MULTI-CHANNEL SIGNAL ENHANCEMENT WITH SPEECH AND NOISE COVARIANCE ESTIMATES COMPUTED BY A PROBABILISTIC LOCALIZATION MODEL

Jörn Anemüller and Hendrik Kayser

Computational Audition Group
Medical Physics Unit and Cluster of Excellence Hearing4all
Carl von Ossietzky Universität Oldenburg
Oldenburg, Germany

ABSTRACT

Classic approaches to multi-channel signal enhancement rely on model assumptions regarding speech source relative transfer functions and noise covariance matrix, or on estimates thereof obtained in, e.g., speech pauses. To alleviate these constraints, we here investigate an approach to adaptive estimation of the speech (target) source and noise related acoustic parameters based on localized speech probability estimates. The latter are computed from a discriminatively trained speech localization algorithm previously proposed [1]. A-priori knowledge of temporal segments that contain noise only, thus, is not required. A standard MVDR system is employed for subsequent signal enhancement.

Evaluation is carried out for anechoic and reverberant conditions using 6-channel input signals recorded with a bilateral hearing-aid geometry. Results indicate that the proposed method outperforms an anechoic, isotropic-noise model when a-priori information is unavailable: I.e., in (a) anechoic conditions with localized interferer in addition to isotropic noise, and (b) reverberant conditions. In these conditions, the proposed method and constrained versions thereof improve upon the free-field isotropic noise model by up to 16.1 dB and 7.7 dB SINR, respectively.

Index Terms— Acoustic source localization, multi-microphone signal enhancement, machine learning for signal processing

1. INTRODUCTION

An important component in multi-channel signal enhancement systems is information about differences in sound propagation from a target source to the sensors of a microphone array. These relative transfer functions (RTF) are used to conduct spatial filtering, e.g., serve as steering vectors in beamforming methods. In the context of hearing aid signal processing, RTFs are influenced by acoustic transfer characteristics of the head such that the interaural transfer function (ITF) has to be considered. The ITF plays an important role for signal enhancement and sound reproduction in hearing aids [2]. In the former, it has to be taken into account during spatial filtering since a free-field sound propagation assumption does not hold. In the latter, preservation of the ITF or binaural cues captured by the ITF is relevant due to their role in speech intelligibility in spatial conditions [3].

ITFs computed from head-related transfer functions (HRTF) can serve as models to derive parameters for signal enhancement [4]. The same holds for assumptions about background noise–as long as the surrounding noise field is diffuse, assuming an isotropic noise model is robust.

However, the assumption of purely head-related information is invalid when the acoustic scenario encompasses real room characteristics including reverberation. In this case, it can be beneficial to estimate ITFs from the data. In case that noise contains localized interfering sources, an isotropic noise model is less effective and spatial filter and noise characteristics need to be estimated from the input data. A single-channel method for noise estimation is minimum statistics [5, 6], which relies on the assumption that speech variations are faster than changes in noise power spectral density (PSD). The case of non-stationary noise has also been tackled by incorporating speech presence probability (SPP), cf. [7, 8, 9] where noise is estimated in minima of speech activity or based on estimates of the SNR [10]. The aforementioned methods have been shown to be robust. If the interfering sound also contains speech, SPP information becomes unreliable and variations in the interfering sound are correlated to variations of the target, making it difficult to find optimal smoothing parameters for minimum statistics estimates. A multi-channel sensor setup provides the possibility to identify a target by location-related information. E.g. the maximum peak position of cross-correlation functions between a sensor pair can be used to identify target source activation [11].

Supported by DFG grants FOR 1732 “Individualized Hearing Acoustics” and SFB/TRR 31 “The Active Auditory System”. The authors wish to thank Daniel Marquardt and Simon Doclo for comments.

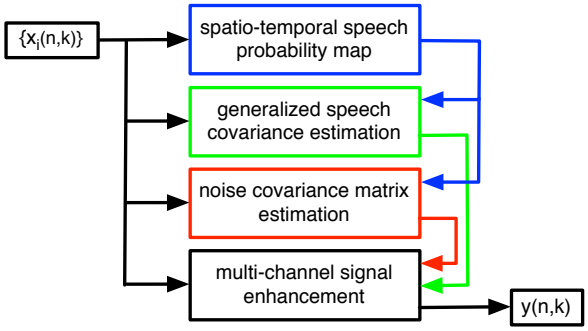

Fig. 1. Processing diagram of the proposed algorithm.

In the proposed approach, we use spatial speech presence probability to obtain a direct estimate of the target source steering vector for a six-channel hearing aid setup. In order to enhance the target signal and reduce interference from other sound sources, minimum-variance-distortionless-response (MVDR) signal enhancement is performed. An estimate of the noise covariance matrix, which is essential for MVDR processing, encodes information about the noise field and is obtained by estimating the noise covariance matrix with a probabilistic weighting that is inversely related to target source activity. Acoustic scenarios including one target source and combinations of a diffuse noise field and a localized interfering talker in an anechoic and a reverberant office room were investigated. The proposed parameter estimation method and use of model-based spatial filters, as well as combinations of both were compared in these scenarios. Results show that probabilistic estimation of the spatial filters outperforms model-based approaches if exact a-priori information is unavailable.

2. METHODS

The method proposed here is conducted in four steps, cf. Fig. 1: First, a spatio-temporal analysis of the acoustic scene is conducted with a probabilistic source localization method that estimates for each time-point n and discretized location index θ the a-posteriori probability of speech and, thus, permits identification of the maximum-a-posteriori speech source location. Estimated speech probability is used subsequently to determine the corresponding (generalized) speech covariance matrix as well as the noise-covariance matrix induced by interfering sources. Multi-channel signal enhancement is then carried out with filter parameters obtained from the estimated covariance matrices.

2.1. Probabilistic source localization

Reliable estimation of spatially localized speech source probability is the first step in the proposed method which the subsequent steps build upon. We here employ the discriminative classification approach to probabilistic sound source localization described in [1]. It estimates the a-posteriori probability of speech for a defined set of source locations θ using short-term generalized cross-correlation [12] with phase transform (GCC-PHAT) as input features. These are used to train a bank of discriminative linear support-vector machine (SVM) classifiers, with presence and absence of a speech source for a given position serving as the training class label. Each SVM is followed by a generalized linear model (GLM) classifier, that converts SVM decision values into the estimated spatial source probability map $p^S(\theta, n)$. Let $\mathcal{G}_\theta(\cdot)$ denote the combined localizer for direction θ as described above, then the source probability map is given by

$$p^S(\theta, n) = \mathcal{G}_\theta(\mathbf{x}(n, k)) \tag{1}$$

for location index θ, time frame index n, spectral band index k and multi-channel STFT input vector $\mathbf{x}(n, k)$.

2.2. Speech and noise covariance matrix estimation

Knowledge of the localizer $\mathcal{G}_\theta(\cdot)$ corresponds to implicit knowledge of a spatial source model. However, a model that is appropriate for source localization does not necessarily imply knowledge of spatial filter parameters that would permit to optimally enhance a target speech source and maximally attenuate interference from other sound sources. For one, the learned localization model may not be precise enough for spatial signal enhancement. In realistic applications, we may further wish to utilize a localiztion model trained in one environment (e.g., under anechoic conditions) also in other more realistic test conditions for signal enhancement. However, the spatial source model learned by the localizer $\mathcal{G}_\theta(\cdot)$ does contain valuable information that should be maximally exploited in order to perform fast and robust spatial signal enhancement in realistic situations.

To this end, we present a novel approach for estimation of spatial filters from the multi-channel input signals without use of an explicit model of sound propagation, while still exploiting the source probability map obtained in Sec. 2.1 and the learned knowledge about spatial source positions that is implicitly encoded in it. The estimated speech probability map values $p^S(\theta, n)$ are used as weights to compute a generalized speech covariance matrix $\mathbf{\Phi}(k|\theta)$ conditioned on speech direction θ, with ij-element

$$[\mathbf{\Phi}(k|\theta)]_{ij} \equiv \frac{1}{N} \sum_{n=1}^{N} p^S(\theta, n)\, c_{ij}(n, k)^{-1}\, x_i^*(n, k)\, x_j(n, k) \tag{2}$$

where the average is computed over N contiguous STFT frames and $c_{ij}(n, k)^{-1}$ are spectral weights. Choosing

$$c_{ij}(n, k) = |x_i(n, k)|\, |x_j(n, k)|, \tag{3}$$

we obtain a measure similar to the normalized cross-power spectrum as used in [13, 14], albeit conditioned on location θ.

	Speech covariance	Noise covariance
PrS+PrN	prob. model (Eq. 2)	prob. model (Eq. 5)
FfS+PrN	free-field HRTF model	prob. model (Eq. 5)
PrS+IsoN	prob. model (Eq. 2)	isotr. model
FfS+IsoN	free-field HRTF model	isotr. model

Table 1. Summary of combined models for generalized speech covariance and noise covariance estimation, that were investigated experimentally.

In order to compute in an analogous way the noise covariance matrix conditioned on speech source direction θ, we define a robust estimate of noise probability $p^{\mathrm{N}}(\theta, n)$ as

$$p^{\mathrm{N}}(\theta, n) = \begin{cases} \gamma\,(1 - p^{S}(\theta, n)), & p^{S}(\theta, n) < p_0 \\ 0, & p^{S}(\theta, n) \geq p_0 \end{cases} \quad (4)$$

with a confidence threshold p_0 and scaling factor γ. The ij-element $[\mathbf{R}(k|\theta)]_{ij}$ of θ-conditioned noise covariance matrix $\mathbf{R}(k|\theta)$ is estimated as

$$[\mathbf{R}(k|\theta)]_{ij} = \frac{1}{N} \sum_{n=1}^{N} p^{\mathrm{N}}(\theta, n)\, x_i^*(n, k)\, x_j(n, k). \quad (5)$$

2.3. Multi-channel signal enhancement

While the proposed scheme is not specific to a particular multi-channel enhancement algorithm, we employ the minimum-variance distortionless-response (MVDR) method. In the spectral domain implementation used here, it uses a projection operator $\mathbf{w}(\theta, k)$ that is applied to the multi-channel short-term Fourier transform $\mathbf{x}(n, k)$ of the input signals. Output signals are obtained as

$$y(n, k|\theta) = \mathbf{w}^{\mathsf{H}}(k|\theta)\, \mathbf{x}(n, k). \quad (6)$$

The projection operator $\mathbf{w}$ is obtained from a steering vector $\mathbf{d}$ and noise covariance matrix $\mathbf{R}$ as

$$\mathbf{w}(k|\theta) = \frac{\mathbf{R}^{-1}(k|\theta)\, \mathbf{d}(k|\theta)}{\mathbf{d}^{\mathsf{H}}(k|\theta)\, \mathbf{R}^{-1}(k|\theta)\, \mathbf{d}(k|\theta)}. \quad (7)$$

The steering vector $\mathbf{d}(k|\theta)$ for speech source direction θ is obtained from the generalized speech covariance matrix Eq. 2 by choosing an arbitrary but fixed reference channel i^* and extracting the normalized i^*-th row elements according to

$$d_j(k|\theta) = [\mathbf{\Phi}(k|\theta)]_{i^* j} \,/\, |[\mathbf{\Phi}(k|\theta)]_{i^* j}|, \quad (8)$$

retaining inter-microphone phase and neglecting (possible) level differences. The maximum-a-posteriori speech position θ^* was chosen as the location value for the MVDR filter.

For baseline comparison, steering vector and noise-covariance were also derived from an anechoic free-field model with head-related transfer functions (in case of $\mathbf{d}$) and a free-field isotropic noise model (in case of $\mathbf{R}$). See Tab. 1 for a summary of investigated conditions for combined source- and noise-model.

3. EXPERIMENTS

We evaluated the signal enhancement performance of the MVDR beamformer (7) with parameters estimated by all approaches summarized in Tab. 1. A six-channel binaural hearing aid geometry setup was used for MVDR beamforming of which four channels (front and rear microphone pairs) were employed for estimation of the spatial source probability map as described in [1] with discrete azimuth angles $\theta = 0^\circ, \ldots, 355^\circ$ in steps of 5°. STFT frame length was 10 ms with 25 % shift. For the estimation of the steering vector and the noise covariance matrix, we utilized the a-priori known target DOA, indicated by $\hat{\theta}$, to select either the according probability weighting from the map for the estimation or the model steering vector. Groundtruth DOA values were used in order to separate localization accuracy [1] from the filter estimation approach pursued here. As reference channel in the spatial filter, the left frontal hearing aid microphone was used. The parameters for the noise covariance estimation were set to $p_0 = 0.99$ and $\gamma = (1 - p_0)/\max_{t \in T}(p(\hat{\theta}, n))$ with T containing all 10 ms-samples from the current test signal. No temporal smoothing, apart from the weighting with $p^S(\theta, n)$ and $p^N(\theta, n)$, was used.

3.1. Acoustic Data

All acoustic signals used in the experiments were generated by filtering single-channel speech signals with head-related impulse responses (HRIR) captured with a binaural hearing aid setup with three microphones on each side of the head [15]. Measurements for various source positions from two different environments were used: an anechoic chamber and an office room. Three-seconds-long speech signals, each from the same (female or male) speaker, were randomly sampled from the TIMIT speech database [16]. A head-related isotropic noise field was obtained by convolution of speech shaped noise [17] with anechoic HRIRs from the whole horizontal plane. Processing was performed at a sampling rate of 16 kHz.

The resulting signals were combined to a set of test scenarios containing a target speech source, an interfering speaker from a different position and isotropic noise. Thereby the energy ratio between target and interferer, signal-to-interference ratio (SIR), was varied between -10 dB, 0 dB, 10 dB and ∞ dB, as well as the energy ratio between target and noise field, signal-to-noise-ratio (SNR). The resulting overall acoustic complexity is then represented by the signal-to-noise-plus-interferer-ratio (SINR). In the anechoic environment, the target was located in the left semi-circle at DOAs ranging from -180° (back) to 0° (front) in steps of 30°. The interfering speaker occurred on the whole circle around the head in the range from -165° to $+165^\circ$ in steps of 30°. In the office environment the source locations were limited to the frontal semi-circle, such that the target position ranged from -90°

Anechoic environment						Office environment					
Input		SINR improvement (dB)				Input		SINR improvement (dB)			
SIR (dB)	SNR (dB)	PrS +PrN	FfS +PrN	PrS +IsoN	FfS +IsoN	SIR (dB)	SNR (dB)	PrS +PrN	FfS +PrN	PrS +IsoN	FfS +IsoN
-10	-10	3.0	**9.6**	-1.0	6.9	-10	-10	**6.0**	4.5	2.5	3.7
-10	0	7.7	**15.1**	-1.5	8.9	-10	0	**8.2**	7.4	1.4	3.4
-10	10	12.9	**20.8**	-0.8	10.0	-10	10	**10.2**	9.8	1.6	3.3
-10	∞	18.6	**26.3**	0.8	10.2	-10	∞	**10.9**	10.6	2.1	3.2
0	-10	1.7	**7.8**	1.4	6.1	0	-10	**5.6**	2.7	5.0	4.3
0	0	2.6	**9.1**	2.2	6.9	0	0	3.8	2.0	**4.0**	3.7
0	10	7.0	**13.4**	2.6	8.8	0	10	**5.0**	3.6	4.2	3.4
0	∞	16.3	**20.8**	3.7	10.2	0	∞	**6.4**	5.4	4.3	3.3
10	-10	1.7	**7.6**	1.7	6.1	10	-10	**6.1**	2.5	5.3	4.4
10	0	1.5	**7.3**	3.5	6.2	10	0	3.1	-0.0	**5.2**	4.3
10	10	2.7	**8.1**	4.6	6.9	10	10	1.0	-1.1	**5.6**	3.8
10	∞	12.9	**14.7**	5.8	10.2	10	∞	1.4	0.1	**6.1**	3.2
∞	-10	1.9	**7.6**	1.7	6.1	∞	-10	**6.3**	2.6	5.1	4.5
∞	0	0.9	**7.1**	3.5	6.1	∞	0	4.1	0.1	**5.2**	4.6
∞	10	2.2	**6.3**	4.6	6.1	∞	10	0.7	-2.5	**6.4**	4.5

Table 2. Improvement in SINR obtained with probabilist estimates of speech (PrS) and noise (PrN) covariance, and with a-priori known free-field HRTF speech (FfS) and isotropic noise (IsoN) models, respectively in all possible combinations. Results shown for the anechoic (left) and reverberant office (right) environment.

to $+90°$ and the interferer from $-75°$ to $+75°$ $30°$ same step size. Four realizations of all possible combinations of target and interferer positions, SIR and SNR were generated resulting in 6832 signals in the anechoic environment and 3472 in the office room.

3.2. Results

In Tab. 2 the signal enhancement performance measured in terms of SINR improvement of the four approaches under test is summarized. For both environments, anechoic and office, average results over all source position combinations, the four realizations of each and both reference channels are shown dependent on the input SIR and SNR. In the anechoic environment (left table) the combination of the a-priori known steering vector and the estimated noise covariance matrix is most successful in all conditions, yielding SINR enhancement up to 26.3 dB. In the office room, where model steering vectors do not provide perfect information, using probabilistic estimates for both parameters (PrS+PrN) yields the best results for conditions with low to moderate SIR followed by the FfS+PrN combination. The latter, however, does not achieve much signal enhancement for SIRs above 0 dB and is even detrimental in some cases. For these SIR conditions the combination of estimated steering vector and noise model (PrS+IsoN) outperforms the other approaches.

4. SUMMARY AND DISCUSSION

In this contribution, we presented an approach to the estimation of steering vector and noise covariance matrix for MVDR beamforming. Based on spatial source presence probability maps, obtained with a machine learning-based localization method, target source activity was measured and steering vectors were estimated in the STFT domain with the resulting probabilistic weights. From target source probability, the inversely related noise probability was derived and used to estimate noise statistics. Incorporating these estimates into the spatial filters of an MVDR beamformer, signal enhancement performance was compared to an entirely HRTF-model-based approach and to two partially model-based approaches, showing that spatial probability delivers suitable information for robust spatial filter estimation. However, the probabilistic estimation-based approach generalized well to a reverberant environment and was shown to be appropriate for real-world scenarios where a-priori knowledge is not available. The proposed scheme for noise covariance estimation may account for mixtures of a diffuse noise field and localized interfering speech without the need for additional parameter estimation.

References

[1] H. Kayser and J. Anemüller, "A discriminative learning approach to probabilistic acoustic source localization," in *Proc. IWAENC 2014 – International Workshop*

on Acoustic Echo and Noise Control, 2014, pp. 100–104.

[2] Daniel Marquardt, Elior Hadad, Sharon Gannot, and Simon Doclo, "Theoretical Analysis of Linearly Constrained Multi-Channel Wiener Filtering Algorithms for Combined Noise Reduction and Binaural Cue Preservation in Binaural Hearing Aids," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 23, no. 12, pp. 2384–2397, 2015.

[3] Adelbert Bronkhorst, "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions," *Acta Acustica united with Acustica*, vol. 86, pp. 117–128, 2000.

[4] Kamil Adiloglu, Hendrik Kayser, Regina M. Baumgärtel, Sanja Rennebeck, Mathias Dietz, and Volker Hohmann, "A Binaural Steering Beamformer System for Enhancing a Moving Speech Source," *Trends in Hearing*, vol. 19, 2015.

[5] Rainer Martin, "Spectral Subtraction Based on Minimum Statistics," in *Proc. EUSIPCO 94*, 1994, pp. 1182–1185.

[6] Rainer Martin, "Noise Power Spectral Density Estimation Based on Optimal Smoothing and Minimum Statistics," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, pp. 504–512, 2001.

[7] Israel Cohen, "Noise Spectrum Estimation in Adverse Environments: Improved Minima Controlled Recursive Averaging," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 5, pp. 466–475, 2003.

[8] Israel Cohen, "Relative Transfer Function Identification Using Speech Signals," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 5, pp. 451–459, 2004.

[9] Ronen Talmon, Israel Cohen, and Sharon Gannot, "Relative Transfer Function Identification Using Convolutive Transfer Function Approximation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 17, no. 4, pp. 546–555, 2009.

[10] Timo Gerkmann and Richard C. Hendriks, "Unbiased MMSE-Based Noise Power Estimation with Low Complexity and Low Tracking Delay," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 4, pp. 1383–1393, 2012.

[11] Bram Cornelis, Marc Moonen, and Jan Wouters, "Binaural Voice Activity Detection for MWF-Based Noise Reduction in Binaural Hearing Aids," in *Proc. EUSIPCO 2011*, 2011, vol. 1, pp. 486–490.

[12] C. Knapp and G. Carter, "The Generalized Correlation Method for Estimation of Time Delay," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976.

[13] M. Omologo and P. Svaizer, "Acoustic Event Localization Using a Crosspower-Spectrum Phase Based Technique," *Proc. ICASSP 1994. IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. ii, no. 2, pp. II/273–II/276, 1994.

[14] M. Omologo and P. Svaizer, "Use of the Crosspower-Spectrum Phase in Acoustic Event Location," *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 3, pp. 288–292, 1997.

[15] 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*, vol. 2009, no. 1, pp. ID 298605, 2009.

[16] John S. Garofolo, Lori F. Lamel, William M. Fisher, Jonathan G. Fiscus, David S. Pallett, Nancy L. Dahlgren, and Victor Zue, "TIMIT Acoustic-Phonetic Continuous Speech Corpus," *CDROM*, 1993.

[17] Wouter a Dreschler, Hans Verschuure, Carl Ludvigsen, and Soren Westermann, "ICRA Noises: Artificial Noise Signals with Speech-Like Spectral and Temporal Properties for Hearing Instrument Assessment," *Audiology*, vol. 40, no. 3, pp. 148–157, 2001.