INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Assessing speech quality in speech-aware hearing aids based on phoneme posteriorgrams

Constantin Spille[1], *Hendrik Kayser*[1], *Hynek Hermansky*[2], *Bernd T. Meyer*[2]

[1]Medizinische Physik and Cluster of Excellence Hearing4all, Carl von Ossietzky Universität, Oldenburg, Germany
[2]Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD, USA
{constantin.spille, hendrik.kayser}@uni-oldenburg.de, {hynek, bernd.t.meyer}@jhu.edu

Abstract

Current behind-the-ear hearing aids (HA) allow to perform spatial filtering to enhance localized sound sources; however, they often lack processing strategies that are tailored to spoken language. Hence, without a feedback about speech quality achieved by the system, spatial filtering potentially remains unused, in case of a conservative enhancement strategy, or can even be detrimental to the speech intelligibility of the output signal. In this paper we apply phoneme posteriorgrams obtained from HA signals processed with deep neural networks to measure the quality of speech representations in spatial scenes. Inverse entropy of phoneme probabilities is proposed as a measure that allows to evaluate if current hearing aid parameters are optimal for the given acoustic condition. We investigate how varying noise levels and wrong estimates of the to-be-enhanced direction affect this measure in anechoic and reverberant conditions and show our measure to provide a high reliability when varying each parameter.Experiments show that entropy as a function of the beam angle has a distinct minimum at the speaker's true position and its immediate vicinity. Thus, it can be used to determine the beam angle which optimizes the speech representation. Further, variations of the SNR cause a consistent offset of the entropy.

Index Terms: speech recognition, hearing aids, beamforming

1. Introduction

Users of hearing aids can substantially profit from adaptive spatial speech signal enhancement [1], which requires estimates such as the sound source direction of arrival (DOA) and noise statistics. Incorrect estimates introduce artifacts in the output signal that decrease speech intelligibility, such that a simpler method less prone to estimation errors should be chosen, e.g., straight-forward better-ear-listening [2]. An important limitation in hearing aid processing arises from the uncertainty of which method to choose and potentially leads to use of conservative processing strategies. Previous studies have shown that normal-hearing listeners can typically identify whether the message in the signal is recognizable [3]. Hearing aids should be able to do the same in order to chose the best processing strategy. Typically, low-level information at sensor level (i.e. hearing aid microphones) about the statistics of the audio signals are used to determine the acoustic scene and select a predefined parameter set in a hearing aid [4]. However, high-level features such as information about spoken language and speech intelligibility are rarely used in hearing aids. Towards this end, we investigate techniques that have been applied in machine listening for estimating the quality of speech representations in enhanced signals suitable for continuous optimization of hearing aid parameters.

Deep learning has resulted in major breakthroughs in speech processing in recent years [5], but has not unfolded its potential in hearing research. We propose to employ phoneme probability posteriorgrams as a representation of speech in order to optimize speech enhancement in hearing aids. An entropy criterion is applied to the posteriorgrams to quantify the speech quality in a given time frame. Entropy-based criteria already have been applied to measure the performance of individual artificial neural networks (ANN) and to estimate weighting factors for ANNs in multi-band and multi-stream ASR [6, 7]. In these approaches generally a set of independent classifiers is trained on different representations of the data and classifiers are subsequently merged to get optimal results. In [6] e.g. different ANNs are trained on phoneme posterior probabilities with different feature combinations and the entropy is used to measure how corrupted a given ANN output is. Then, an inverse entropy weighting is applied to the ANNs and the probabilities are summed. In [7] different systems are trained for different frequency subbands of the whole frequency range and mutual information (which is based on entropy) is used as a weighting factor of the individual subband systems.

Automatic speech recognition (ASR) with deep neural networks (DNN) in complex spatial acoustic scenes where azimuth angles of speakers were used to steer a beamforming algorithm to enhance target speech was applied e.g. in [8, 9]. In these studies DOA estimates were used to steer a beamformer to the direction of the target speaker which shows to improve ASR accuracies in most scenes. However, in some cases it turned out to be beneficial to use simpler signal enhancement strategies such as better-ear listening [2]. These approaches are for the first time combined in this work to exploit performance monitoring with the aim of increasing hearing aid settings and ultimately speech intelligibility for hearing-impaired listeners.

The remainder of this paper is structured as follows: In Section 2 the experimental setup including the acoustic scenes and spatial filtering as well as the speech recognition framework and the entropy measure are introduced. After presenting the results of the experiments in Section 3 the discussion and conclusions follow in Section 4.

2. Methods

2.1. Acoustic scenes and beamforming method

Spatially localized and diffuse sound sources are simulated using a database of head-related impulse responses (HRIR), which

http://dx.doi.org/10.21437/Interspeech.2016-1318

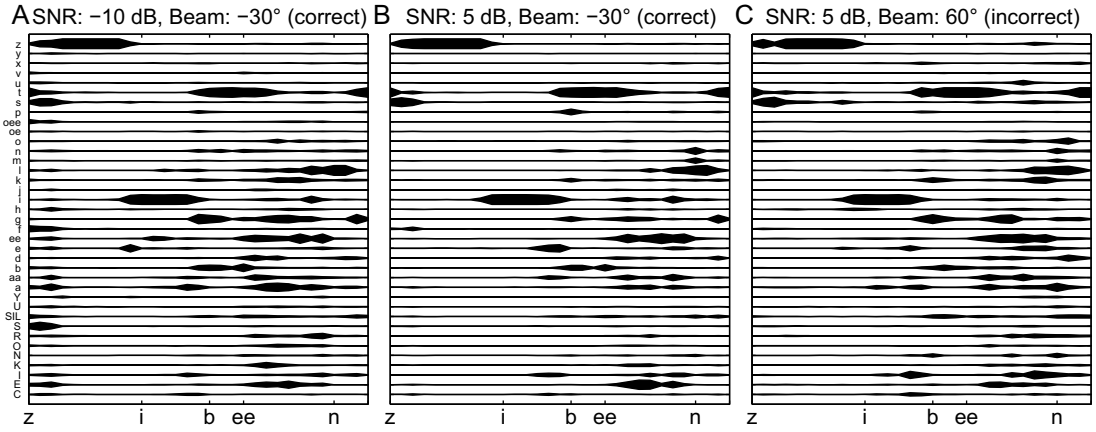

Figure 1: *DNN-based phoneme posteriorgrams of the German word "sieben" (seven) obtained from hearing aid signals and subsequent beamforming. A: -10 dB SNR, correct beam angle (−30°); B: 5 dB SNR and correct beam angle (−30°); C: 5 dB SNR and incorrect beam angle (60°).*

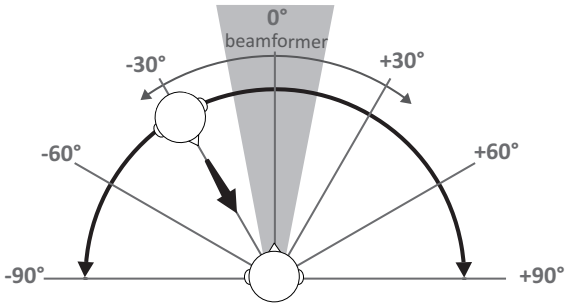

Figure 2: *Overview of the experimental setup: a target speaker is fixed at an azimuth angle of -30° from the listener (center). A beamformer (indicated by a grey shadow) operating on six behind-the-ear hearing aid microphones is steered over the whole azimuth range.*

features impulse responses recorded with three microphones from each of two behind-the-ear (BTE) hearing aids attached to left and the right ear. The HRIRs used in this study are a subset of the database described in [10]: Anechoic free-field HRIRs and reverberated HRIRs from the frontal horizontal half-plane were measured at a distance of 3 m and 1 m between microphones and loudspeaker, respectively. All HRIRs (anechoic and reverberated) from the database were measured with 5° resolution for the azimuth angles, which was interpolated to obtain a resolution of 0.5°. Reverberated HRIRs were measured in a typical office environment with a reverberation time of ∼300 ms. Figure 2 shows a sketch of the acoustic scenes under consideration. One target speaker at a fixed azimuth angle of -30°, i.e. 30° to the left, was mixed with random parts of an additional spatially diffuse stationary speech-shaped noise at signal-to-noise ratios (SNR) from -10 to 5 dB in 5 dB steps.

The beamformer employed is a super-directive beamformer based on the minimum variance distortionless response (MVDR) principle [11] that uses the six BTE microphone inputs jointly. Let $\boldsymbol{W}$ be the matrix containing the frequency domain filter coefficients of the beamformer, $\boldsymbol{d}$ the vector containing the transfer functions to the microphones of the target speaker and $\boldsymbol{\Phi}_{VV}$ the noise power-spectral density (PSD) matrix. Then, the following minimization problem has to be solved

$$\begin{aligned} \min_{\boldsymbol{W}} &: \boldsymbol{W}^H \boldsymbol{\Phi}_{VV} \boldsymbol{W} \\ \text{with } &\boldsymbol{W}^H \boldsymbol{d} = 1. \end{aligned} \tag{1}$$

The solution is the MVDR beamformer [12]. In our approach, $\boldsymbol{d}$ contains an anechoic transfer function according to the steering direction of the beamformer. The noise coherence matrix which is required to solve Eq. (1) is estimated using the same impulse responses. By these means, solely head-related characteristics of sound propagation are included in the signal enhancement setup, but no further information about room acoustics.

2.2. Speech recognition framework and phoneme posteriorgrams

The speech recognition system is a deep neural network (DNN) with five hidden layers, 2048 units per layer and an additional softmax output layer. The DNN was trained as a stack of restricted Boltzmann machines with an unsupervised pre-training and a supervised fine-tuning of the parameters with triphone targets. Every phone was modeled with three Hidden-Markov-Model (HMM) states except for the silence phone which was modeled with five states. After training the DNN, discriminant sequence training with Minimum Bayes-Risk was performed.

Input features are calculated by converting the time signals to Mel-Frequency Cepstral Coefficients (MFCCs) [13] with additional cepstral mean and variance normalization, resulting in a 13-dimensional feature vector per 10 ms step. These features were spliced with a temporal context of ±5 frames, resulting in 143 features per frame.

Phoneme posteriorgrams were derived from the activations of the softmax output layer. All triphones belonging to the same phone were grouped and activations were summed resulting in monophone activations, which can be interpreted based on visual inspection (cf. Fig. 1), which is in contrast to high-dimensional triphone activations. Note that for a regularly-trained DNN speech recognition system with high-SNR input speech, the activations should be high for the current phone and activations for all other phones should be close to zero. For low SNRs, the uncertainty of the system increases and activations

of other phones will also increase, resulting in noisy posteriorgrams as observable by comparing the first panel in Fig. 1 to the second one. The same is true if the beamformer is steered to the correct and incorrect angle, respectively. It is assumed that in the acoustic conditions considered here, speech quality is optimal if the beamformer is steered to the correct angle.

2.3. Speech data

The speech data used in these experiments was taken from a German matrix sentence test, the Oldenburg sentence test (OLSA) [14]. The speech material has a fixed syntactical structure: 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. For training the ASR system, a speech corpus of 20 hours of speech from 20 different speakers (10 male, 10 female) was used keeping the syntactical structure of the OLSA [15]. For multi-condition ASR training, clean and noisy files were used, the latter being obtained by mixing signals with a stationary speech-shaped noise at various SNRs ranging from -10 to 20 dB in 5 dB steps. The test set consists of 8 speakers (4 m, 4 f) each uttering 100 sentences which results in a total duration of about 24 minutes of speech.

2.4. Phoneme entropy as speech quality measure

To quantify how well speech is represented in a time frame j, the entropy of the posteriorgram of this frame is calculated as follows:

$$H = -\sum_{i=1}^{N} a_{ij} \log_2(a_{ij}) \,, \qquad (2)$$

where a_{ij} is the activation of the i-th phone in time frame j and N is the total number of phones. In the case of one active phoneme and all other activations being close to zero, the entropy is minimal, whereas the entropy is maximal when all activations are equally distributed. Hence, the entropy should be low if speech is well-represented in a given time frame, whereas the entropy should increase with an increasing amount of signal noise. With a total number of 37 phones the maximum entropy in one time frame amounts to $H = 5.209$.

3. Results

Fig. 1 shows exemplary posteriorgrams for high and low SNRs as well as correct (A, B) and incorrect (C) steering angles of the beamformer. An increase of the noise level and an incorrect steering direction both lead to a more noisy posteriorgram which is also reflected in an increase of the mean frame entropy of the respective posteriorgrams. At -10 dB SNR and a correct beam angle, the mean entropy of the posteriorgram in Fig. 1 A is 0.917. Increasing the SNR to 5 dB results in a decrease of the entropy to 0.612 (cf. Fig. 1 B). If the beam is steered to 60° (Fig.1 C) the entropy increases to 0.804.

The entropy as a function of SNR and beam angle was analyzed in more detail in spatial acoustic scenes described in Section 2.1. The MVDR beamformer was steered from -90 to 90° in 5° steps and additionally to -28° and -32° to get an accurate analysis in the vicinity of the target beam angle. For each beam angle the whole test set was processed with the beamformer and the entropy was averaged over all frames in the test set.

Figure 3 shows the entropy when varying the beam angle and the SNR in the anechoic condition as well as in a typical

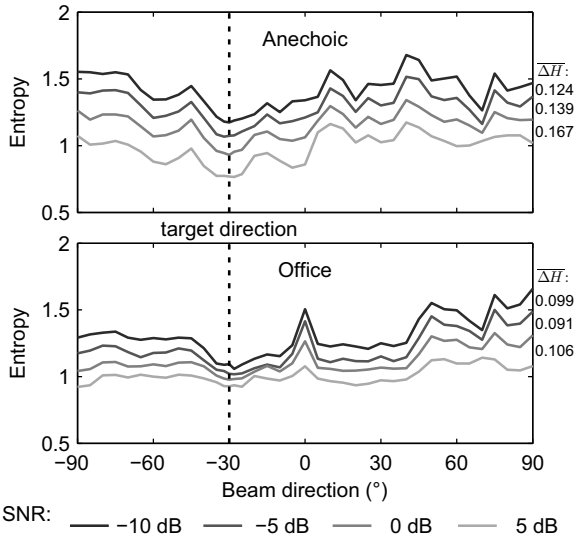

Figure 3: *Average entropy depending on the level of diffuse noise and the steering angle of the beamformer (with the speech target at $-30°$) in an anechoic environment (upper panel) and a typical office environment (lower panel). Mean offsets $\overline{\Delta H}$ between curves are given at the right side next to the figure.*

office environment. Several local minima and maxima over the whole range of beam angles are observed (cf. Fig 3) which presumably arises due to characteristics of the room, the dummy head and also the frequency-dependent beampattern that are captured in the audio signal. In the anechoic condition, starting with a beam angle of -90°, the entropy stays constant for beam angles of up to -70°, then decreases to a local minimum at around -55° to -60°. At -45° the entropy reaches a local maximum before decreasing again to the global minimum at -30° when the beam is steered to the speaker's position. At beam angles above -30° the function has a similar shape with maxima at -15° and 10° and local minima between -5° and -10° and at 20°. A change in SNR mainly results in a constant offset of the entropy but the shape of the function remains almost unchanged.

In the reverberant office condition there is a similar trend but the curves are more flattened with less prominent local minima compared to the curves in anechoic conditions. There are prominent maxima at 0°, 50° and 75°. However, although the speech recognition system is trained on anechoic signals and the beamformer uses a coherence matrix based on anechoic HRIRs, the global minimum of the entropy is also around -30°, except for 5 dB SNR where lowest entropy is at -90° (0.922 compared to 0.924 and 0.927 at -25 and -30°, respectively). Due to sound reflections from the walls of the room, speech is arriving at the microphones from more than just the speaker's position. This might cause the entropy to flatten because there is, compared to the anechoic case, always a higher amount of speech in the output of the beamformer independently of the steering direction.

A shift in SNR seems to cause an almost constant offset ΔH of the entropy curve, which is analyzed in more detail in the following. We computed the mean entropy offset $\overline{\Delta H}$ which is the difference of two entropy curves averaged over all beam angles and the corresponding standard deviation $\sigma\Delta H$. Table 1 shows $\overline{\Delta H}$ and $\sigma\Delta H$ between adjacent entropy curves shown in Fig. 3 in both room conditions (anechoic and office

environment). The small standard deviations indicate that the offset is fairly constant over the whole range of different beam angles. Given the entropy curve at a certain SNR, these offsets can be used to predict the entropy curve at another SNR. This results in quite accurate predictions with root mean squared errors of 0.023 between -10 and -5 dB up to 0.074 between -10 and 5 dB for anechoic conditions and 0.027 (-10 to -5 dB) up to 0.107 (-10 to 5 dB) for the office condition.

Table 1: *Mean offsets $\overline{\Delta H}$ between adjacent entropy curves with corresponding standard deviation $\sigma\Delta H$.*

	anechoic		office	
SNR	$\overline{\Delta H}$	$\sigma\Delta H$	$\overline{\Delta H}$	$\sigma\Delta H$
-10 dB	0.124	0.023	0.099	0.028
-5 dB	0.139	0.032	0.091	0.045
0 dB	0.167	0.038	0.106	0.047
5 dB				

4. Discussion and conclusions

In this paper DNN-based phoneme posteriorgrams are investigated to assess the quality of speech representation in a given time frame. The entropy of phoneme posteriorgrams' is used as a measure of speech quality. The phoneme posteriorgrams entropy of speech samples in anechoic and reverberant spatial scenes with diffuse noise fields was analyzed as a function of signal-to-noise ratio (SNR) and steering angle of an MVDR beamformer.

Experiments show that the entropy captures speech quality (as reflected in posteriorgrams) and is sensible to a mismatch between the true position of a target speaker and the steering direction of the beamformer as well as to a change in SNR. Under anechoic conditions the entropy reaches a global minimum independent of the SNR if the beamformer is steered to the true target source position. By this means entropy could be used to find the optimal steering angle - at least on the limited data set used in this study and matched conditions which are given by the availability of the identical impulse responses used for signal generation and signal enhancement. Furthermore the ASR system was trained in anechoic scenes with the same stationary speech shaped noise. Nevertheless, experiments conducted in reverberant conditions, but with exactly the same anechoic beamformer setup, i.e, anechoic steering vectors and noise covariance matrix, and the same ASR model show promising results. Despite this strong mismatch entropy shows a qualitatively similar behavior, particularly for more challenging SNR conditions. This indicates that the proposed system comprises some generalization capabilities and robustness against mismatch in environmental conditions which is very likely to also occur in real-world scenarios. Of course, different noise types and different configurations of spatial scenes have to be analyzed to further support these findings.

Furthermore, the entropy shows a characteristic curvature that is independent of the SNR and exhibits only a constant offset between different SNR conditions. This constant offset allows inference of SNR at input level (hearing aid microphones) from high-level information (entropy of posteriorgrams), e.g. by steering a beamformer to different angles and calculate the mean entropy over these angles. With this information one could potentially identify the working point at which a certain signal enhancement algorithm is operating, which in turn can be used to improve the enhancement capability of this algorithm. E.g., a more sophisticated method which can exploit explicit knowledge of the input SNR (or an estimate thereof) can be utilized an its success can be monitored by comparison to the previously used, more conservative processing strategy. Further research on extended data sets will be conducted to verify this hypothesis.

The results of this work show that the entropy can be used to find the optimal steering direction of a beamformer. With this measure the hearing aid is capable of comparing different signal enhancement strategies and decide which one is optimal given the microphone input signals. If there is e.g. no distinct minimum in the entropy when steering the beamformer over the whole azimuth range, then it is likely that there is no specific direction to be enhanced and thus, spatial filtering should be turned off. If multiple sources are active at the same time, entropy could be used to distinguish speech from non-speech sources. If there are two speech sources active, however, entropy should be low for both sources. In this case speaker adaptation techniques such as the use of i-vectors could help to adapt the speech recognition system to a specific speaker [16] which in turn presumably could improve posteriorgrams and lower the entropy for this speaker.

The main findings of the present study are:

- DNN-based phoneme posteriorgrams as representations of speech can be utilized to measure speech quality which is captured and quantified by the entropy of the posteriorgrams.
- Phoneme posteriorgram entropy is dependent on the steering direction of the beamformer. This can be used to find the optimal steering direction that maximizes speech quality.
- Variation of the SNR causes a constant offset of the entropy which potentially can be used to infer the SNR of a given input signal at sensor-level.

In future work the promising results of this paper have to be confirmed in different acoustic scenes with multiple sound sources and with a larger variety of noise types, e.g., by including modulated noise in the analysis. When combined with DOA estimation (e.g. [17, 18]) that is not speech-specific, but instead produces estimates for any localized sound source in the acoustic scene, the entropy estimate could be used to differentiate between speech and non-speech sources, which might result in a benefit compared to methods that are not directly tailored to speech. In order to investigate more natural scenarios for hearing-impaired listeners, including additional data from accelerometers can be employed to compensate beam angles during head movements. Also, the effect of different DNN input features on posteriograms and their entropy has to be investigated. Auditory features such as perceptual linear predictive (PLP) features [19] and spectro-temporal Gabor features [20] that are potentially more speech-specific than MFCCs might also lead to lower entropy values when speech is active. This could help to better identify the optimal hearing aid parameters.

5. Acknowledgements

This work was funded by the DFG (Research Unit FOR 1732 "Individualized Hearing Acoustics", Cluster of Excellence 1077/1‘ "Hearing4all" and the SFB/TRR 31 "The Active Auditory System") and by Google via a Google faculty award to Hynek Hermansky.

6. References

[1] Adiloglu, K., Kayser, H., Baumgärtel, R. M., Rennebeck, S., Dietz, M., Hohmann, V., (2015). "A binaural steering beamforming system for enhancing a moving speech source," Trends in Hearing, Volume 19.

[2] Kayser, H., Spille, C., Marquardt, D., Meyer, B.T. (2015a). "Improving automatic speech recognition in spatially-aware hearing aids," in Proc. Interspeech, pp. 175-179.

[3] Hermansky, H., Burget, L., Cohen, J. Dupoux, E., Feldman, N., Godfrey, J., Khudanpur, S. (2015). "Towards machines that know when they do not know," Proc. ICASSP, pp. 5009-5013.

[4] J. M. Kates (1995). "Classification of background noises for hearing-aid applications," The Journal of the Acoustical Society of America, 97(1), 461–470.

[5] Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T., Kingsbury, B. (2012). "Deep Neural Networks for Acoustic Modeling in Speech Recognition," IEEE Signal Process. Mag.

[6] Misra, H., Bourlard, H., and Tyagi, V. (2003). "New entropy based combination rules in HMM/ANN multistream ASR," in Proc. ICASSP, pp. II-741).

[7] Okawa, S., Nakajima, T., Shirai, K. (1999). "A recombination strategy for multi-band speech recognition based on mutual information criterion," Eur. Conf. Speech Commun. Technol., 1999, 2, 603-606.

[8] Spille, C., Meyer, B.T., Dietz, M., Hohmann, V. (2013b). Chapter "Binaural scene analysis with multi-dimensional statistical filters," in "The Technology of Binaural Listening" (Ed. Blauert, J.), Springer, Berlin.

[9] Spille, C., Dietz, M., Hohmann, V., Meyer, B. (2013a). "Using binaural processing for automatic speech recognition in multi-talker scenes," proc. ICASSP, pp. 7805-7809.

[10] Kayser, H., Ewert, S. D., Anemüller, J., Rohdenburg, T., Hohmann, V., Kollmeier, B. (2009). "Database of Multichannel In-Ear and Behind-the-Ear Head-Related and Binaural Room Impulse Responses," EURASIP Journal on Advances in Signal Processing, 2009.

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

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

[13] Davis, S. B., Mermelstein, P. (1980). "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," IEEE Trans. Acoust., Speech, Signal Process., 28, 357–366.

[14] Wagener, K., Brand, T., Kollmeier, B. (1999). "Development and evaluation of a German sentence test Part III: Evaluation of the Oldenburg sentence test," Z Audiol, 38(3), 5–15.

[15] Meyer, B. T., Kollmeier, B., Ooster, J. (2015). "Autonomous Measurement of Speech Intelligibility Utilizing Automatic Speech Recognition," Proc. Interspeech, pp. 2982–2986.

[16] Saon, G., Soltau, H., Nahamoo, D., Picheny, M., "Speaker adaptation of neural network acoustic models using i-vectors," Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on, Olomouc, 2013, pp. 55-59.

[17] Kayser, H., Hohmann, V., Ewert, S.D., Kollmeier, B., Anemüller, J. (2015b). "Robust auditory localization using probabilistic inference and coherence-based weighting of interaural cues," J. Acoust. Soc. Am. 138 (5), pp. 2635-2648.

[18] Kayser, H., Anemüller, J. (2014). "A discriminative learning approach to probabilistic acoustic source localization," in Proc IWAENC, pp. 99–103.

[19] Hermansky, H. (1990). "Perceptual linear predictive (PLP) analysis of speech," Journal of the Acoustical Society of America, 87(4), 1738–1752.

[20] Schädler, M.R., Meyer, B.T., Kollmeier, B. (2012). "Spectro-temporal modulation subspace-spanning filter bank features for robust automatic speech recognition", J. Acoust. Soc. Am. Volume 131, Issue 5, pp. 4134-4151