FEATURES FOR SPEAKER LOCALIZATION IN MULTICHANNEL BILATERAL HEARING AIDS

Joachim Thiemann, Simon Doclo, and Steven van de Par

Dept. of Medical Physics and Acoustics and Cluster of Excellence “Hearing4All”,
University of Oldenburg

ABSTRACT

Modern hearing aids often contain multiple microphones to enable the use of spatial filtering techniques for signal enhancement. To steer the spatial filtering algorithm it is necessary to localize sources of interest, which can be intelligently achieved using computational auditory scene analysis (CASA). In this article, we describe a CASA system using a binaural auditory processing model that has been extended to six channels to allow reliable localization in both azimuth and elevation, thus also distinguishing between front and back. The features used to estimate the direction are one level difference and five inter-microphone time differences of arrival (TDOA). Initial experiments are presented that show the localization errors that can be expected with this set of features on a typical multichannel hearing aid in anechoic conditions with diffuse noise.

Index Terms— Computational Auditory Scene Analysis, Localization, Multichannel Hearing Aids

1. INTRODUCTION

The human auditory system is remarkable in its ability to recognize and understand sounds in very complex acoustic scenes, with reverberation and multiple interfering sources present. This ability is termed the cocktail party effect which is studied in the field of Auditory Scene Analysis (ASA) [1]. The ability of separating a sound from a mixture is greatly enhanced if the listener is presented with a binaural signal, that is, if the listener can localize the source of interest (the “target” source) as well as the interfering sources, the enhancement in comparison to monaural presentation usually quantified as the spatial release from masking [2].

Coupled with research into the underlying processes that enable this ability in humans is research into mimicking this ability by signal processing algorithms (Computational Auditory Scene Analysis, or CASA). In this article, we focus on the sound localization aspect, based on work which uses a probabilistic model with a binaural auditory front-end [3]. The model presented in [3] is able to determine the azimuth of multiple sources based on interaural level differences (ILD) and interaural time differences (ITD) computed from a human auditory model (consisting of a gammatone filterbank and a simple neuronal transduction model). It was shown that the model is quite robust to reverberation and diffuse noise.

One interesting application of CASA for sound localization would be the use in assistive hearing devices (or hearing aids, HA). In particular, modern HAs typically use multiple microphones to enable spatial filtering (e.g. beamforming) to increase the Signal-to-Noise Ratio (SNR) [4]. This filtering is usually designed to only enhance sound coming from the front of the HA user. CASA-based localization could benefit HA users since the spatial filtering could be optimized based on the direction of the target and the interfering sources.

The problem we are considering is the localization of multiple sound sources using a microphone array formed by a bilateral hearing aid with multiple microphones on each hearing aid. There are numerous well-established algorithms for sound localization with microphone arrays [4, 5]. However, based on the results in [3], the particular geometry, variability, and presence of obstructions (the head and pinnae) suggest that a probabilistic auditory model based approach would be beneficial.

In this paper, we present investigations on extending the binaural model of [3] into a six-channel localization algorithm. We use the 6-channel HA described in [6], and expand the location space to an upper-hemispherical grid with 10° resolution in azimuth and elevation. In particular, we examine how the presence of noise affects the localization performance.

2. MODEL DESCRIPTION

The CASA localization model presented here is based on human audition such that it can be combined with further human auditory based processing. The localization model can be divided into four distinct stages. First, the acquisition of the audio signals with the multichannel HA, then the auditory-model-based processing converting the audio signal into features for localization. Next, the feature analysis uses a probabilistic classifier, and finally the probabilities are evaluated to make a localization decision.

We assume that the audio signal is captured by a bilateral HA with three microphones per side, denoting them LF for

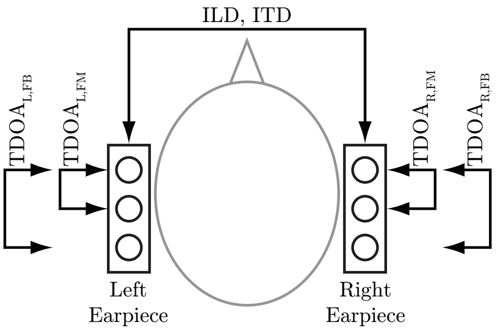

Fig. 1. Schematic of features for localization. ITD and ILD are computed from signals of the left and right front microphones. The TDOA features are computed for each side.

the frontmost microphone on the left side, LM for the middle microphone and LB for the rearmost microphone. We assume a symmetrical arrangement of microphones is located on the opposite ear (RF, RM, and RB).

In the peripheral auditory model processing, the six channels of audio from the bilateral HA are first separately passed through a $F = 32$ band fourth-order gammatone filterbank (GTFB). The GTFB uses phase-compensated filters to align temporal cues across bands. The center frequencies of the filters are equally distributed on an effective rectangular bandwidth (ERB) scale [7]. Neural transduction is simulated using half-wave rectification and square-root compression. We denote the resulting signals $h_{t,f}^{ch}$, where t is a time frame index, f is the gammatone filter index and ch indicates the channel (LF, LM, ..., RB).

After the peripheral processing, the features $\overrightarrow{X}_{t,f}$ for localizing sources are computed. Building upon the features used in [3], the additional microphone channels can provide features that allow for determining source elevation and resolve front-back confusion. As shown in Fig. 1, we use a set of 6 features ($\mathrm{ILD}_{t,f}, \mathrm{ITD}_{t,f}, \mathrm{TDOA}_{t,f}^{\mathrm{LFM}}, \mathrm{TDOA}_{t,f}^{\mathrm{RFM}}$, $\mathrm{TDOA}_{t,f}^{\mathrm{LFB}}$, and $\mathrm{TDOA}_{t,f}^{\mathrm{RFB}}$) per t, f bin. $\mathrm{ILD}_{t,f}$ represents the inter-aural level difference and is expressed as the energy difference between $h_{t,f}^{\mathrm{LF}}$ and $h_{t,f}^{\mathrm{RF}}$ in dB. The remaining features are computed using the normalized cross-correlation between two channels. Like the $\mathrm{ILD}_{t,f}$, the $\mathrm{ITD}_{t,f}$ is computed from $h_{t,f}^{\mathrm{LF}}$ and $h_{t,f}^{\mathrm{RF}}$. The TDOA features are computed only from signals within each side of the HA: $\mathrm{TDOA}_{t,f}^{\mathrm{LFM}}$ from $h_{t,f}^{\mathrm{LF}}$ and $h_{t,f}^{\mathrm{LM}}$, $\mathrm{TDOA}_{t,f}^{\mathrm{LFB}}$ from $h_{t,f}^{\mathrm{LF}}$ and $h_{t,f}^{\mathrm{LB}}$, and similar-for the right side. In order to compute the TDOA features with the required inter-sample accuracy, exponential interpolation [8] has been used to determine the maximum of the cross-correlation function.

2.1. Gaussian mixture model classifier

We consider the problem of localizing a sound in a probabilistic fashion. We assign to each direction a point on a sphere centred on the head of the HA user. For each combination of azimuth ϕ and elevation θ direction of a sound source, denoted by $\lambda_{(\phi,\theta)}$ we train a Gaussian Mixture Model that predicts the probability to observe the feature vector $\overrightarrow{X}_{t,f}$. This probability is denoted as $p(\overrightarrow{X}_{t,f}|\lambda_{(\phi,\theta)_k})$.

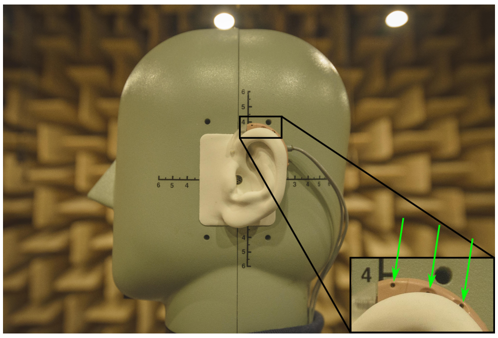

Fig. 2. One side of the binaural hearing aid with six microphones used for this study. There are three microphones on each side, the locations of which are indicated by the green arrows in the cutout.

The probabilities are modelled using a Gaussian mixture model (GMM) with $\mathcal{V}$ components, where each direction $\lambda_{(\phi,\theta)_k}$ is a class. All combinations of azimuth and elevation results in a total of $k = 1, \ldots, K$ classes. A separate GMM is trained for each frequency band and each class.

2.2. Localization decision

From the above described GMM, a decision for the location of a sound source can be made for each t, f bin by likelihood maximisation, as

$$\hat{\lambda}(t,f)_{(\phi,\theta)} = \underset{1\leq k\leq K}{\operatorname{argmax}}\, p(\overrightarrow{X}_{t,f}|\lambda_{(\phi,\theta)_k}). \tag{1}$$

In some cases, it is sufficient to make a localization decision per frame only, in which case it is possible to improve estimation by combining estimates over frequency using

$$\hat{\lambda}_T(t)_{(\phi,\theta)} = \underset{1\leq k\leq K}{\operatorname{argmax}} \sum_{f=1}^{F} \log\left(p(\overrightarrow{X}_{t,f}|\lambda_{(\phi,\theta)_k}) + \epsilon\right), \tag{2}$$

where ϵ is a small constant to limit the effect of very unlikely feature combinations on the estimated probability.

3. EVALUATION

In the present study, the goal is to assess the proposed localization scheme, to examine if the proposed set of features can be used for localizing sounds in the azimuth and elevation direction.

For our experiments, we use a new database of anechoic head-related impulse responses (HRIR) [9] recorded using the same hearing aid as in [6]. The left side of this device is shown in Fig. 2, showing the microphones with a distance of 15.6 mm from LF microphone to the LB microphone. The LM microphone is approximately in the center between the front and back microphones, the three microphones forming a shallow triangle. In contrast to the database of [6], the new recordings cover elevations from -64° to 90°. For this initial study, a subset of points is used to reduce computation time, covering only the upper hemisphere at 10° resolution in azimuth and elevation, for a total of 283 points. The grid is sparser at elevations of 70° (20° azimuth resolution) and above (30° azimuth resolution at 80° elevation, single point at 90° elevation) to avoid having a high density of points at the pole.

Speech samples are taken from the TIMIT database [10], with a sampling frequency of 16 kHz. After the peripheral auditory processing, features are computed using frames of 20 ms with 50% overlap, for a frame shift of 10 ms.

We use GMMs with $\mathcal{V} = 15$ components in all bands, with diagonal covariance matrices. The GMMs are trained by spatializing 10 randomly chosen sentences from the TIMIT database at all 283 points, then extracting the features only for frames where the energy for all microphone channels exceeds a given threshold, to avoid training on noise. Variance normalisation is used to equalize the dimensions during training. Training used the Expectation-Maximisation algorithm [11], with k-means clustering [12] to initialize the parameters.

Testing is performed by randomly selecting a male and a female speech sample from the TIMIT database excluding the samples used for training the model. The testing samples are spatialized at each point $\lambda_{\phi,\theta}$, then multichannel random gaussian diffuse speech-shaped noise (SSN) at 0, 5, 10, 15, 20 and 25 dB SNR is added. Using the same energy thresholding as used during training on the clean speech (in effect, an ideal voice activity detector, VAD), the features from active t, f bins are classified using the GMM.

4. RESULTS

The main result is shown in Fig. 3, showing the percentage of t, f bins correctly localized in anechoic conditions with various level of interfering SSN. The triangles (green dashed line) show the percentage of frames where localization was correct in both azimuth and elevation. By considering the azimuth and elevation components separately, we can see that especially at higher SNR azimuth estimation is significantly more robust than elevation estimation.

The effect of using additional features over the ones in [3] is shown in Fig. 4. Using ILD and ITD only, the performance over the entire semisphere is poor (16.1% at 25 dB), but adding one pair of TDOA features improves classification performance. The dashed line shows using $\mathrm{TDOA}_{t,f}^{\mathrm{LFB}}$ and $\mathrm{TDOA}_{t,f}^{\mathrm{RFB}}$ (40.0% at 25 dB), but similar performance is seen with $\mathrm{TDOA}_{t,f}^{\mathrm{LFM}}$ and $\mathrm{TDOA}_{t,f}^{\mathrm{RFM}}$ (39.2% at 25 dB).

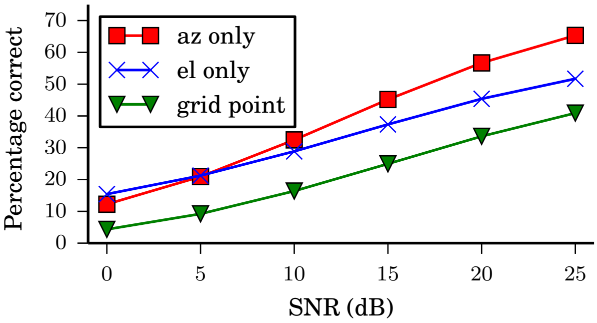

Fig. 3. Localization performance at different levels of diffuse SSN. The triangles show percentage of frames exactly localized. The squares and crosses show the performance if only azimuth or elevation are considered.

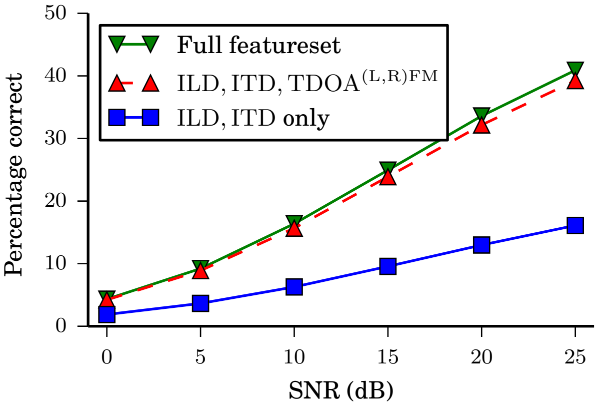

Fig. 4. Grid localization performance for subsets of features.

Using all TDOAs raises performance only marginally (40.9% at 25 dB).

In Fig. 5 the localization errors are shown in the azimuth and elevation direction as well as for the central angle estimated per t, f bin. It can be seen that localization errors are considerably smaller for elevation than for azimuth. This is due to the different ranges of errors; for elevation errors can be maximally 90°, while for azimuth they can be 180°. The central angle shows errors of maximally 50 degrees at an SNR of 0 dB. These relatively large errors are due to the fact that no specific provisions have been included in the localization algorithm to accommodate for the mismatch between the training which was done for sources without noise, and the evaluation which was done for sources presented in diffuse background noise. In fact errors decrease to about 15 degrees for high SNRs where the mismatch is much smaller.

The rate of correct localization on the grid points shown in Fig. 3 appears very low, but we observe that the localization accuracy is also strongly dependent on which frequency band is considered. As can be seen in Fig. 6, performance is

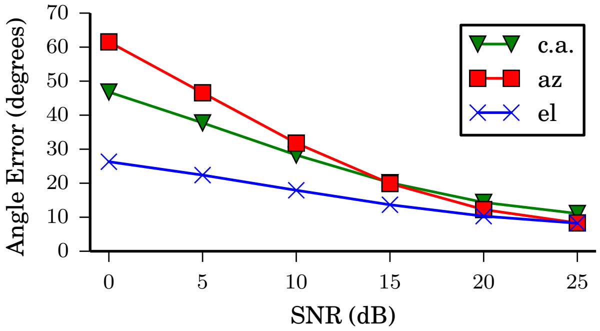

Fig. 5. Average localization error in degrees. Triangles show the error in the central angle (or great circle) sense, squares and crosses show the average error in the azimuth and elevation components respectively.

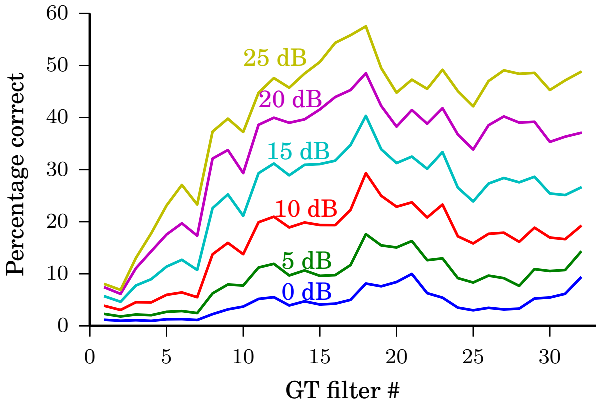

Fig. 6. Localization performance (for combined azimuth and elevation estimation) as a function of frequency.

very poor in low frequency bands. This can be explained by the fact that speech signals are less likely to have sufficient energy for reliable feature extraction in those bands, and thus the GMMs for those bands are less well trained.

Results of the combined azimuth and elevation estimation improve slightly if the log-likelihood results are combined over frequency as described by Eq. (2). For example, at 25 dB, correct classification raises from 40.9% to 52.2%, and at 20 dB from 33.7% to 44.7%. At SNR 15 dB and below, the percentage of correct classification is about 1.4 times higher with frequency integration.

For visualizing the method for a practical application, Fig. 7 shows the t, f maps for a segment of speech which was rendered at azimuth 40° and elevation 20°. The signal was mixed with diffuse noise at 10 dB SNR. Panel A shows the energy in each t, f bin, with blank areas omitted using the VAD. Panel B shows the azimuth estimation error, while panel C shows the error in the elevation estimate.

Two issues can be illustrated by panels B and C. First, how localization performance is dependent on the signal energy: where panel A shows high energy, the localization error shown in B and C tends to be low. Second, it can be seen that in many cases the localization error is only offset by a single step in the azimuth and elevation grid, predominantly in the elevation estimate. This skews the results shown in Fig. 3 and 6 which do not include a measure of estimation error.

5. DISCUSSION

In this study we are presenting an initial assessment of a method to localize sound sources using a six-channel bilateral HA. Localization is performed using a probabilistic framework, specifically a set of GMMs that classify six-dimensional feature vectors. The GMMs (one for each frequency band of a auditory model analysis) compute the probabilities that the observed features originate from any of 283 points on a grid of a hemisphere.

It was shown previously that the ILD and ITD features work well for determining the azimuth of a sound source provided the source is located in the front and near the equator [3]. The aim of this study is to extend the method of [3] to resolve the front-back confusion, and if possible, perform localization on the vertical axis as well. While the results presented here are limited in scope (significant sources of errors, such as reverberation and localizable interfering sources are not considered), we show that using two microphones per side near the ear and the described features can discriminate directions to at least 10° resolution in elevation and thus of course also solve the front-back confusion problem.

As expected, we find that it is important to have sufficient training data. In our experiments, we find that especially at the lower frequencies, speech energy was too sparse to properly train the GMMs from our training set. Combining estimates across frequency to get per-time frame localization only provides a small benefit, but this may also be explained by the poor quality of training at low frequencies.

Further research is required however to allow this scheme to be practical with current HA technology. One issue is the large number of classes to be classified caused by the two-dimensional azimuth-elevation grid. Reducing the number of grid points would reduce complexity during classification (important due to power constraints in HAs) as well as during training. Without reducing the localization accuracy, this could be achieved by eg. using a hierarchical approach, where the location of the source is first limited to the left or right hemisphere, followed by a more fine grained classifier.

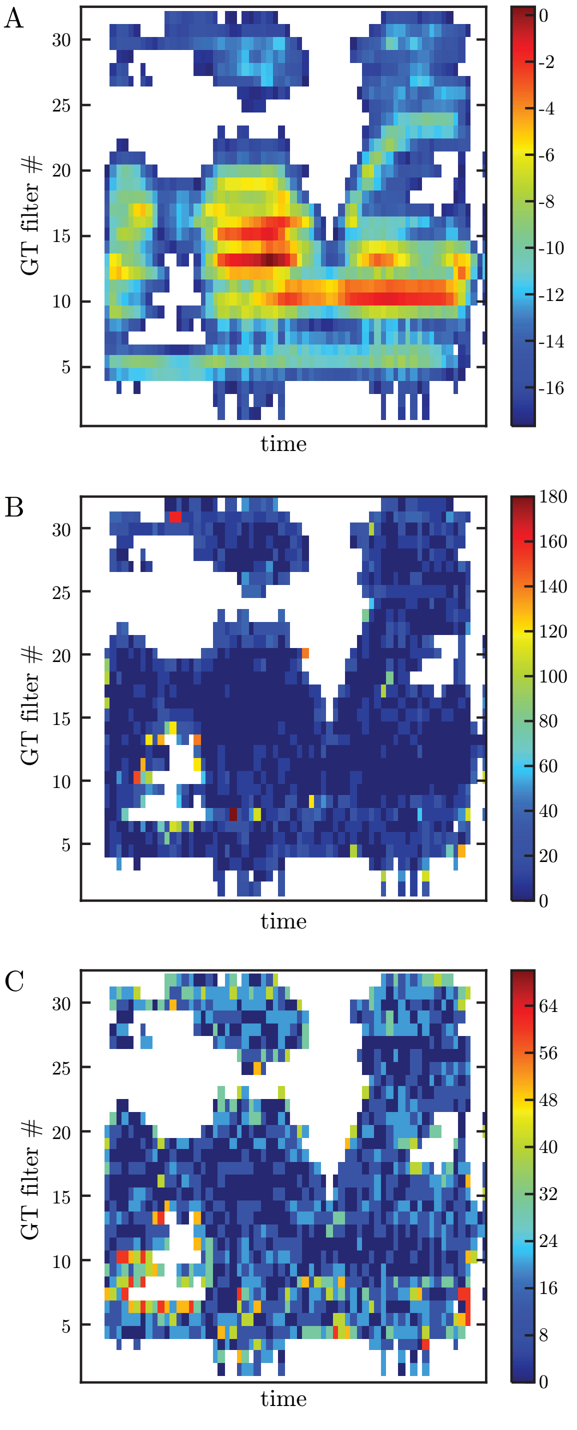

Fig. 7. Example localization error for a segment of speech, showing only those time/frequency (t, f) regions that the (ideal) VAD detected. Panel A shows the energy in the t, f bins (in dB), panel B shows the azimuth estimation error, panel C the elevation estimation error. This speech sample was rendered at an azimuth of 40° and elevation of 20°, and mixed with 10 dB SSN.

REFERENCES

[1] A. S. Bregman, *Auditory scene analysis: The perceptual organization of sound*, MIT press, 1994.

[2] M. L. Hawley, R. Y. Litovsky, and J. F. Culling, “The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer,” *The Journal of the Acoustical Society of America*, vol. 115, no. 2, pp. 833–843, 2004.

[3] T. May, S. van de Par, and A. Kohlrausch, “A probabilistic model for robust localization based on a binaural auditory front-end,” *IEEE Trans. Audio, Speech, and Language Processing*, vol. 19, no. 1, pp. 1–13, 2011.

[4] S. Doclo, W. Kellermann, S. Makino, and S. Nordholm, “Multichannel signal enhancement algorithms for assisted listening devices,” *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 16–17, Mar. 2015.

[5] M. S. Brandstein and D. B. Ward, *Microphone Arrays: Signal Processing Techniques and Applications*, Springer, 2001.

[6] 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 Journal on Advances in Signal Processing*, 2009.

[7] B. R. Glasberg and B. J. C. Moore, “Derivation of auditory filter shapes from notched-noise data,” *Hearing Research*, vol. 47, no. 1-2, pp. 103–138, 1990.

[8] L. Zhang and X. Wu, “On cross correlation based discrete time delay estimation,” in *Proc. IEEE Int. Conf. Acous., Speech and Sig. Proc. (ICASSP)*, 2005, vol. 4, pp. 981–984.

[9] J. Thiemann and S. van de Par, “Multiple model high-spatial resolution HRTF measurements,” in *Proc. DAGA 2015*, Nürnberg, Germany, Mar. 2015, to be published.

[10] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, “Timit acoustic-phonetic continous speech corpus,” Tech. Rep. NISTIR 4930, National Institute of Standards and Technology, Gaithersburg, MD, 1993.

[11] A. Dempster, N. Laird, and D. Rubin, “Maximum likelihood estimation from incomplete data via the EM algorithm,” in *Journal of the Royal Statistical Society: Series B*, 1977, vol. 39, pp. 1–38.

[12] S. P. Lloyd, “Least squares quantization in PCM,” *IEEE Trans. Information Theory*, vol. 28, no. 2, pp. 129–137, 1982.