated condition (-1.6 dB difference, t(204)=-2.36, p<.177).

Listening effort

A first look at the raw data of the ACALES test can be found in Figure 2.

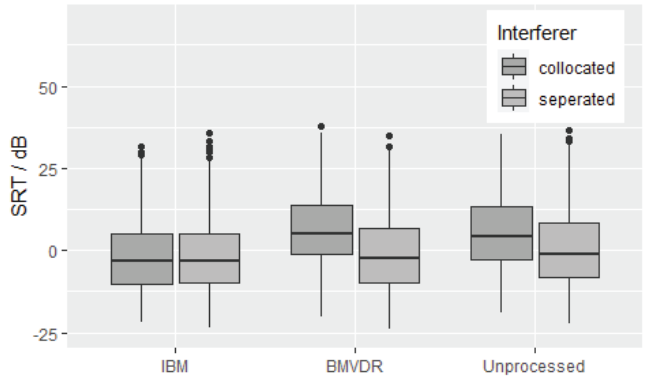

Figure 2: Boxplot of the ACALES results over all effort scale units (ESCUS). Algorithms are grouped along the x-axis; spatial conditions are colour-coded.

We fitted a linear mixed model (estimated using REML and nloptwrap optimizer) to predict listening effort with spatial location (collocated, separated), algorithmic processing (IBM, BMVDR, Unprocessed), test location (booth, lobby) and the ESCUS scale. The model included a random intercept for each participant and it also included interactions between the algorithms and spatial conditions. The model explained about 88% of the observed variance (conditional R^2=0.88) and the part related to the fixed effects alone was 61%.
The model's intercept was defined as the IBM, collocated, and tested at the booth. The intercept was estimated to be at 14.6 dB (95% CI [11.77, 17.43], t(2954)=10.12, p<.001). Within this model we found the following effects.
The effect of spatial location was statistically non-significant and positive (beta=0.26, 95% CI [-0.26, .79], t(2954)=0.99, p=0.323).
The effect of the BMVDR-N was statistically significant and positive (beta=8.14, 95% CI [7.62, 8.67], t(2954)=30.44, p<.001).
The effect of the unprocessed condition was statistically significant and positive (beta=7.22, 95% CI [6.70, 7.75], t(2954)=27.00, p<.001).
The effect of test location was statistically non-significant and negative (beta=-0.06, 95% CI [-0.36, .24], t(2954)=-0.40, p=0.689).
The effect of ESCUS was statistically significant and negative (beta=-2.34, 95% CI [-2.38, -2.30], t(2954)=-113.29, p<.001).

The interaction effect of BMVDR and spatially separated speech was statistically significant and negative (beta=-7.46, 95% CI [-8.20, -6.72], t(2954)=-19.72, p<.001).
The interaction effect of the unprocessed speech and spatial separation was statistically significant and negative (beta=-4.85, 95% CI [-5.59, -4.11], t(2954)=-12.82, p<.001).

Of particular interest for the present study was the question how the algorithms performed compared to each other. Therefore, we did a marginal contrast analysis to see how the algorithms compared within the same spatial condition. We found that when all speakers were collocated that the IBM outperformed the BMVDR (-8.1 dB difference, t(2939)=-30.45, p<.001) and the unprocessed condition (-7.2 dB difference, t(2939)=-27, p<.001). Further, the unprocessed condition outperformed the BMVDR (0.9 dB difference, t(2939)=3.44, p<.01).
When target and interferers were separated we found a non-significant difference between IBM and BMVDR (-0.7 dB difference, t(2939)=-2.55, p<.110) and a significant difference between IBM and the unprocessed condition (-2.4 dB difference, t(2939)=-8.87, p<.001). Under this condition the BMVDR outperformed the unprocessed condition (-1.7 dB difference, t(2939)=-6.32, p<.001).

Discussion

We compared three different speech enhancement algorithms under two different spatial conditions in their ability to improve SRT as well as subjective listening effort. As expected, we found that spatial separation generally improved performance and that the best SRT was achieved with the IBM (which requires perfect knowledge of the unmixed target and interferer stimuli). The interaction effects showed that spatial separation improved the performance of the BMVDR as well as the unprocessed condition significantly. A comparison between the BMVDR and the unprocessed condition showed no significant differences for the SRT, although the mean SRT of BMVDR-processed stimuli were slightly lower (by 1.6 dB), which is in line with [1].

The comparison of listening effort also showed that the IBM performed best, but we found no significant improvement for spatial separation in general.
The interaction effects, however, showed a significant improvement for the BMVDR and the unprocessed condition. When we looked at the marginal means we found that in the collocated condition the IBM outperformed the BMVDR and the unprocessed condition significantly, and that the BMVDR was also outperformed by the unprocessed condition. On the other hand, we found no significant difference between the IBM and the BMVDR in the spatially separated condition and both algorithms outperformed the unprocessed condition.

Acknowledgments:

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Projektnummer 352015383 - SFB 1330 A1.

Literatur

[1] Van Veen, B. D., and Buckley, K. M.: Beamforming: A versatile approach to spatial filtering. IEEE assp magazine 5 (1988), 4-24

[2] Doclo, S., Kellermann, W., Makino, S., and Nordholm, S. E.: Multichannel signal enhancement algorithms for assisted listening devices: Exploiting spatial diversity using multiple microphones. IEEE Signal Processing Magazine 32 (2015), 18-30

[3] Brungart, D. S., Chang, P. S., Simpson, B. D., and Wang, D.: Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation. The Journal of the Acoustical Society of America 120 (2006), 4007-4018

[4] Brungart, D. S., Chang, P. S., Simpson, B. D., and Wang, D.: Multitalker speech perception with ideal time-frequency segregation: Effects of voice characteristics and number of talkers. The Journal of the Acoustical Society of America 125 (2009), 4006-4022

[5] Hauth, C. F., Gößling, N., and Brand, T.: Performance Prediction of the BinauralMVDR Beamformer with Partial Noise Estimation using a Binaural Speech IntelligibilityModel. Speech Communication, 13th ITG-Symposium, 2018, 1-5,

[6] Rennies, J., Best, V., Roverud, E., and Kidd Jr, G.: Energetic and informational components of speech-on-speech masking in binaural speech intelligibility and perceived listening effort. Trends in hearing 23 (2019), 1-21

[7] Kidd Jr, G., Mason, C. R., Swaminathan, J., Roverud, E., Clayton, K. K., and Best, V.: Determining the energetic and informational components of speech-on-speech masking. The Journal of the Acoustical Society of America 140 (2016), 132-144

[8] Gößling, N., Marquardt, D., and Doclo, S.: Perceptual evaluation of binaural MVDR-based algorithms to preserve the interaural coherence of diffuse noise fields. Trends in hearing 24 (2020), 1-18

[9] Van den Bogaert, T., Doclo, S., Wouters, J., and Moonen, M.: Speech enhancement with multichannel Wiener filter techniques in multimicrophone binaural hearing aids. The Journal of the Acoustical Society of America 125 (2009), 360-371

[10] Van den Bogaert, T., Doclo, S., Wouters, J., and Moonen, M.: The effect of multimicrophone noise reduction systems on sound source localization by users of binaural hearing aids. The Journal of the Acoustical Society of America 124 (2008), 484-497

[11] Krueger, M., Michael, S., and Inga, H.: Entwicklung einer adaptiven Skalierungsmethode zur Ermittlung der subjektiven Höranstrengung. Proceedings of German Annual Conference on Acoustics, 2015.

[12] Wagener, K., Brand, T., and Kollmeier, B.: Entwicklung und Evaluation eines Satztests in deutscher Sprache Teil III: Evaluation des Oldenburger Satztests (in German).(Development and evaluation of a German sentence test–Part III: Evaluation of the Oldenburg sentence test). Zeitschrift für Audiologie 38 (1999), 44-56

[13] Hochmuth, S., Kollmeier, B., and Shinn-Cunningham, B.: The relation between acoustic-phonetic properties and speech intelligibility in noise across languages and talkers. Proceedings of German Annual Conference on Acoustics, 2018, 628-629,

[14] Rennies, J., Best, V., Roverud, E., and Kidd Jr, G.: Energetic and informational components of speech-on-speech masking in binaural speech intelligibility and perceived listening effort. Trends in hearing 23 (2019), 2331216519854597

[15] Kayser, H., Ewert, S. D., Anemüller, J., Rohdenburg, T., Hohmann, V., and Kollmeier, B.: Database of multichannel in-ear and behind-the-ear head-related and binaural room impulse responses. EURASIP Journal on Advances in Signal Processing 2009 (2009), 1-10

[16] Ewert, S. D.: AFC—A modular framework for running psychoacoustic experiments and computational perception models. Proceedings of the international conference on acoustics AIA-DAGA, 2013, 1326-1329,

[17] Brand, T., and Hohmann, V.: An adaptive procedure for categorical loudness scaling. The Journal of the Acoustical Society of America 112 (2002), 1597-1604

[18] Team, R. C.: R: A language and environment for statistical computing. (2013),

[19] Bates, D. M.: lme4: Mixed-effects modeling with R. Springer New York (2010)

[20] Makowski, D., Lüdecke, D., and Ben-Shachar, M.: Modelbased: Estimation of model-based predictions, contrasts and means. CRAN. https://github. com/easystats/modelbased (2020)