Identifying the human-machine differences in complex binaural scenes: What can be learned from our auditory system

Constantin Spille and Bernd T. Meyer

Department of Medical Physics and Acoustics, Cluster of Excellence Hearing4All
University of Oldenburg, Germany
constantin.spille@uni-oldenburg.de, bernd.meyer@uni-oldenburg.de

10.21437/Interspeech.2014-150

Abstract

Previous comparisons of human speech recognition (HSR) and automatic speech recognition (ASR) focused on monaural signals in additive noise, and showed that HSR is far more robust against intrinsic and extrinsic sources of variation than conventional ASR. The aim of this study is to analyze the man-machine gap (and its causes) in more complex acoustic scenarios, particularly in scenes with two moving speakers, reverberation and diffuse noise. Responses of nine normal-hearing listeners are compared to errors of an ASR system that employs a binaural model for direction-of-arrival estimation and beamforming for signal enhancement. The overall man-machine gap is measured in terms for the speech recognition threshold (SRT), i.e., the signal-to-noise ratio at which a 50 % recognition rate is obtained. The comparison shows that the gap amounts to 16.7 dB SRT difference which exceeds the difference of 10 dB found in monaural situations. Based on cross comparisons that use oracle knowledge (e.g., the speakers' true position), incorrect responses are attributed to localization errors (7 dB) or missing spectral information to distinguish between speakers with different gender (3 dB). The comparison hence identifies specific ASR components that can profit from learning from binaural auditory signal processing.

Index Terms: speech synthesis, unit selection, join costs

1. Introduction

The segregation and comprehension of concurrent speakers is one of the key features in cocktail-party processing and a task normal-hearing listeners can easily perform [1]. While most of today's ASR systems work well in situations with high signal-to-noise ratios (SNR), e.g., in situations with close-talk microphones, performance rapidly decreases with increasing distance where reverberation and surrounding noises have a detrimental effect. This gap in performance between human speech recognition (HSR) and automatic speech recognition (ASR) in complex scenes motivates the use of auditory-inspired processing [2] and the incorporation of features estimated from the acoustic scene, such as prior information about the position of speakers. Previous studies dealing with a comparison of ASR and HSR focused on monaural signals. Lippmann [3] presented a review of various human and machine results for different corpora such as the Wall Street Journal and TiDigits corpora and ASR error rates were reported to be an order of magnitude higher compared to HSR. In phoneme recognition ASR error rates were smaller but still about 2.5 times higher than those of humans, as reported in [4]. Robertson et al. [5] presented a digit-in-noise test to model normal-hearing and hearing-impaired listeners speech intelligibility which was based on the Aurora-2 speech material [6]. Leonard [7] developed the TIDigits database which is also used in this study and reported human error rates of 0.1 % for clean speech resynthesized from linear prediction coefficients. Motivated by recent technological advances that put ASR in everyday situations within reach, the scope of this paper is to extend the man-machine comparison to binaural signals and complex acoustic scenes in which auditory scene analysis (ASA) plays an important role. In this study we compare results obtained by nine normal-hearing listeners to results of a previously developed ASR system ([8, 9]) in complex acoustic scenes consisting of two moving speakers, reverberation and additive diffuse noise at different target-to-masker ratios (TMR). For a fair comparison, the same signals are used for HSR and ASR experiments. Besides a quantification of the man-machine gap and thereby answering the question of how far have we come in reaching human speech recognition performance, a further analysis of the processing chain is carried out to partition the man-machine gap; the aim is to identify critical processing steps for which the performance difference of man and machine is especially high. Hence, these steps have a high potential of being improved by applying knowledge about the auditory system to machine listening. In order to identify these processing steps, the analysis focuses on factors that increase the complexity of the scene, specifically, we investigate the influence of the gender of target and masking speaker which has been shown to have a significant effect on recognition performance of human listeners [10], as well as the impact of spatial distance between target and masker on humans and machines. The paper is structured as follows. In Sec. 2 the speech data as well as the listening and ASR experiments are described. In Sec. 3 the results are presented and discussed and in the end the study is summarized and concluded in Sec. 4.

2. Methods

2.1. Speech database for ASR and HSR

The same test speech data was used both for ASR and HSR. In the following, the test data, as well as the training data used for ASR (and partly for HSR) is described. The speech data used for the experiments consists of sentences produced by 10 speakers (4 male, 6 female) that were recorded using close-talk microphones in our lab. The syntactical structure and the vocabulary were adopted from the Oldenburg Sentence Test (olsa) [11]: Each sentence contains five words with 10 possible response choices for each word category and a syntax that follows the pattern <name><verb><number> <adjective><object>, which results in a vocabulary size of 50 words. By using this fixed grammatical structure, the focus of the comparison is laid on the lexical level: Both in HSR and ASR ex-

periments, the fixed structure is known to the ASR system/the listener, so that possible differences of language models - which are out of scope for this paper, but have been analyzed in [12] - are not investigated. The original recordings with a sampling rate of 44.1 kHz were downsampled to 16 kHz and concatenated (using three sentences produced by one speaker) to obtain utterances with a duration of 5 to 10 s, suitable for speaker tracking. The head-related impulse responses (HRIR) used in this study are a subset of the database described in [13]: Reverberant HRIRs from the frontal horizontal half-plane measured at a distance of one meter between microphones and loudspeaker were selected. The HRIRs from the database were measured in a typical office room with a 5° resolution for the azimuth angles, which was interpolated to obtain a 0.5° resolution. Moving speakers where simulated by employing a frame-wise processing scheme: 64 ms Hann windows with 50 % overlap were applied and each time frame was convolved with the respective HRIR. The speakers' initial position, speed and direction of movement on a semi-circle with a radius of 1 m were randomly chosen, so that typical walking speeds were simulated and the two speakers crossed their path in 50 % of the simulations. This allows the creation of different datasets with all different speaker configurations. The ASR training set was created by converting the original speech to reverberated one-speaker signals and mixing these signals with random parts of a stationary speech-shaped noise at signal-to-noise ratios (SNR) ranging from -20 to 0 dB in 5 dB steps. Additionally, the same signals without noise were created which we refer to as clean. The training signals were processed with the beamformer also used in the ASR system which was steered with the true angles of the speaker. To obtain a reasonable amount of training data, the original speech material was processed five times with this method (resulting in 3.8 hours of training data); the randomized speaker movement and addition of noise should result in a better modeling of the variability to be expected during the testing. An ASR test set (part of which is later used for HSR experiments as well) was created in a similar manner: Two-speaker signals were generated by mixing two different moving speakers producing the same sound pressure level and random parts of a diffuse noise at target-to-masker ratios (TMR) ranging from -20 to 0 dB in 5 dB steps. This procedure was carried out ten times to obtain test data of 10650 sentences (approximately 7.5 hours duration).

2.2. HSR experiments

Nine normal-hearing subjects participated in the experiment. Their hearing thresholds did not exceed +20 dB at any data point in the audiogram, and not more than +10 dB at more than three data points. Signals were presented in a soundproof booth via audiological headphones (Sennheiser HDA200). Before the experiment, listeners were verbally instructed and were also asked to read the written instructions to prepare for the task. Additionally, each listener completed a short training phase to get familiar with the speech material, the acoustic scenes as well as the graphical user interface of the experimental software. This training procedure took about 5-10 minutes. For ASR experiments, the initial position of one speaker is supplied to the system and used to define the desired target speaker. For HSR experiments, a visual marker indicating the initial azimuth of one speaker was presented to the participants before the playback started. Participants were instructed to focus on the source initially located at that angle (ranging from -90 to $+90^\circ$). All subjects listened to the first of the ten created test sets containing ca. 40 minutes of speech data with 1065 sentences, and to enter the recognized words produced by the target speaker via the graphical user interface. To mimic the ASR setup with the known vocabulary and a fixed grammar, subjects were presented all word alternatives as a grid of 5 (word groups) $\times$ 10 (words per group) on the screen, resulting in a closed-test setup. Subjects could enter the responses for each sentence of the three-sentence presentations in arbitrary order. To avoid incorrect responses due to memory effects (which would be relevant with 15-word sentences of an average duration of 6.8 seconds), subjects could re-listen to the presentation as often as desired. The total measurement time per subject was approximately six hours, which was divided in multiple sessions that did not exceed 2 hours of measurement per session and listener. Word error rates were averaged over all signals and all subjects for each TMR.

2.3. ASR experiments

The ASR system used here is described in more detail in [9] and [8] so that we will only briefly describe the relevant features here. Figure 1 shows a block diagram of the whole processing chain. Acoustic scenes are simulated by convolving signals with recorded 6-channel head-related impulse respones (HRIR) (3 channels from each of two behind-the-ear (BTE) hearing aids). In the binaural processing step, the signals of the front microphones are fed into the binaural model developed by Dietz et al. [14] that is employed to estimate the direction of arrival of spatially distributed speakers. A particle filter [15, 16] is then used to keep track of the positions of the moving sources. Its output is used to steer a minimum variance distortionless response (MVDR) beamformer [17, 18], enhancing the 6-channel speech signal that is to be transcribed by an ASR system. The enhanced signal is then converted to high dimensional Gabor features by convolving the log-mel spectrogram with a filterbank of two-dimensional spectro-temporal Gabor filters as described in [19]. As a baseline, standard mel-frequency cepstral coefficients (MFCCs) with cepstral mean and variance normalization are used. The ASR system is a hidden Markov model (HMM) with Gaussian mixture models (GMM) based classifier implemented using HTK [20]. The ASR system was trained on data of nine speakers and testing was carried out with data of the remaining speaker. This was done for each of the speakers and word error rates (WER) were averaged over all ten speakers.

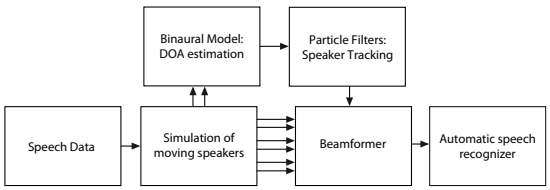

Figure 1: Block diagram of the ASR setup.

3. Results and discussion

3.1. Overall performance and speech recognition threshold

The word error rate (WER) obtained in HSR listening experiments is shown in Figure 2. Although all participants had normal hearing, a large range of individual error rates was obtained. For instance, at -20 dB target-to-masker ratio (TMR), the highest WER (64 %) is higher by a factor of 2.4 compared to the lowest WER (27 %). On average, this factor is 2.3, which is

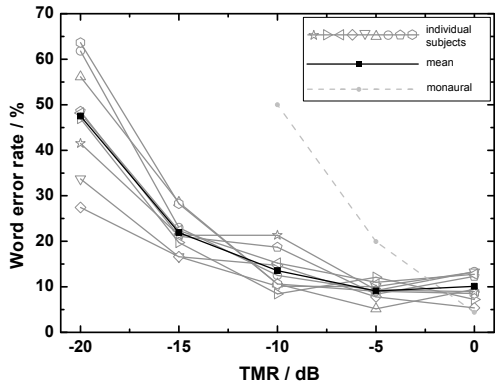

Figure 2: Word error rates for all subjects and the corresponding mean at all TMRs in grey and black lines, respectively.

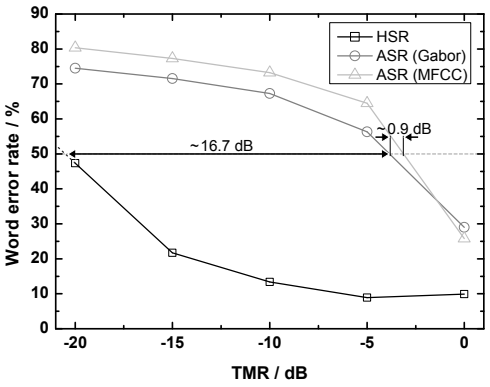

Figure 3: Word error rate vs. TMR for human speech recognition (HSR, black lines) and automatic speech recognition (ASR, grey lines). Circles refer to the ASR system with subsequently used Gabor features and triangles refer to the ASR with MFCCs.

higher than the factor of 1.9 obtained for *monaural* noisy speech in [21]. This indicates that the cognitive load is more important in binaural scenes, thereby increasing inter-individual differences. Even at a TMR of -20 dB subjects still produce less than 50 % errors (47.5 % WER). By increasing the TMR, word error rates considerably decrease to 9.1 % at -5 dB. Although the mean slightly increases at 0 dB TMR to 10.1 %, this increase is not significant (p-value: 0.18). Most of the errors at -5 and 0 dB TMR occur due to speaker confusions where the subjects attended the wrong speaker. When these cases are excluded from the analysis, word error rates are reduced to about 3 % for both TMRs. The human-machine gap in terms of the *factor* between error rates highly depends on the specific TMR (cf. [21]), whereas the difference in speech recognition thresholds (SRTs) is a more stable measure. The SRT is the SNR or TMR at which 50 % of the speech is recognized. Since the average WER in HSR does not reach 50 % a linear extrapolation based on the data points at -20 dB and -15 dB TMR and their corresponding standard errors is used. The SRT in this case was found to be -20.5 $\pm$ 4.3 dB. We now compare this human performance to the performance our ASR system achieved. For ASR features, we used MFCCs with cepstral mean and variance normalization and also physiologically motivated Gabor features, as described in Section 2. [21, 8, 9, 19]. The results are shown in Fig. 3. Fig. 3 shows that there is a large gap between ASR and human performance. This gap is relatively small at -20 dB where WERs increase by a factor of 1.6 and 1.5 for MFCCs and Gabor features, respectively, compared to HSR. While error rates rapidly decrease in HSR, there is only a moderate decrease in ASR, leading to a relative increase of error rates at -5 dB by a factor of 7.7 (MFCCs) and 6.7 (Gabor features). The SRT of the ASR system with Gabor features (-3.8 $\pm$ 0.7 dB) is 16.7 dB higher than that of humans. By using MFCCs the SRT again increase by 0.7 dB to -3.1 $\pm$ 0.5 dB.

3.2. Analysis of the performance gap

In the following, a further analysis of the 16.7 dB gap is presented with the aim of identifying the error sources that cause the differences between man and machine. One obvious candidate for this is the localization of speakers. Previous studies have shown that recognition performance considerably decrease with degrading localization performance [8, 9]. If the binaural model is not sufficient to accurately localize the speaker, the beamformer cannot enhance the signal and hence, WERs increase. To quantify this effect, the true speaker positions (instead of estimated angles as required for a working application) were used for beamforming. The removal of localization errors had a major effect on overall performance, with the SRT gap being reduced by 7.3 dB (from 16.7 to 9.4 dB) (see Fig. 4). The remaining 9.4 dB gap is consistent with the 10 dB reported

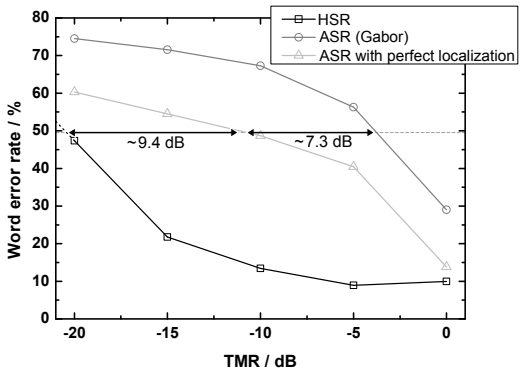

Figure 4: WER vs. TMR. Squares refer to human performance, circles to ASR performance with Gabor features and triangles to ASR performance with Gabor features and perfect knowledge of the speakers position.

as the man-machine difference for monaural noisy digit strings [21]. To improve localization performance in practice, parts of the binaural model could be changed. First, the frequency range which is used for localization could be extended. So far only frequencies up to 1.4 khz are used [14], which was motivated by physiological and psychoacoustic measurements, but which also means that a large frequency range is neglected in terms of localization. Further, a more sophisticated way of selecting proper DOA estimations could be used to improve localization performance, e.g. a statistical analysis of all estimations with a maximum-likelihood estimator of the "'true"' DOA [22]. An explicit incorporation of the precedence effect which helps humans to localize sounds in reverberant conditions [23] might also help to improve localization. For multi-speaker scenes such as the ones investigated, human listeners perform significantly better when voice characteristics of the target and the concurrent

speaker are different [10]. In particular, it was shown that word error rates for different-gender speakers was substantially lower than for same-gender speakers. This is also reflected by the data collected for this study: The same-gender error rates are consistently higher than the different-gender WERs (by 8.4 % on average). However, in case of ASR, a clear trend does not emerge, and the differences between both categories are very small in comparison to HSR. Darwin et al. [24] showed that HSR in two-speaker scenes was significantly improved if the difference in fundamental frequency (F0) was greater than 2 semitones and if the ratio of vocal-tract lengths was greater than 1.08. When altering both F0 and vocal-tract length, which simulates a shift in gender, substantially larger improvements than differences in one of the two (F0 and vocal-tract length) alone were produced. These improvements were found to be similar to the improvements obtained by different-gender speakers. Differences in fundamental frequency and vocal-tract length are not exploited in the ASR system here. One possible approach to make use of these differences is to apply a pitch estimation algorithm (see, e.g., [25]) to spectrally separate the two speakers, or to use vocal-tract parameters to generate speaker clusters and use speaker-clustered HMMs to perform speech recognition [26]. One additional factor that influences the complexity of binaural scenes is the spatial distance of target and masker. Since we are interested in how complexity influences HSR and ASR, an analysis of situations with speakers being temporarily at the same position (i.e., the constantly moving speakers cross their ways) was performed. For HSR, the WER for non-crossing speakers is 5.3 % lower than for crossing speakers, which was expected since a non-crossing implies a spatial distance that should increase source separability. In ASR, this difference is even larger. With perfect localization, WERs are 13.7 % higher for crossing speakers than for non-crossing speakers. This can be partly attributed to distortions that are generated when the main lobe and the attenuation of the beamformer fall on the exact same angle. At these time instances the signal is completely attenuated and thus, a reliable recognition of speech cannot be performed. This is a general problem in ASR systems with beamforming approaches using a main lobe and attenuation such as our system; in contrast to this, human listeners perform a selective suppression of the undesired source, which operates on the complete signal. Two approaches to mimic this behaviour are to either selectively change the beamformer processing when the target-masker distance falls below a specific threshold or to use other source separation methods such as non-negative matrix factorization which already was successfully applied to speech recognition [27]. To bridge the man-machine gap, ASR algorithms should be applied that take the same cues into account that are employed by the auditory system. However, since these algorithms are currently unavailable in our ASR system, we analyzed the man-machine gap in situations in which the mentioned cues are unavailable to the human listener, i.e., scenes with crossing speakers with the same gender. In these situations, the SRT is raised from the original -20.5 dB by 3.2 dB to -17.3 dB. The gender and position cues hence explain for a SRT shift of 3.2 dB. Assuming that ASR algorithms are developed that optimally exploit these cues, the gap of 9.4 dB (which was already based on optimal localization) could potentially be further reduced to 6.2 dB. There are numerous reasons for the remaining 6 dB gap, some of which are briefly discussed here: Even for a small-vocabulary task as the one presented here, the exposure to training material is essential. We assume that an extension of the training data (in combination with different classification algorithms, see below) is beneficial; future experiments will therefore take into account the amount of training samples required for saturated recognition scores, and its comparison to HSR. ASR oracle experiments with the true positions of source and masker showed that the beamformer is suboptimal in comparison to stream segregation of the auditory system: As discussed earlier the attenuation of the beamformer can attenuate the signal completely when steered to the same position as the main lobe, which could be fixed by putting additional constraints on beamforming. Other source separation schemes like non-negative matrix factorization (NMF) or binary-mask based approaches do not have this problem [28, 27] but are fragile in situations with moving sources, since the spectral change induced by moving sources requires the estimation of binary masks from very short observations. Regarding the backend, one obvious difference between HSR and ASR is the classifier, i.e., typically applied HMM/GMMs are quite different compared to the higher auditory pathway and the cortical stages of hearing both in terms of structure and functionality. Recent progress in machine learning was based on deep neural networks (see ,e.g., [29]), which still exhibit much lower complexity than its biological example, but are structurally more similar to auditory processing which could contribute to a further reduction between HSR and ASR.

4. Summary and conclusion

This study compared human speech recognition with automatic speech recognition in complex acoustic scenes, with the main goals of a) quantifying the man-machine gap to answer the question of how far have we come in reaching human speech recognition performance and b) to identify the critical processing steps for which the performance difference of man and machine is especially high. The focus was laid on the lexical level by using speech material with a known grammatical structure, which was used both for HSR by performing listening experiments with nine normal-hearing listeners and ASR. The main measure for quantification is the difference of the speech recognition threshold (SRT). In total, the SRT gap between HSR and ASR was found to be 16.7 dB. When the true speakers position is supplied to the ASR system, this gap is reduced by 7.3 dB to 9.1 dB which highlights the importance of robust estimation of direction of arrival. A further important difference between HSR and ASR was found when analyzing scenes with two speakers that are either of equal or different gender: When speech was presented in a situation with two speakers of different gender, the SRT in HSR was decreased by 3 dB, while this effect was not observed in ASR. We assume that spectral differences between the target and the interfering speaker are exploited by humans to separate both speakers but are neglected in ASR. If these differences could be optimally exploited by ASR, this could further reduce the observed gap by 3 dB. Candidates to perform an efficient separation between speakers are features that make use of spectral fine-structure, whereas the spectral smoothing that is applied in most current ASR systems aggravates this kind of speaker separation.

5. Acknowledgment

Supported by the DFG (SFB/TRR 31 'The active auditory system'; URL: http://www.sfb-trr31.uni-oldenburg.de/). The authors would like to thank Mathias Dietz, Volker Hohmann and Daniel Marquardt for valuable contributions to this work and Marc René Schädler for sharing the code of the recognition system.

6. References

[1] A. S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound*. MIT Press, 1990.

[2] O. Scharenborg, "Reaching over the gap: A review of efforts to link human and automatic speech recognition research," *Speech Commun.*, vol. 49, no. 5, pp. 336–347, May 2007,

[3] R. Lippmann, "Speech recognition by machines and humans," *Speech Commun.*, vol. 22, no. 1, pp. 1–15, 1997,

[4] B. T. Meyer and B. Kollmeier, "Robustness of spectro-temporal features against intrinsic and extrinsic variations in automatic speech recognition," *Speech Commun.*, vol. 53, no. 5, pp. 753–767, May 2011,

[5] M. Robertson, G. J. Brown, W. Lecluyse, M. Panda, and C. M. Tan, "A speech-in-noise test based on spoken digits: comparison of normal and impaired listeners using a computer model." *Proc. Interspeech*, pp. 2470–2473, 2010,

[6] H. Hirsch and D. Pearce, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," *ASR-2000*, vol. 2000, no. October, pp. 16–19, 2000,

[7] R. Leonard, "A database for speaker-independent digit recognition," in *Proc. ICASSP 1984*, 1984,

[8] C. Spille, B. T. Meyer, M. Dietz, and V. Hohmann, "Binaural scene analysis with multi-dimensional statistical filters," in *Technol. binaural List.*, J. Blauert, Ed. Springer, Berlin-Heidelberg-New York NY, 2013, ch. 6,

[9] C. Spille, M. Dietz, V. Hohmann, and B. T. Meyer, "Using binarual processing for automatic speech recognition in multi-talker scenes," in *Proc. ICASSP 2013*, 2013, pp. 7805–7809,

[10] D. S. Brungart, "Informational and energetic masking effects in the perception of two simultaneous talkers." *J. Acoust. Soc. Am.*, vol. 109, no. 3, pp. 1101–9, Mar. 2001,

[11] K. C. Wagener and T. Brand, "Sentence intelligibility in noise for listeners with normal hearing and hearing impairment: influence of measurement procedure and masking parameters." *Int. J. Audiol.*, vol. 44, no. 3, pp. 144–156, 2005,

[12] W. Shen, J. Olive, and D. Jones, "Two protocols comparing human and machine phonetic recognition performance in conversational speech." in *Proc. Interspeech*, 2008, pp. 1630–1633,

[13] H. Kayser, S. D. Ewert, J. Anemüller, T. Rohdenburg, V. Hohmann, and B. Kollmeier, "Database of Multichannel In-Ear and Behind-the-Ear Head-Related and Binaural Room Impulse Responses," *EURASIP J. Adv. Signal Process.*, no. 1, p. 298605, 2009,

[14] M. Dietz, S. D. Ewert, and V. Hohmann, "Auditory model based direction estimation of concurrent speakers from binaural signals," *Speech Commun.*, vol. 53, no. 5, pp. 592–605, May 2011,

[15] S. Särkkä, A. Vehtari, and J. Lampinen, "Rao-Blackwellized particle filter for multiple target tracking," *Inf. Fusion*, vol. 8, no. 1, pp. 2–15, Jan. 2007,

[16] J. Hartikainen and S. Särkkä, "RBMCDAbox-Matlab Toolbox of Rao-Blackwellized Data Association Particle Filters," Department of Biomedical Engineering and Computational Science, Helsinki University of Technology, Tech. Rep., 2008,

[17] H. Cox, R. Zeskind, and M. Owen, "Robust adaptive beamforming," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 35, no. 10, pp. 1365–1376, 1987,

[18] J. Bitzer and K. U. Simmer, "Superdirective Microphone Arrays," in *Microphone Arrays*, M. Brandstein and D. Ward, Eds. Springer, 2001, pp. 1021–1042,

[19] M. R. Schädler, B. T. Meyer, and B. Kollmeier, "Spectro-temporal modulation subspace-spanning filter bank features for robust automatic speech recognition." *J. Acoust. Soc. Am.*, vol. 131, no. 5, pp. 4134–4151, May 2012,

[20] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, "The HTK book," *Cambridge Univ. Eng. Dep.*, vol. 3, 2002,

[21] B. T. Meyer, "What s the difference ? Comparing humans and machines on the Aurora 2 speech recognition task," in *Proc. Interspeech 2013*, 2013, pp. 2634–2638,

[22] H. Kayser, S. D. Ewert, V. Hohmann, and J. Anemüller, "Probabilistic modeling of auditory cues for binaural speaker localization," in *Proc. AIA-DAGA 2013*, 2013,

[23] H. Wallach, E. B. Newman, and M. R. Rosenzweig, "A Precedence Effect in Sound Localization," *J. ...*, vol. 21, no. 4, pp. 468–468, 1949,

[24] C. J. Darwin, D. S. Brungart, and B. D. Simpson, "Effects of fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers," *J. Acoust. Soc. Am.*, vol. 114, no. 5, p. 2913, 2003,

[25] S. D. Ewert, C. Iben, and V. Hohmann, "Robust fundamental frequency estimation in an auditory model," in *Proc. Int. Conf. Acoust. AIA-DAGA*, 2013, pp. 271–274,

[26] M. Naito, L. Deng, and Y. Sagisaka, "Speaker clustering for speech recognition using vocal tract parameters," *Speech Commun.*, vol. 36, no. 3-4, pp. 305–315, Mar. 2002,

[27] N. Moritz and M. Schädler, "Noise robust distant automatic speech recognition utilizing NMF based source separation and auditory feature extraction," *Proc. CHiME-2013, ...*, pp. 1–6, 2013,

[28] N. Roman, S. Srinivasan, and D. Wang, "Binaural segregation in multisource reverberant environments," *J. Acoust. Soc. Am.*, vol. 120, no. 6, p. 4040, 2006,

[29] A.-r. Mohamed, G. E. Dahl, and G. Hinton, "Acoustic Modeling Using Deep Belief Networks," *IEEE Trans. Audio. Speech. Lang. Processing*, vol. 20, no. 1, pp. 14–22, Jan. 2012,