A Study on Online Source Extraction in the Presence of Changing Speaker Positions

Jens Heitkaemper[1(✉)], Thomas Fehér[2], Michael Freitag[2], and Reinhold Haeb-Umbach[1]

[1] Department of Communications Engineering, Paderborn University, Pohlweg 47-49, 33098 Paderborn, Germany
{heitkaemper,haeb}@nt.upb.de
[2] voice INTER connect GmbH, Ammonstr. 35, 01067 Dresden, Germany
{thomas.feher,michael.freitag}@voiceinterconnect.de

Abstract. Multi-talker speech and moving speakers still pose a significant challenge to automatic speech recognition systems. Assuming an enrollment utterance of the target speakeris available, the so-called SpeakerBeam concept has been recently proposed to extract the target speaker from a speech mixture. If multi-channel input is available, spatial properties of the speaker can be exploited to support the source extraction. In this contribution we investigate different approaches to exploit such spatial information. In particular, we are interested in the question, how useful this information is if the target speaker changes his/her position. To this end, we present a SpeakerBeam-based source extraction network that is adapted to work on moving speakers by recursively updating the beamformer coefficients. Experimental results are presented on two data sets, one with artificially created room impulse responses, and one with real room impulse responses and noise recorded in a conference room. Interestingly, spatial features turn out to be advantageous even if the speaker position changes.

Keywords: Robust speech recognition and Multi-channel speech enhancement · Speaker adaptation · Conference scenario

1 Introduction

In recent years, robust multi-channel Automatic Speech Recognition (ASR) has been a major focus of research which led to large improvements in transcription accuracy [1]. These gains are mainly due to the development of novel neural network (NN) architectures [2,3] and the combination of neural network (NN)s with well-known speech enhancement techniques like statistical beamforming [4,5] and dereverberation [6]. However, realistic application environments often still present a challenge to Automatic Speech Recognition (ASR) systems because of overlapped speech and moving speakers [7].

C. Martín-Vide et al. (Eds.): SLSP 2019, LNAI 11816, pp. 198–209, 2019.
https://doi.org/10.1007/978-3-030-31372-2_17

Recently, several promising approaches for source separation [8–10] and source extraction [11–14] in the presence of multiple simultaneously active speakers were presented. This contribution focuses on source extraction, where one is interested in only one of the speakers in a mixture.

Different techniques have been proposed to identify the target speaker. In the so-called SpeakerBeam (SB) approach, the target speaker is identified by an enrollment, also called adaptation utterance (AU), which the speaker has to provide in advance and from which his/her spectral characteristics are obtained [11,13]. This information is then used to guide a neural network for mask estimation to focus on the target speaker.

The desired speaker can also be identified by the speaker's position as in [14], where a neural network uses oracle information of the target speaker location to focus on a specific source, assuming the speaker does not move. In [12] a beamforming vector is estimated on a keyword preceding the user's command. While this setting may be appropriate for operating a digital home assistant, in many other application scenarios, such as a meeting, it would be very inconvenient if utterances had to start with a keyword to identify and locate the target speaker. Additionally, a fixed beamformer estimated on a AU or a keyword cannot capture changes in the speaker position or noise statistics.

In this contribution we are concerned with the extraction of a target speaker from multi-talker speech. We would like to take advantage of the spatial diversity present in the speech mixture while facing the problem that the spatial characteristics of the target speaker may change. To be specific, we allow speakers to change their position from one utterance to the next. The proposed system is based on the SpeakerBeam concept developed in [11], which we extend to a block-online source extraction system. We assume that an AU has been recorded for each speaker in advance, when no competing speakers are present. This AU is used to estimate a beamforming vector, which is applied to the AU itself to improve the extraction of the speaker embedding vector, which captures the target speaker's spectral characteristics. It is further used to enhance the distorted input signal of the neural network. Thereby, emphasizing all signal components originating from the position of the target speaker during the AU. To cope with subsequent changing speaker positions, the beamformer coefficients are recursively updated.

Spatial features have proven very effective in enhancing the performance of neural network supported acoustic beamforming [15–17]. It is, however, unclear, to which extent they are also useful if speaker positions change. We therefore test the effectiveness of those features by comparing results for stationary speakers and speaker position changes between utterances. It will be shown that spatial features computed on the speech mixtures remain to be effective.

The paper is structured as follows: In Sect. 2 a short overview over the system is presented, where Sect. 2.1 focuses on the beamforming vector estimation and Sect. 2.2 explains the neural network structure used for mask estimation. In Sect. 3 the systems are evaluated on a database presented in Sect. 3.1. Final conclusions are drawn in Sect. 4.

2 System Overview

We assume a multi-channel signal captured by D microphones. In the short-time Fourier transform (STFT) domain the overlapped speech $\mathbf{Y}$ and the adaptation utterance $\mathbf{A}$ can be expressed as

$$\mathbf{Y}(t,f) = \mathbf{X}_i(t,f) + \sum_{j \neq i} \mathbf{X}_j(t,f) + \mathbf{N}(t,f) \tag{1}$$

$$\mathbf{A}(t,f) = U(t,f) + \mathbf{N}(t,f). \tag{2}$$

Here, $\mathbf{Y}(t,f)$, $\mathbf{N}(t,f)$ and $\mathbf{X}_k(t,f)$ are the STFT coefficient vectors of the speech mixture, of the noise and of the k-th source image at the microphones. $\mathbf{A}(t,f)$ represents the distorted and $U(t,f)$ the clean AU. The time and frequency indices t and f will be dropped wherever possible without sacrificing clarity.

2.1 Beamforming

Speech enhancement is done using the well known Minimum Variance Distortionless Response (MVDR) beamformer, which minimizes the noise power without introducing distortions on signals originating from a target direction, by optimizing the cost function [18]:

$$\mathbf{F}_{\text{MVDR}} = \underset{\mathbf{F}}{\operatorname{argmin}}\, \mathbf{F}^{\mathsf{H}} \boldsymbol{\Phi}_{\mathbf{NN}} \mathbf{F} \quad \text{s.t.} \quad \mathbf{F}^{\mathsf{H}} \tilde{\mathbf{H}} = 1, \tag{3}$$

where $\tilde{\mathbf{H}} = [1, ..., \tilde{H}_D]^{\mathsf{T}}$ is the target speaker acoustic transfer function (ATF) normalized to a reference microphone, which is called relative transfer function (RTF), and $\boldsymbol{\Phi}_{\mathbf{NN}}$ is the noise spatial correlation matrix (SCM).

We employ the solution of the MVDR cost function in the form presented in [19]:

$$\mathbf{F}_{\text{MVDR}} = \frac{\boldsymbol{\Phi}_{\mathbf{NN}}^{-1} \boldsymbol{\Phi}_{\mathbf{XX}}}{\operatorname{tr}\left\{\boldsymbol{\Phi}_{\mathbf{NN}}^{-1} \boldsymbol{\Phi}_{\mathbf{XX}}\right\}} \mathbf{u}, \tag{4}$$

where $\mathbf{u}$ is a unit vector pointing to the reference microphone, $\operatorname{tr}\{\cdot\}$ is the trace operator and $\boldsymbol{\Phi}_{\mathbf{XX}}$ is the target speech SCM. Here, the target speech SCM is forced to follow the rank-1 approximation [20] by using:

$$\tilde{\boldsymbol{\Phi}}_{\mathbf{XX}} = \mathbf{a}\mathbf{a}^H \cdot \operatorname{tr}\{\boldsymbol{\Phi}_{\mathbf{XX}}\} / \operatorname{tr}\{\mathbf{a}\mathbf{a}^H\} \tag{5}$$

with $\mathbf{a} = \boldsymbol{\Phi}_{\mathbf{NN}} \mathcal{P}\left\{\boldsymbol{\Phi}_{\mathbf{NN}}^{-1} \boldsymbol{\Phi}_{\mathbf{XX}}\right\}$ and $\mathcal{P}\{\cdot\}$ as the principal component of the matrix given in parentheses. Both the noise and target speaker SCMs are estimated using speech and noise masks M_ν, where $\nu \in [\mathbf{X}, \mathbf{N}]$. In case of block-wise estimation a recursive update of the SCM is applied [21]:

$$\boldsymbol{\Phi}_{\nu\nu}(nN) = \beta_\nu \boldsymbol{\Phi}_{\nu\nu}((n-1)N) + (1-\beta_\nu) \widehat{\boldsymbol{\Phi}}_{\nu\nu}(nN), \tag{6}$$

with n as the block-index, β_ν as the forgetting factor and

$$\widehat{\boldsymbol{\Phi}}_{\nu\nu}(nN) = \frac{1}{\sum_{l=0}^{N-1} M_\nu(nN-l)} \sum_{l=0}^{N-1} M_\nu(nN-l)\mathbf{Y}(nN-l)\mathbf{Y}^H(nN-l). \quad (7)$$

In the offline (batch) case, $\boldsymbol{\Phi}_{\nu\nu}(nN)$ is estimated on the whole utterance, i.e., $\beta_\nu = 0$ and N is set to the number of frames in the utterance.

Equation (6) requires an initialization. The noise SCM is initialized either by assuming white noise and thereby a diagonal matrix or by estimating the SCM of diffuse noise:

$$\boldsymbol{\Phi}_{\text{diff}}(f) = \varphi_{\mathbf{N}} \cdot \text{sinc}\left(2\pi f \cdot \frac{F_{\max}}{F} \cdot \mathbf{d}/c\right), \quad (8)$$

where $\mathbf{d}$ is the matrix of distances between the microphones, c is the velocity of sound, $F_{\max}$ the Nyquist frequency, F the number of frequency bins, and $\varphi_{\mathbf{N}}$ is the noise power.

The target speech SCM may either be initialized using the RTF of the speaker position and the rank-1 approximation $\tilde{\boldsymbol{\Phi}}_{\mathbf{XX}} = \varphi_{\mathbf{X}}\tilde{\mathbf{H}}\tilde{\mathbf{H}}^{\mathsf{H}}$ with $\varphi_{\mathbf{X}}$ as the speech power, or using the SCM of the AU.

For comparison purposes, a second speech enhancement method is employed using non-adaptive beamforming. A set of MVDR beamforming coefficient vectors is precomputed, assuming concentrated sources at fixed, predefined positions and a diffuse noise field, as described in [22]. The predefined positions for the FixedBF are set in a circular form around the array with 10° distance, a radius of 1.5 m and 0.4 m height relative to the array, resulting in 36 positions. During the AU phase, an acoustic source localization is performed using the Steered Response Power - Normalized Arithmetic Mean (SRP-NAM) algorithm, as described in [23], and the beamforming vector corresponding to the estimated position is selected for source extraction. This method will be referred to as FixedBF.

2.2 Mask Estimation

In this section we describe the mask estimation required for SCM updates given in Eq. (6). It is a modified version of the SB source extraction network introduced in [11].

The neural network for mask estimation can be split in three parts: a recurrent neural network (RNN) layer, followed by an adaptation layer and a classification layer, consisting of two feed forward layers (FFs). In the adaptation layer one larger feed forward layer is split into several sub-layers. The outputs of these sub-layers are combined prior to the application of the non-linearity σ, using weights α:

$$h_k^{(\ell)} = \sigma\left(\sum_{j=1}^{N^{(\ell-1)}} h_j^{(\ell-1)} \sum_{m=1}^{M} \alpha_m W_{mjk}\right), \qquad k = 1, \ldots, N^{(\ell)} \quad (9)$$

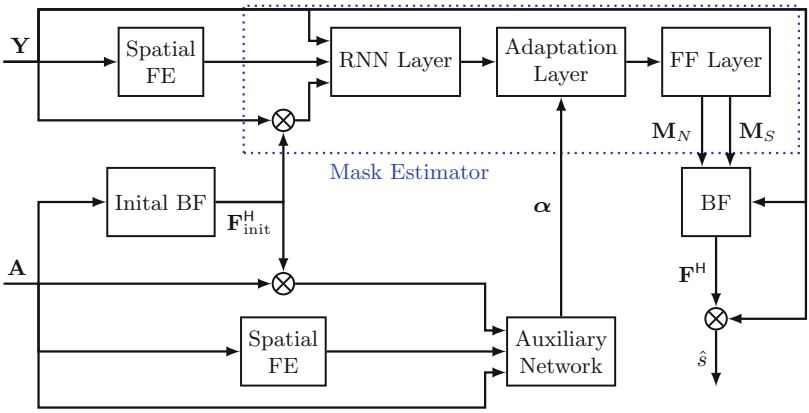

Fig. 1. System overview of the presented spatial speaker extractor.

where $h_j^{(\ell-1)}$ is the output of the j-th node in the preceding, $(\ell-1)$-st, layer, and $h_k^{(\ell)}$ the k-th node output in the ℓ-th layer. $N^{(\ell)}$ is the number of nodes in layer ℓ, W_{mjk} the learn-able weight matrix coefficients, where m indicates the sub-layer, and M the number of adaptation weights. Here, $\boldsymbol{\alpha} = [\alpha_1, ..., \alpha_M]^T$ is provided by an Auxiliary Network (AUX), to which the AU is used as input. This enables the mask estimator (ME) to focus on the speaker which was present during the AU.

The SB approach shows a degradation in performance when applied in a scenario with overlapping speakers with similar spectral characteristics as is observed in speakers of the same gender. To alleviate this problem spatial information is employed, assuming that the target speaker spoke the AU and his contribution to the speech mixture $\mathbf{Y}$ from the same position in the room. First, both the AU and the distorted signal $\mathbf{Y}$ are enhanced using a beamformer estimated from the SCM calculated on the AU as described above. Additionally, spatial features as described in [16] are extracted from both the AU and $\mathbf{Y}$:

$$\text{cosIPD}(t, f, p, q) = \cos\left(\angle y_{t,f,p} - \angle y_{t,f,q}\right), \tag{10}$$

$$\text{sinIPD}(t, f, p, q) = \sin\left(\angle y_{t,f,p} - \angle y_{t,f,q}\right), \tag{11}$$

where p, q are channel indices and $\angle$ is the phase operator. In the case of more than two channels all combinations of channel pairs are employed. However, at the output of the auxiliary network a mean pooling over the channel pairs is carried out to allow a more robust estimation in case of defective channels.

Furthermore, a beamformer is estimated on the AU. This beamformer, called "initial beamformer" in the following, is used to enhance the AU and the mixed speech to compute enhanced features.

To summarize, three sets of features are input to the AUX and mask estimation network: first, log-spectral features computed from the observed microphone signals, second, enhanced log-spectral features obtained after applying the initial beamformer to the microphone signals, and third, the aforementioned spatial features.

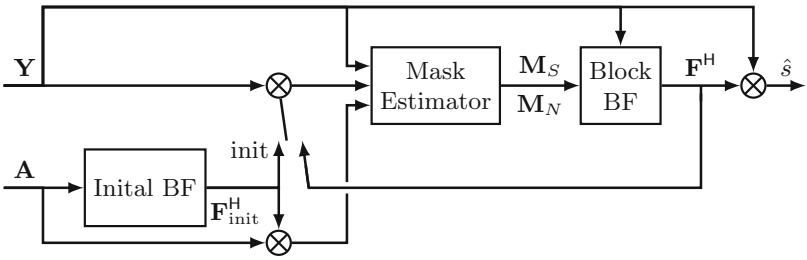

Fig. 2. System overview of the spatial speaker extractor reusing the estimated beamforming vector as initial beamformer for the next block of frames.

A block diagram of the presented system is depicted in Fig. 1.

Both the features computed from the initial beamformer and the spatial features computed on the AU are informative only under the assumption that both the speech of the target speaker in the speech mixture and the AU originate from the same position in the room. Therefore, a system dependent on these features will probably fail in a moving speaker scenario. However, the spatial information computed from the speech mixture can still be beneficial to extract the target speech, in particular if the competing speaker has similar spectral characteristics.

We propose to use a block-online recursive mask estimation system as depicted in Fig. 2. The initial beamformer estimated on the AU is used to enhance the first block of input frames which in turn are used to update the SCMs and estimate a new beamforming vector. This new beamforming vector then replaces the initial beamformer coefficients to compute the above mentioned set of enhanced features on the next block of frames. By this recursive update the enhanced feature set remains able to capture valid information in the presence of speaker movement or changes in the noise statistics.

3 Experiments

The presented systems are compared using four evaluation metrics: signal to distortion ratio (SDR) following the implementation presented in [24], an "invasive" SDR (InvSDR) [25], whereby the speech and the distortion are separately processed by the beamformer, and the SDR is computed as the power ratio of the resulting two outputs, the intelligibility measure STOI [26] and the perceptual speech quality metric PESQ [27]. All systems will be evaluated in terms of their gain compared to the signal at a reference microphone prior to the enhancement. Additionally, the systems are evaluated in terms of Word Error Rate (WER) of a subsequent Automatic Speech Recognition (ASR) system.

All signals are recorded or resampled with 8kHz. For the STFT computation, a 512-point FFT is used with a Hann window and an 75% overlap, resulting in 257 frequency bins for each time frame. The ME consists of an LSTM layer of 1024 units, two feedforward layers with 1024 units each and one output layer.

The first feedforward layer is split into 30 sub-layers for the SB approach. The auxiliary network has two feed-forward layers of 50 units each and an output layer of 30 units, as in [11]. Finally, for the block-online estimation we use a block size of $N = 5$ frames, corresponding to 80 ms.

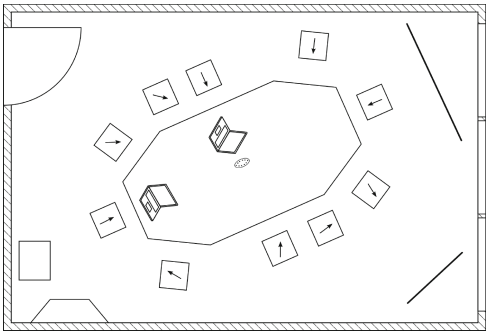

Fig. 3. Sketch of one of the meeting rooms the impulse responses and noises were recorded in. Room size approx. 4m × 6m. Drawn true to scale.

3.1 Database Description

We evaluate the proposed source extraction system on two databases. The first is the one described in [28], which consists of 30000 training, 500 development and 1500 evaluation examples. Each example is created by randomly choosing two utterances from the Wall Street Journal (WSJ) database and convolving the signals with six-channel room impulse responses (RIRs) with reverberation times $T_{60} \in [20\,\text{ms}, 500\,\text{ms}]$ simulated by the Image Methode [29]. The shorter of the generated multi-channel signals is padded with zeros to arbitrarily fall in the duration of the longer signal. The observation utterance then consists of the sum over both utterances, to which white Gaussian noise with an Signal to Noise Ratio (SNR) of 15 to 25 dB is added. The speaker sets of training, development and evaluation sets are mutually exclusive. Therefore, we characterize the database as open. For the AU we convolve a second utterance spoken by the target speaker with the same RIR and add white Gaussian noise. This database will be referred to as RirSim and is used for all parameter tuning and network training.

The second database is created similarly to the one described above, however the RIRs and the noise are replaced by real signals recorded in a conference scenario. The real RIRs and noises were recorded using a flat 8-channel Microelectromechanical systems (MEMS) microphone array, 7 cm × 10 cm in size and of elliptic shape. The recordings took place in two different meeting rooms with reverberation times of $T_{60} \approx 1$s at the premises of voice INTER connect GmbH in Dresden. Figure 3 shows the floor plan of one of these rooms. The microphone array was flush-mounted at the center of the meeting room table in both cases. The table height is 0.73 m. Impulse responses for ten different

Table 1. Gains of the beamformer output compared to the signal at a reference microphone w.r.t. different performance measures, and word error rate for different feature sets of the speaker extraction system on RirSim.

Method	Add. features		ΔSDR	ΔInvSDR	ΔSTOI	ΔPESQ	WER
	Enhanced	Spatial	dB	dB			%
Offline	–	–	6.48	6.49	0.10	0.26	32.66
	–	✓	9.54	9.36	0.14	0.40	29.43
	✓	–	10.16	10.22	0.16	0.46	27.32
	✓	✓	11.09	11.07	0.16	0.51	23.50
Online	✓	✓	7.57	9.00	0.15	0.41	30.61

lateral speaker positions per room were recorded using a coaxial loudspeaker at an assumed human speaker's mouth height of 1.15 m. The speaker positions for the depicted room, together with their directions of view, are shown as squares with arrows in Fig. 3. Four different types of typical meeting room noise sources (air-conditioning, paper shuffling, projector, typing noises) were recorded using the microphone array. The database thus created will be called RirReal.

3.2 ASR Backend

The Automatic Speech Recognition (ASR) backend used the wide residual network structure proposed in [30] with logarithmic mel filterbank input features and two Long-Short-Term-Memory (LSTM) layers. This acoustic model is combined with a trigram language model from the WSJ baseline script provided by the KALDI toolkit [31]. All hyper-parameters were taken from [30]. The same neural acoustic model, trained on the artificially reverberated WSJ utterances of RirSim, is used for both databases. The network is trained on alignments extracted with a HMM model trained in KALDI. The decoding is performed without language model rescoring.

3.3 Source Extraction in Static Speaker Scenario

In Table 1 the performance of different feature sets for the extraction systems described above are compared on the RirSim database. All systems use the log-spectral magnitude of the observation. As additional features we compare the log-spectral magnitude of the observation enhanced using an initial beamforming vector estimated on the AU, spatial features according to Eqs. (10) and (11), or both the spatial features and the enhanced signals. If the method is offline, both the beamforming vector and mask estimation are carried out in batch mode on the whole utterance.

All described features achieve better results than the original SpeakerBeam system, whose performance is given in the first results row of Table 1. Even the online system achieves better results using the additional features compared to the original offline SpeakerBeam system. Therefore, we conclude that using

Table 2. Gains of the beamformer output compared to the signal at a reference microphone w.r.t. different performance measures, and word error rate for non-stationary speaker on RirReal. Here Position (Pos.) 0 symbolizes the first speaker position which is equal to the position during the AU whereas Position 1 indicates a change in the position. "only ME" indicates that the additional spatial features are used as input to the mask estimation network only.

Method	Add. features		Pos.	ΔInvSDR	ΔSTOI	ΔPESQ	WER
	Enhanced	Spatial		dB			%
FixedBF	–	–	0	−1.72	0.03	−0.04	63.27
			1	−6.71	−0.08	−0.05	94.51
Offline	✓	✓	0	2.76	0.05	0.11	36.26
			1	0.12	−0.02	0.03	88.82
Online	✓	✓	0	3.93	0.07	0.13	34.79
			1	1.41	0.01	0.05	63.94
		only ME	0	3.38	0.06	0.13	35.34
			1	1.71	0.02	0.06	62.18
	$\mathbf{F}(\ell-1)$	only ME	0	3.43	0.05	0.11	35.12
			1	2.29	0.03	0.08	50.44

spatial information is beneficial for our source extraction system in case of static speakers. In [17] we present an in-depth evaluation of the described features in case of static speakers.

3.4 Source Extraction in the Presence of a Speaker Position Change

To simulate a change in speaker position, we divided the WSJ database in pairs of two utterances, where the first is convolved with the same set of RIRs as the AU and second is convolved with a different set of RIRs than the first, while keeping the competing speaker in the speech mixture and his/her position in the room fixed in both utterances.

The change of the target speaker position calls for adaptive beamforming. We thus expect the online beamformer to outperform the offline beamformer.

While the target speaker position in the first of the two utterances coincides with the one present in the AU, this no longer holds for the second. This renders the spatial information gained from the AUX incorrect. Table 2 displays the extraction results achieved with different features for online and offline systems. Note that neither the Acoustic Model (AM) nor the ME is retrained on the new RIR and noise.

Using spatial features during mask estimation but not in the AUX improves the extraction in case of changes in the target speaker position as can be seen in the entry with "only ME" in the column "spatial". Similarly, can be concluded that it is beneficial to update the initial beamforming vector for each block of frames, see the entry with $\mathbf{F}(\ell-1)$ under the column "enhanced".

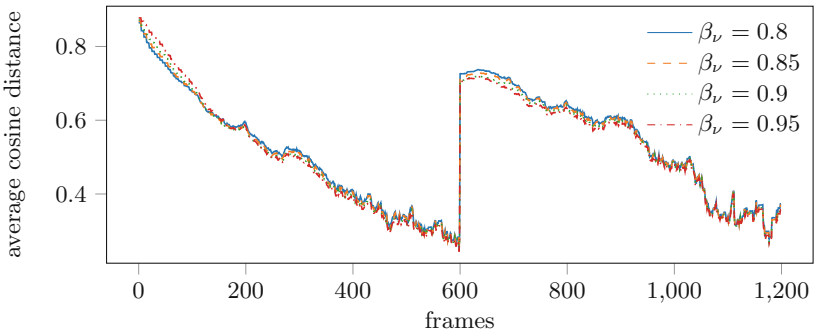

Fig. 4. Cosine distance between the block-online beamforming vector and an oracle offline beamforming vector calculated on the speech and noise image averaged over 500 utterances. Speaker positions changed at frame #600.

Additionally, the results confirm that the extraction achieved by a recursively updated beamforming vector is only slightly impeded by the change in speaker position, whereas a fixed beamformer estimated once for the concatenated utterances suffers significantly from changes in the speaker position. This is especially true for the fixed beamforming vector estimated on the AU since no information about the concurrent speaker is included in the noise SCM estimation.

To emphasize the benefits of recursive beamformer adaptation the cosine distance between the recursively estimated beamforming vector and an oracle offline beamformer is depicted in Fig. 4. Here, the coefficients of the offline beamformer have been obtained separately on the first and second utterance using the oracle speech and noise images at the microphones. The displayed tracking curves are averaged over multiple utterances.

The figure showcases the ability of the online beamforming vector to adapt to a change in speaker position. Furthermore, the recursive update displays an invariance concerning the forgetting factor β_ν

4 Conclusion

This paper offers a thorough investigation of speaker extraction systems guided by an AU in case of changes in the speaker position. We showcased the benefits of recursively updating a beamforming vector and investigated the usefulness of spatial features in case of target speaker position changes. While the spatial characteristics of the target speaker extracted from the adaptation utterance becomes outdated, the use of spatial features for mask estimation to extract a target speaker from a speech mixture remains beneficial. This can be attributed to the fact that they allow to separate speakers based on their spatial diversity, thus not relying solely on different spectro-temporal properties of the speakers.

Acknowledgements. The work was in part supported by DFG under contract number Ha3455/14-1. Computational resources were provided by the Paderborn Center for Parallel Computing.

References

1. Vincent, E., Watanabe, S., Nugraha, A.A., Barker, J., Marxer, R.: An analysis of environment, microphone and data simulation mismatches in robust speech recognition. Comput. Speech Lang. **46**, 535–557 (2017)
2. Sainath, T.N., Weiss, R.J., Wilson, K.W., Narayanan, A., Bacchiani, M.: Factored spatial and spectral multichannel raw waveform CLDNNs. In: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) (March 2016)
3. Wang, Y., Fan, X., Chen, I.F., Liu, Y., Chen, T., Hoffmeister, B.: End-to-end anchored speech recognition. CoRR abs/1902.02383 (2019)
4. Heymann, J., Drude, L., Chinaev, A., Haeb-Umbach, R.: BLSTM supported GEV beamformer front-end for the 3rd CHiME challenge. In: Proceedings Workshop Automatic Speech Recognition, Understanding, pp. 444–451 (2015)
5. Higuchi, T., Ito, N., Yoshioka, T., Nakatani, T.: Robust MVDR beamforming using time-frequency masks for online/offline ASR in noise. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (March 2016)
6. Yoshioka, T., Nakatani, T.: Generalization of multi-channel linear prediction methods for blind mimo impulse response shortening. IEEE Trans. Audio, Speech, Lang. Process. **20**(10), 2707–2720 (2012)
7. Yoshioka, T., Erdogan, H., Chen, Z., Xiao, X., Alleva, F.: Recognizing overlapped speech in meetings: a multichannel separation approach using neural networks. In: Interspeech (2018)
8. Luo, Y., Mesgarani, N.: Tasnet: surpassing ideal time-frequency masking for speech separation. CoRR abs/1809.07454 (2018)
9. Yu, D., Kolbaek, M., Tan, Z., Jensen, J.: Permutation invariant training of deep models for speaker-independent multi-talker speech separation. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 241–245 (March 2017)
10. Chen, Z., Luo, Y., Mesgarani, N.: Deep attractor network for single-microphone speaker separation. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 246–250 (March 2017)
11. Zmolíková, K., Delcroix, M., Kinoshita, K., Higuchi, T., Ogawa, A., Nakatani, T.: Speaker-aware neural network based beamformer for speaker extraction in speech mixtures. Proc. Interspeech **2017**, 2655–2659 (2017)
12. Kida, Y., Tran, D., Omachi, M., Taniguchi, T., Fujita, Y.: Speaker selective beamformer with keyword mask estimation. In: 2018 IEEE Spoken Language Technology Workshop (SLT), pp. 528–534 (Dec 2018)
13. Wang, Q., et al.: Voicefilter: targeted voice separation by speaker-conditioned spectrogram masking. arXiv e-prints arXiv:1810.04826 (2018)
14. Chen, Z., Xiao, X., Yoshioka, T., Erdogan, H., Li, J., Gong, Y.: Multi-channel overlapped speech recognition with location guided speech extraction network. In: 2018 IEEE Spoken Language Technology Workshop (SLT), pp. 558–565 (Dec 2018)
15. Liu, Y., Ganguly, A., Kamath, K., Kristjansson, T.: Neural network based time-frequency masking and steering vector estimation for two-channel MVDR beamforming. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (April 2018)

16. Wang, Z., Le Roux, J., Hershey, J.R.: Multi-channel deep clustering: discriminative spectral and spatial embeddings for speaker-independent speech separation. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5 (April 2018)
17. Martín-Doñas, J.M., Heitkaemper, J., Haeb-Umbach, R., Gomez, A.M., Peinad, A.M.: Multi-channel block-online source extraction based on utterance adaptation. In: 20th Annual Conference of the International Speech Communication Association. Graz, Austria (September 2019)
18. Gannot, S., Vincent, E., Markovich-Golan, S., Ozerov, A.: A consolidated perspective on multi-microphone speech enhancement and source separation. IEEE/ACM Trans. Audio, Speech, Lang. Process. **PP(99)**, 1 (2017)
19. Souden, M., Benesty, J., Affes, S.: On optimal frequency-domain multichannel linear filtering for noise reduction. IEEE Trans. Audio, Speech, Lang. Process. **18**(2), 260–276 (2007)
20. Wang, Z., Vincent, E., Serizel, R., Yan, Y.: Rank-1 constrained multichannel wiener filter for speech recognition in noisy environments. Comput. Speech Lang. **49**, 37–51 (2018)
21. Heitkaemper, J., Heymann, J., Haeb-Umbach, R.: Smoothing along frequency in online neural network supported acoustic beamforming. In: ITG 2018, Oldenburg, Germany (October 2018)
22. Fehér, T., Freitag, M., Gruber, C.: Real-time audio signal enhancement for hands-free speech applications. In: 16th Annual Conference of the International Speech Communication Association, pp. 1246–1250. Dresden, Germany (September 2015)
23. Salvati, D., Drioli, C., Foresti, G.L.: Incoherent frequency fusion for broadband steered response power algorithms in noisy environments. IEEE Signal Process. Lett. **21**(5), 581–585 (2014)
24. Raffel, C., et al.: mir_eval: a transparent implementation of common MIR metrics. In: Proceedings of the 15th International Society for Music Information Retrieval Conference, ISMIR (2014)
25. Tran Vu, D.H., Haeb-Umbach, R.: Blind speech separation employing directional statistics in an expectation maximization framework. In: 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 241–244 (March 2010)
26. Taal, C.H., Hendriks, R.C., Heusdens, R., Jensen, J.: An algorithm for intelligibility prediction of time-frequency weighted noisy speech. IEEE Trans. Audio, Speech, Lang. Process. **19**(7), 2125–2136 (2011)
27. Rix, A.W., Beerends, J.G., Hollier, M.P., Hekstra, A.P.: Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In: 2001 Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing. (Cat. No.01CH37221). vol. 2, pp. 749–752 (2001)
28. Drude, L., Haeb-Umbach, R.: Integration of neural networks and probabilistic spatial models for acoustic blind source separation. In: IEEE Journal of Selected Topics in Signal Processing (2018)
29. Allen, J.B., Berkley, D.A.: Image method for efficiently simulating small-room acoustics. J. Acoust. Soc. Am. **65**(4), 943–950 (1979)
30. Heymann, J., Drude, L., Haeb-Umbach, R.: Wide residual BLSTM network with discriminative speaker adaptation for robust speech recognition. In: CHiME4 Workshop (2016)
31. Povey, D., et al.: The kaldi speech recognition toolkit. In: IEEE 2011 Workshop on Automatic Speech Recognition and Understanding. No. Idiap-RR-04-2012, IEEE Signal Processing Society, Rue Marconi 19, Martigny (Dec 2011)