Warping: Data-driven Mixture Preprocessing to Boost the Performance of Blind Speech Separation

Jiri Malek, Zbynek Koldovsky and Jaroslav Cmejla
Technical University in Liberec,
Faculty of Mechatronics, Informatics and Interdisciplinary Studies,
Email: jiri.malek@tul.cz

***Abstract*—Blind source separation (BSS) can be used to recover speech signals from mixtures recorded by microphones. However, their performances show significant limitations because of the deviations between the instantaneous mixing model and real audio mixtures transformed into the short-time Fourier domain (STFT). This paper presents a data-driven preprocessing technique called *mixture warping*, which aims to adjust the mixture to obey the instantaneous model as much as possible. As a proof of concept, we demonstrate its effect on a set of reverberant mixtures of two speakers. Warping implemented through a deep neural network is trained to estimate mixtures ideally modified towards the instantaneous model in the least-squares sense. By applying it as a preprocessing stage, it boosts the BSS performance by up to 6.4 dB of signal-to-interference (SIR) and 4.7 dB of signal-to-distortion (SDR) on average without requiring any modification to the BSS methods.**

I. Introduction

Speech originating in a real-world environment is often measured along with unwanted noises and competing utterances. One way to remove the unwanted components is the application of separation or extraction techniques. Separation [1] attempts to estimate each active sound source, while extraction aims to recover one target source only [2]. This work addresses methods using multiple microphones.

There are two significant branches of separation and extraction. *Data-driven techniques* [3]–[6] exploit a large amount of training data and achieve high separation quality, provided that this data closely match the considered scenario. *Model-based methods* [1], [7] are based on physical and information-theoretical assumptions about the input mixtures and, where possible, usage of reference information. Newly emerging *hybrid methods* [4], [8] aim to combine the complementary strengths of data and model-based approaches.

This work is focused on the use of blind methods (BSS/BSE) [7], [9]–[12], which are model-based techniques that require minimum prior knowledge and no training data. In particular, Independent Component Analysis (ICA) [13] and Independent Component Extraction (ICE) [14], as well as their extensions to vector components IVA [15] and IVE [14], assume only linearity of mixtures and statistical independence of original sources. Due to the low requirements, their applicability is far-reaching. However, significant performance limitations occur if the mixing models do not sufficiently match reality.

The most widely used BSS/BSE methods assume the determined instantaneous mixing model where the number of original sources corresponds to the number of inputs [13]. However, the real-world recorded audio mixtures are convolutive due to the finite speed of sound propagation and reverberation [16]. To enable IVA/IVE processing, the short-term Fourier transform (STFT) is applied. Each frequency band is then processed assuming the instantaneous model, and the bands are processed jointly through IVA/IVE [15]. In other words, convolution is *approximated* by multiplication in the frequency domain, which has been referred to as the multiplicative transfer function (MTF) approximation [17].

This MTF model is accurate when the acoustic paths between sources and microphones are short enough and the targeted speakers behave like point sources. However, these assumptions often do not match the real acoustic environment, where the reverberation time is long, and the speaker is usually slightly moving. These effects cause the instantaneous model to capture the reality insufficiently [16]. With data precisely obeying the mixing model, blind methods can achieve strong interference suppression, e.g., by 20 or even 30 dB; see, e.g., [18]. They also show the equivariance property, meaning their accuracy does not depend on mixing parameters [19]. However, they do not show such strong capabilities on real audio mixtures, which leaves significant room for improvement.

The convolutive transfer function (CTF) mixing model [17] has been proposed as the remedy for the inaccuracy caused by MTF. It allows long RIRs to be approximated within short STFT frames. In the context of BSS/BSE [20], [21], it is assumed that the past frames of the source signals (corresponding to the reverberant part of the RIR) are additional sources. However, this approach increases the complexity and brings additional uncertainties.

In this paper, we explore a novel data-driven approach that addresses the problem of deviations of real audio mixtures from the instantaneous model in the STFT domain, an approach that we refer to as *mixture warping*. The method transforms the multi-channel STFT spectrogram to adhere to the instantaneous mixing in the least square sense. It is implemented through a neural network trained in a supervised manner using artificially augmented acoustic mixtures. We demonstrate the efficiency of this concept by applying it to highly reverberant mixtures of two speakers. The results are promising in that the mixture warping can significantly boost the performance of BSS/BSE without interfering with these methods and increasing their computational complexity.

II. Problem description

The multi-channel mixture of D static point sources in a reverberant environment can be described in the time domain by the *convolutional mixing* model

$$x_m(n) = \sum_{d=1}^{D} \{a_{m,d} * s_d\}(n), \tag{1}$$

where $x_m(n)$ denotes the mth channel of the mixture, $m = 1 \dots M$, s_d represent the dth acoustic source, $d = 1 \dots D$, and $a_{m,d}$ is the room impulse response (RIR) between the mth sensor and dth source. The signals transformed by the STFT are often approximated by the instantaneous model

$$\mathbf{x}(k, \ell) = \mathbf{A}(k)\mathbf{s}(k, \ell), \tag{2}$$

where $k = 1 \dots K$ is the frequency index, $\ell = 1 \dots L$ is the frame index, $\mathbf{x}$ denotes the $M \times 1$ vector of the mixed signals, $\mathbf{s}$ is the $D \times 1$ vector of all sources, and $\mathbf{A}$ is the $M \times D$ mixing matrix whose elements corresponds to the MTFs. The determined model corresponds to the assumption that $D = M$, which makes $\mathbf{A}$ square and invertible.

The MTF approximation is justified for a sufficiently long analysis frame, where the length of the STFT window is significantly longer compared to the lengths of the RIRs $a_{m,d}$. If this requirement is not met, (2) is not sufficiently accurate, which deteriorates the BSS/BSE performance. Unfortunately, typical RIRs can have thousands of taps, even in mildly reverberated rooms. The effective length of a RIR also significantly grows with the source-microphone distance. Given how effective BSS methods can be if (2) is sufficiently accurate and how they lose performance if the opposite is true, we are motivated to find a transformation that would be capable of correcting the inaccuracy in (2).

III. Mixture warping

For the sake of simplicity, let us first consider a mixture of two sources recorded by two microphones. Then, (2) can be rewritten as

$$x_1(k, \ell) = s_1^1(k, \ell) + s_2^1(k, \ell), \tag{3}$$

$$x_2(k, \ell) = H_1(k) \cdot s_1^1(k, \ell) + H_2(k) \cdot s_2^1(k, \ell), \tag{4}$$

where $s_1^1(k, \ell) = A_{1,1}(k)s_1(k, \ell)$ and $s_2^1(k, \ell) = A_{1,2}(k)s_2(k, \ell)$ are the images of the first/second source on the first microphone, $H_1(k) = \frac{A_{2,1}(k)}{A_{1,1}(k)}$ and $H_2(k) = \frac{A_{2,2}(k)}{A_{1,2}(k)}$ are relative transfer functions between source images recorded on the first (reference) and second microphone, respectively, and $A_{i,j}(k)$ denotes the ijth element of $\mathbf{A}(k)$.

The inaccuracies of the instantaneous model (2) can be interpreted as implying that while (3) is true, (4) is inaccurate. We therefore propose to substitute $x_2(k, \ell)$ by

$$y_2(k, \ell) = G_1(k) \cdot s_1^1(k, \ell) + G_2(k) \cdot s_2^1(k, \ell), \tag{5}$$

where $G_1(k)$ and $G_2(k)$ are the solutions of

$$\underset{G_1(k), G_2(k)}{\arg\min} \; \|x_2(k, \ell) - G_1(k) \cdot s_1^1(k, \ell) - G_2(k) \cdot s_2^1(k, \ell))\|_2. \tag{6}$$

We call the transformation from $\mathbf{x}(k, \ell) = \binom{x_1(k,\ell)}{x_2(k,\ell)}$ to $\tilde{\mathbf{x}}(k, \ell) = \binom{x_1(k,\ell)}{y_2(k,\ell)}$ as *mixture warping*. The ideally transformed mixture $\tilde{\mathbf{x}}(k, \ell)$ is similar to the original one, while it perfectly obeys the instantaneous mixing model, which satisfies our goal. However, the ideal transformation is a function of the true source images $s_1^1(k, \ell)$ and $s_2^1(k, \ell)$, which we aim to separate/extract. Our strategy is to approximate the transformation by a deep neural network and train it to estimate $y_2(k, \ell)$ given the input $x_1(k, \ell)$ and $x_2(k, \ell)$, with the training target $y_2(k, \ell)$ computed according to (5),(6). The approach can be extended to any number of microphones by applying the transform to each pair of signals $x_1(k, \ell)$ and $x_m(k, \ell)$, $m = 2 \dots M$.

A. The warping network

The proposed warping network is a variant of U-net [22] depicted in Fig. 1. We use complex-valued parameters; the network is implemented in Pytorch with partial support of the ComplexNN toolbox [23].

The input consists of 19 frames (about 600 ms) of two-channel input mixture, i.e., the reference channel $x_1(k, \ell)$ and the to-be warped channel $x_2(k, \ell)$. The context is non-causal, containing 9 frames before and 9 frames after the current frame. Our preliminary experiments found advantageous to input frames with two frequency resolutions: here $K_1 = 1024$ and $K_2 = 4096$ bins. The K_1 corresponds to the desired output resolution. The motivation for inclusion of K_2 is that the MTF approximation becomes more accurate with a longer frame and the network is expected to learn from this information.

Most of the benefits of mixture warping remain when applied only in the frequency bins, which concentrate most of the signal power. This significantly reduces the number of trainable parameters and, consequently, the training time. To demonstrate this, we present two variants of the warping network: 1) *The full-band variant* (FB) depicted in Fig. 1, where the warping is applied to the complete frequency range for both input frequency resolutions. The FB variant requires 2.1 millions of trainable parameters. 2) *The sub-band variant* (SB) optimized for speech signals, where the warping is applied only to frequencies $0 \dots 2000$ Hz, i.e., to bins $0 \dots 128$ for resolution K_1 and $0 \dots 512$ for resolution K_2. The SB variant requires 831 thousands of trainable parameters. For the SB variant, the number of frequency bins (dimension 1 of tensors in Fig. 1) is reduced by 4.

The first block of layers realizes the merging of the two input frequency resolutions. The features of the larger one are two times sub-sampled by a convolutional layer with frequency stride 2. Moreover, the time-context is consecutively reduced by omitting zero-padding in the frame dimension of the convolutional kernel. At the output of this block, the two sets of feature maps are concatenated and have time-context equal to 1.

The following block of layers represents the encoder part of the network. Two types of layers alternate here: a sub-sampling layer that reduces the frequency resolution via kernel stride 2

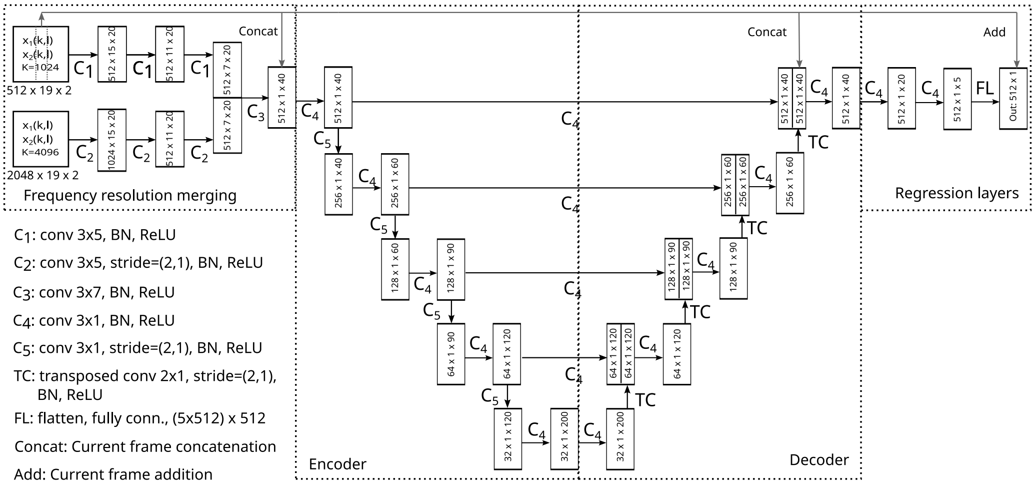

Fig. 1. Warping network architecture (the full-band variant): Each arrow represents a layer or an operation, each rectangle indicates the resulting dimensions of the data (frequency resolution × number of frames × number of feature maps). All data concatenations perform a merging of feature maps from different network branches. For the sub-band variant, the first dimension of the data, corresponding to the frequency range, is always four times smaller.

and a classical convolutional layer. The frequency resolution is reduced four times.

The subsequent layers implement decoding. The frequency resolution is consecutively increased back to the original size via transposed convolution with the stride of two. The encoder and the decoder layers are interconnected with skip connections containing the convolutional layer, as suggested in [24]. This prevents the concatenation of semantically different features between the encoder and decoder.

The final block of layers performs the regression using the features computed by the decoder. The network output estimates the warped frame $y_2(k,\ell)$, obtained by a fully connected layer without a nonlinearity.

All convolutional layers in the network are followed by batch normalization (BN) and ReLU nonlinearity. A skip connection using the input frame is concatenated with the feature maps at the output of the resolution merging block and the decoder. The network is trained by minimization of the mean square error loss function.

B. Dataset

The training of the network and the experimental evaluation are performed using reverberated mixtures of two speakers originating in the multi-channel Wall Street Journal dataset (MC-WSJ0-2mix [25], [26]). These are spatialized versions of the single-channel dataset described in [27]. The dataset contains $20,000$ training, 5000 validation, and $3,000$ test mixtures simulated in a reverberant environment using a microphone array containing eight microphones; our experiments utilize the first two or four microphones. The sources are mixed at SIR between $\langle -5, +5\rangle$ dB. The recordings are highly reverberant ($T_{60} \in \langle 200, 600\rangle$ ms) and captured in rooms with variable dimensions. The geometry of the microphone array is varying, as well as the source-microphone distance, which is 1.3 m with 0.4 m standard deviation. The sampling frequency is 16 kHz.

True source components are available for the training mixtures. Using these, the ideal warped mixtures are computed using equations (5) and (6) and used to train the warping network. The first channel of WSJ0-2mix serves as the reference channel $x_1(k,\ell)$. Other channels are used as $x_2(k,\ell)$, i.e., one four-channel recording produces three warped two-channel mixtures. In this way, $3 \times 20,000 = 60,000$ training mixtures are created.

IV. Experiments

The mixture warping can improve the performance of either blind source separation (BSS) or blind source extraction (BSE). To demonstrate both, we present results achieved by AuxIVA (Auxiliary function Independent Vector Analysis, [9]), which is a BSS algorithm, and AuxIVE (Auxiliary function Independent Vector Extraction [28]), a BSE algorithm. Both algorithms run for 200 iterations per mixture. AuxIVE requires guidance to extract the target speaker through external information; we utilize embedding features encoding the speaker's characteristics. The embeddings are computed using a pre-trained neural network; our implementation uses the feed-forward sequential memory network (FSMN, [29]); details are provided in [30].

Three variants of the test mixtures in the time-frequency domain (with STFT frame length of 1024 and the shift of 512 samples) are processed. 1) *Original mixture* contains unmodified channels $x_i(k,\ell), i = 1\ldots4$. 2) *Ideal mixture* is computed by (5) and contains $x_1(k,\ell)$ and ideally warped $y_2(k,\ell)$. 3) *Warped mixture* contains $x_1(k,\ell)$ and one or three additional channels estimated by the warping network, either by the full-band (FB) or the sub-band (SB) variant. Note that the ideal warping is applicable only for the two microphone settings because the mixtures contain only two sources. With a larger number of microphones, the ideally warped mixtures would be rank deficient (overdetermined), which is an artificial situation that does not occur in practice.

The experiments are evaluated using the BSS_EVAL toolbox [31]. The presented measures are SIR, which quantifies suppression of the unwanted sources, and SDR, which measures both the suppression and the distortion of the desired source. Table I[1] summarizes the results achieved by AuxIVA (separation) and AuxIVE (extraction) under various settings.

A. Separation by AuxIVA

For 2 input channels, the performance on mixtures preprocessed with the FB warping is improved by 5.5 dB SIR and 3.8 dB SDR on average compared to that achieved on the original mixtures. Detailed distributions of the results are presented in Fig. 2 where the results obtained on the warped mixtures are distinctly shifted towards higher values for both SIR and SDR. With 4 microphones, the FB warping is even more beneficial than with 2 microphones, bringing 6.4 dB SIR and 4.7 dB SDR improvements on average by AuxIVA.

When comparing the SB and FB warping variants, the SB warping yields lower results by up to 1dB in both observed metrics. Therefore, the sub-band variant appears to be an economical alternative to FB, as its slight performance loss results in a 60% reduction of the network's number of parameters.

The results achieved with the ideal mixtures show a significant improvement of 17.9 dB SIR and 16.7 dB SDR on average compared to the original mixtures. This points to the large potential for warping and considerable room for further improvements.

B. Extraction by AuxIVE

Similarly to separation, warping appears to be beneficial as it improves SIR and SDR by 5.2 dB SIR and 3.8 dB SDR, on average, compared to the performance achieved with original mixtures. With 4 microphones, warping improves the extraction accuracy only by 0.9 dB SIR and 1.0 SDR.

Overall, the extraction by AuxIVE shows slightly lower performance than the separation by AuxIVA. There are two explanations for this. First, extraction fails when a source different from the target speaker is extracted. This does not happen in case of separation because the correct output channel containing the target speaker is selected automatically. Second, the significant difference in results with 4 microphones is because AuxIVE does not implement dimension reduction, while AuxIVA does. Dimension reduction may be necessary for our experiments precisely because the mixtures contain only 2 speakers, which may be even more important for warped mixtures than for the original mixtures. Answering these questions requires further testing, where AuxIVE will also be implemented with dimension reduction.

When the extraction accuracies of the SB and FB warping applied to 2 microphones are compared, the SB warping achieves results lower by up to 2 dB SIR and 1.3 dB SDR. Therefore, the computationally less demanding SB variant still represents an effective alternative to FB as a way to improve BSE performance.

TABLE I
MC-WSJ0-2MIX: SIR/SDR [DB] AVERAGED OVER ALL 3000 SEPARATION (OR 6000 EXTRACTION) EXPERIMENTS USING 2 OR 4 MICROPHONES.

Task	**Mic. #**	**Mixture**	**SIR [dB]**	**SDR [dB]**
-	-	Original	0.1	0.1
Separation	2	Original	8.0	3.8
	2	Warped(SB)	12.8	7.3
	2	Warped(FB)	13.5	7.6
	2	Ideal	25.9	20.5
	4	Original	9.5	4.7
	4	Warped(SB)	15.0	8.8
	4	Warped(FB)	15.9	9.4
	4	Ideal	N/A	N/A
Extraction	2	Original	4.4	1.6
	2	Warped(SB)	7.6	4.1
	2	Warped(FB)	9.6	5.4
	2	Ideal	16.3	11.7
	4	Original	10.0	3.2
	4	Warped(SB)	10.3	3.8
	4	Warped(FB)	10.9	4.2
	4	Ideal	N/A	N/A

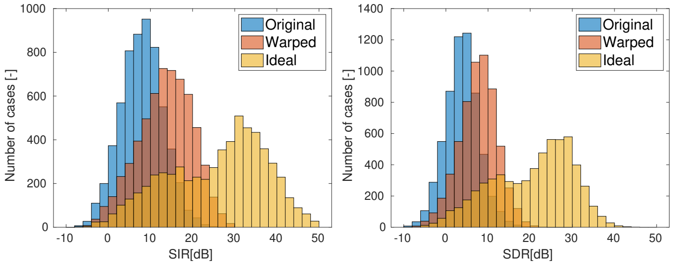

Fig. 2. MC-WSJ0-2mix: Separation using 2 channels - Histograms of SIR and SDR values achieved on mixtures with/without the FB warping.

V. Conclusion

A data-driven preprocessing for BSS in the STFT domain was proposed. The procedure is called mixture warping and remedies deviations of audio mixtures from the instantaneous mixing model. As a proof of concept, we apply BSS to ideally warped reverberant mixtures of two speech sources. The resulting performance is very high: compared to the unmodified mixtures, it yields a performance gain of 18dB SIR and 17dB SDR. Applying the trained warping network to the mixtures and applying BSS yields performance gains of up to 6.4dB SIR and 4.7 SDR, demonstrating the effectiveness of the approach and also pointing to room for possible improvements.

There are several open problems worth exploring in the future. Warping should be trained and analyzed for mixtures containing a varying number of active speakers, other sources, and environmental noise. Also, the sequence of warping+BSS should be compared to end-to-end data-driven systems in terms of data requirements, performance, and robustness to training-test mismatch (off-domain test data). Results show that a small network is sufficient to perform successful warping, which may indicate less data/computationally demanding training of warping. The presence of BSS has the potential to be less vulnerable to mismatching training-test conditions.

[1] Different values of SIR/SDR were presented for AuxIVE in [30] when applied to WSJ0-2mix dataset. The difference is caused by a different input sampling frequency; the 16 kHz version of the dataset is processed here.

Acknowledgment

The authors were supported by a project No. 25-18485S by the Czech Science Foundation (GAČR).

References

[1] Emmanuel Vincent, Tuomas Virtanen, and Sharon Gannot, *Audio Source Separation and Speech Enhancement*, Wiley Publishing, 2018.

[2] Katerina Zmolikova, Marc Delcroix, Tsubasa Ochiai, Keisuke Kinoshita, Jan Černockỳ, and Dong Yu, "Neural target speech extraction: An overview," *IEEE Signal Processing Magazine*, vol. 40, no. 3, pp. 8–29, 2023.

[3] DeLiang Wang and Jitong Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.

[4] Reinhold Haeb-Umbach, Tomohiro Nakatani, Marc Delcroix, Christoph Boeddeker, and Tsubasa Ochiai, "Microphone array signal processing and deep learning for speech enhancement: Combining model-based and data-driven approaches to parameter estimation and filtering," *IEEE Signal Processing Magazine*, vol. 41, no. 6, pp. 12–23, 2025.

[5] Christoph Boeddeker et al., "Convolutive transfer function invariant sdr training criteria for multi-channel reverberant speech separation," in *ICASSP 2021*. IEEE, 2021, pp. 8428–8432.

[6] Rongzhi Gu, Shi-Xiong Zhang, Yuexian Zou, and Dong Yu, "Towards unified all-neural beamforming for time and frequency domain speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 849–862, 2022.

[7] Sharon Gannot, Emmanuel Vincent, Shmulik Markovich-Golan, and Alexey Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 692–730, 2017.

[8] Nir Shlezinger, Jay Whang, Yonina C Eldar, and Alexandros G Dimakis, "Model-based deep learning," *Proceedings of the IEEE*, vol. 111, no. 5, pp. 465–499, 2023.

[9] Nobutaka Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *WASPAA 2011*, 2011, pp. 189–192.

[10] Lele Liao, Guoliang Cheng, Kai Chen, Zhanzhong Cao, and Jing Lu, "Improvement of independent vector analysis for closely spaced sources," *Applied Acoustics*, vol. 212, pp. 109575, 2023.

[11] Daichi Kitamura and Kohei Yatabe, "Consistent independent low-rank matrix analysis for determined blind source separation," *EURASIP Journal on Advances in Signal Processing*, vol. 2020, no. 1, pp. 1–35, 2020.

[12] Kouhei Sekiguchi et al., "Autoregressive fast multichannel nonnegative matrix factorization for joint blind source separation and dereverberation," in *ICASSP 2021*. IEEE, 2021, pp. 511–515.

[13] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, John Wiley & Sons, 2001.

[14] Z. Koldovsky and P. Tichavsky, "Gradient algorithms for complex non-gaussian independent component/vector extraction, question of convergence," *IEEE Transactions on Signal Processing*, vol. 67, no. 4, pp. 1050–1064, Feb 2019.

[15] T. Kim, H. T. Attias, S.-Y. Lee, and T.-W. Lee, "Blind source separation exploiting higher-order frequency dependencies," pp. 70–79, Jan. 2007.

[16] Shoko Araki, Ryo Mukai, Shoji Makino, Tsuyoki Nishikawa, and Hiroshi Saruwatari, "The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 2, pp. 109–116, 2003.

[17] 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.

[18] Z. Koldovský, V. Kautský, and P. Tichavský, "Double nonstationarity: Blind extraction of independent nonstationary vector/component from nonstationary mixtures—Algorithms," *IEEE Transactions on Signal Processing*, vol. 70, pp. 5102–5116, 2022.

[19] J. F. Cardoso and B. H. Laheld, "Equivariant adaptive source separation," vol. 44, no. 12, pp. 3017–3030, Dec 1996.

[20] Taihui Wang, Feiran Yang, and Jun Yang, "Convolutive transfer function-based multichannel nonnegative matrix factorization for overdetermined blind source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 802–815, 2022.

[21] Taihui Wang, Feiran Yang, Nan Li, Chen Zhang, and Jun Yang, "Low-latency real-time independent vector analysis using convolutive transfer function," *Applied Acoustics*, vol. 197, pp. 108931, 2022.

[22] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention (MICCAI 2015)*. Springer, 2015, pp. 234–241.

[23] Xinyuan Liao, "Complexnn: Complex neural network modules," https://github.com/XinyuanLiao/ComplexNN, Accessed: 7.1.2025.

[24] Nabil Ibtehaz and M Sohel Rahman, "Multiresunet: Rethinking the u-net architecture for multimodal biomedical image segmentation," *Neural networks*, vol. 121, pp. 74–87, 2020.

[25] Zhong-Qiu Wang, Jonathan Le Roux, and John R Hershey, "Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation," in *ICASSP 2018*. IEEE, 2018, pp. 1–5.

[26] "Scripts to generate the wsj0-mix multi-speaker dataset [online]," https://www.merl.com/demos/deep-clustering, Accessed: 18.12.2024.

[27] John R Hershey, Zhuo Chen, Jonathan Le Roux, and Shinji Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *ICASSP 2016*. IEEE, 2016, pp. 31–35.

[28] Jakub Janský, Zbyněk Koldovský, Jiří Málek, Tomáš Kounovský, and Jaroslav Čmejla, "Auxiliary function-based algorithm for blind extraction of a moving speaker," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2022, no. 1, pp. 1–16, 2022.

[29] Shiliang Zhang, Cong Liu, Hui Jiang, Si Wei, Lirong Dai, and Yu Hu, "Feedforward sequential memory networks: A new structure to learn long-term dependency," *arXiv preprint arXiv:1512.08301*, 2015.

[30] Jiri Malek, Jaroslav Cmejla, and Zbynek Koldovsky, "Blind extraction of target speech source: Three ways of guidance exploiting supervised speaker embeddings," in *2022 International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE, 2022, pp. 1–5.

[31] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE transactions on audio, speech, and language processing*, vol. 14, no. 4, pp. 1462–1469, July 2006.