INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Cancellation of Local Competing Speaker with Near-field Localization for Distributed Ad-Hoc Sensor Network

Pablo Pérez Zarazaga[1], *Mariem Bouafif Mansali*[2], *Tom Bäckström*[1], *Zied Lachiri*[2]

[1]*Department of Signal Processing and Acoustics, Aalto University, Espoo, Finland*
[2]*SITI Lab, National Engineering School of Tunis, El Manar University, Tunis, Tunisia*

pablo.perezzarazaga;tom.backstrom@aalto.fi, mariem.bouafif@gmail.com, zied.lachiri@enit.utm.tn

Abstract

In scenarios such as remote work, open offices and call centers, multiple people may simultaneously have independent spoken interactions with their devices in the same room. The speech of competing speakers will however be picked up by all microphones, both reducing the quality of audio and exposing speakers to breaches in privacy. We propose a cooperative cross-talk cancellation solution breaking the single active speaker assumption employed by most telecommunication systems. The proposed method applies source separation on the microphone signals of independent devices, to extract the dominant speaker in each device. It is realized using a localization estimator based on a deep neural network, followed by a time-frequency mask to separate the target speech from the interfering one at each time-frequency unit referring to its orientation. By experimental evaluation, we confirm that the proposed method effectively reduces crosstalk and exceeds the baseline expectation maximization method by 10 dB in terms of interference rejection. This performance makes the proposed method a viable solution for cross-talk cancellation in near-field conditions, thus protecting the privacy of external speakers in the same acoustic space.
Index Terms: Crosstalk Cancellation, Source Separation, Localization, Acoustic Sensor Network

1. Introduction

The actual state of telecommunication technologies leaves something to be desired; when multiple speakers hold different conversations in the same space, interference inevitably leaks into each other's transmissions, which is not only a cause of quality degradation but can violate the users' privacy. This effect is known as cross-talk [1, 2] or microphone leakage [3] and requires multi-channel speaker interference rejection to extract a target source from an observed mixture. This task is known as cross-talk cancellation (CTC) and it is getting much attention not only to reduce the noise but also to protect the privacy of external speakers.

For the CTC task, a multi-channel meeting scenario was designed towards an ad-hoc sensor network [4, 2] where we only have access to the sensor observations of all other users, which are as well corrupted by crosstalk. In this context, adaptive filtering in time and frequency domain has been adopted to estimate the room impulse responses (RIRs) from the interfering source's microphone to the target microphone [5, 6, 7]. However, it requires a speaker activity detector or prior knowledge of the signal-to-interference ratio (SIR) to control the adaptation process. Other methods address a non-negative signal factorization [8], kernel additive modeling [2], or a Gaussian probabilistic framework [9]. CTC is also tackled as a multi-channel denoising task using a multi-channel Wiener filter [3], focusing only on the competing speaker's energy. An extended version using a Kalman-based Wiener filter [10] improves the interference rejection by considering a prior estimation of the RIRs [11]. However, CTC performance still depends on the RIR estimation accuracy. Other studies address the CTC issue in a typical meeting scenario as a blind source separation (BSS) task [12, 13]. On the other side, CTC has been also studied in the binaural context as a BSS task [14, 15] by localization-based approaches to reduce the effect of competing signals coming from estimated directions. These localization-based BSS approaches are not focusing on close-talk and near-field scenarios since they are potentially limited by the features used for localization, which are usually adapted to the far-field context.

Unlike already established filtering-based CTC methods depending on prior knowledge, we study the reliability of adapting a localization-based BSS technique for the CTC task in a near-field telecommunication scenario detailed in Sec. 2. We propose a two-channel CTC algorithm based on masking inferred from the speakers' orientations. As described in Sec. 3, we make use of a deep neural network (DNN)-based architecture to infer the orientation of each time-frequency (TF) bin in different frequency blocks in the spectrum of the observed mixture, assuming the W-disjoint orthogonality property of speech signals [16]. Therefore, CTC is iteratively obtained for each device by multiplying the observed mixture by the mask associated with the dominant direction of arrival (DOA) in its dedicated microphone channel. Comparing our system with a state-of-the art method in a realistic and challenging scenario, our experiments in Sec. 4 show that better interference rejection is obtained, making our framework a good candidate for privacy preservation in telecommunication applications by attenuating competing speakers.Note however that our method relies on collaboration between local devices, such that we need to both authenticate and trust local devices. For such authentication, we can apply for example acoustic fingerprints [17].

2. Telecommunication scenario

We propose a telecommunication scenario where each device detects other devices in close proximity [17] and share their recorded signals. In particular, we consider a sensor network comprising two independent recording devices that can be seen as a linear array with two omni-directional microphones. Assuming an anechoic and noiseless context, the recorded mixture of M speakers in a specific device can be modeled in the discrete-time Fourier transform (DTFT) domain as

$$X_m(\omega) = S_m(\omega) + \sum_{\mu \in \mathfrak{I}} I_{m,\mu}(\omega), m = 1, 2, \qquad (1)$$

http://dx.doi.org/10.21437/Interspeech.2021-1329

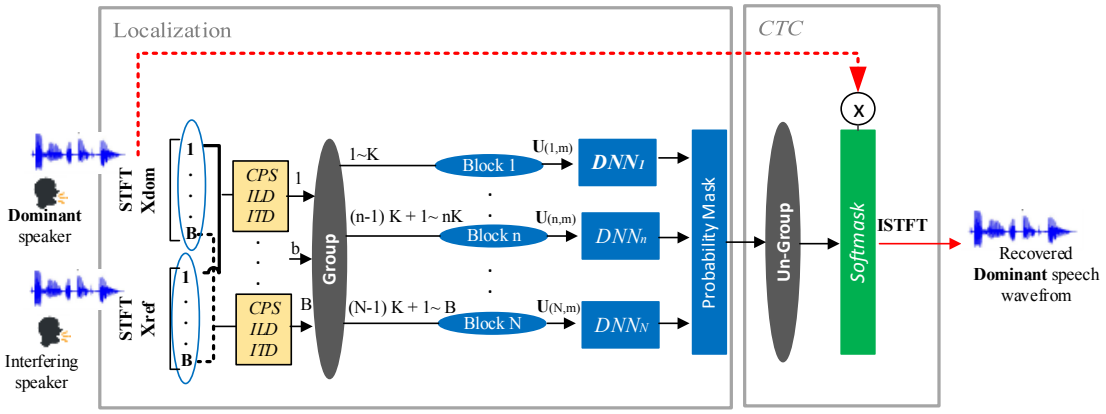

Figure 1: *The architecture of the proposed system using DNNs based time-frequency masking for the crosstalk cancellation (CTC) task.*

where $S_m(\omega)$ stands for the spectrum of the m^{th} specific target speaker and $I_{m,\mu}(\omega)$ denotes the signals from the μ^{th} interferer, with $\mu \in \Im = [1..M]$ when $\mu \neq m$. The speaker signal is attenuated with a positive amplitude decay factor $\beta_m \in \mathbb{R}_+$ and delayed by a transmission time of flight $\tau_m \in \mathbb{R}_+$ relative to the speaker position, which can be represented as

$$S_m(w) = \beta_m S(w) \exp(-jw\tau_m), m = 1, 2. \quad (2)$$

where $j = \sqrt{(-1)}$. The CTC problem consists in computing estimates $\tilde{S}_m$ for S_m, which depends on the delay parameter τ_m which in turn represents the source's spatial location. Under the sparsity assumption, contributing sources in each T-F bin can be classified to their spatial locations. Then, similar to [18], the target dominant speech can be estimated by exploiting the fact that each source is considered dominant in its dedicated microphone channel.

3. Proposed framework

Considering our two-device model, we propose a DNN-based system for the joint localization, and classification of sources. The system architecture of the proposed framework is shown in Fig. 1 and it consists of the following four stages: (1) extraction of the low-level features; (2) training of the DNN; (3) estimation of the probabilities that each T-F unit belongs to different sources and generation of the soft mask; and (4) recovering of the target signal from the soft mask and mixture signal.

3.1. Inputs to DNNs

We adopt the raw short time Fourier transform (STFT) as the input which contains required information for the localization task. The resulting spectrogram is a $(T \times B)$ complex-valued matrix, where T represents the amount of time frames contained in the audio sequence and B stands for the number of frequency bins. We obtain then the T-F representation for each dominant and interferer sources $X_{dom}(t, b)$ and $X_{int}(t, b)$ where $t = [1..T]$ and $b = [1..B]$. The input features are then estimated at each T-F unit and grouped into N uniformly distributed frequency blocks, each of them containing $K = \lceil \frac{B}{N} \rceil$ frequency bins. To localize speech sources, a DoA classifier is built using N DNNs, each of which is fed with its corresponding input vector of features referring to the n^{th} frequencyblock in the range $((n-1)K + 1, ..., nK)$.

As we assume a near-field telecommunication scenario, the inter-channel level difference (ILD) is expected to be close to 0 dB in low frequency bands as it is a function of frequency and thus invalid, because the sound wave period is larger than the microphone spacing. However, inter-channel time differences (ITD) perform well in the frequency range 500 Hz to 2000 Hz [19]. Nevertheless, ITD estimators underpinning the cross-correlation algorithm as well as its alternatives rely on a free-field propagation model of the sound waves. Therefore, the presence of multiple, simultaneously active sources can cause severe ambiguities in the distinction of peaks due to the direct path of the dominant source from peaks arising due to the interfering source, which is more challenging in the case of the near-field context [20]. Since cross-power-spectrum (CPS) involves ITD information, the combination of CPS and ILD can make the estimation of sound source direction more accurate. Thus, unlike [21], which uses mixing vector inter-channel phase difference (IPD) and ILD for the same DNN based separation method, we use a feature vector combining ILD and CPS derived from the observed input spectrum in each frequency block, consecutively defined as:

$$ILD(t, b) = 20 \log_{10} \left(\frac{|X_{int}(t, b)|}{|X_{dom}(t, b)|} \right), \in \mathbb{R}^{TXB} \quad (3)$$

$$CPS(t, b) = \frac{X_{dom}(t, b) X^*_{int}(t, b)}{|X_{dom}(t, b) X^*_{int}(t, b)|}, \in \mathbb{R}^{TXB} \quad (4)$$

$$ITD(t, b) = \underset{t}{\operatorname{argmax}} (CPS(t, b)), \in \mathbb{R}^{TXB} \quad (5)$$

where $|.|$ takes the absolute value of their arguments and argmax computes the time lag of the maximum peak. Concatenating CPS, ILD and ITD features, a vector is obtained at each T-F unit $U(t, b) = [CPS^T(t, b), ILD^T(t, b), ITD^T(t, b)]^T \in \mathbb{R}^{Tx3B}$.

All extracted vectors of features are split into N blocks, where each block involves only the information from K frequency bins as follows:

$$U_{(t,n)} = [U^T(t, (n-1)K + 1), ..., U^T(t, nK)]^T \in \mathbb{R}^{3K} \quad (6)$$

where n is the index of the n^{th} frequency block fed into the n^{th} DNN. For raw data with a sampling rate of 16 kHz, the STFT of the observed mixtures are computed in frames of 2048 samples (128 ms) with 50 % overlap. A sliding frequency window with 64 samples and 50 % overlap is then used to extract 31 CPS vectors in each time frame, which are then cropped to a number of samples equivalent to a maximum microphone distance around 2.5 m. Therefore, the CPS vector within a lag range of

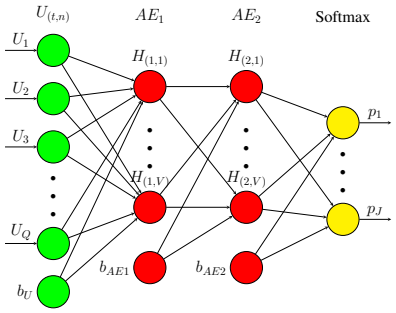

Figure 2: *Architecture of the used DNN_n: $U_{n,m}$ is the input vector of features and a bias vector b_U. AE_1 and AE_2 are two auto encoders and the bias vectors b_{AE_1} and b_{AE_2}. The softmax classifier gives the probability $p_j = j|U_{t,m}$.*

$[-128, +128]$ samples has size 256. Adding then ILD for 64 T-F units and one ITD value results in an input feature vector $U_{(t,n)}$ with length $Q = 321$ for each DNN.

3.2. The DNN architecture

The N DNNs in our proposed framework have the architecture shown in Fig. 2. It is composed of one input layer, two hidden layers using an unsupervised learning algorithm based on backpropagation namely sparse autoencoder (AE) [22] and an output layer using a softmax classifier.

Given a two-channels observed mixture, the input vector of features is computed and fed into the first layer. Therefore, the number of input neurons equals the dimension of the feature vector, $Q = 321$ neurons, and one bias unit. In order to extract high-level features from low level ones, similar to [23], we use two hidden sparse AEs which are composed of $V = 256$ neurons and one bias unit employing a sigmoid activation function. The estimated high level-features will serve as input for the DoA classifier based on a softmax layer containing 19 neurons corresponding to the J considered orientations ranges. Thus, our DNN outputs 19 nodes, which stand for the azimuth directions θ_j from –90° to +90° with steps of 10°, representing the likelihood of the presence of a sound source $P = \{p_j\}$ in the j^{th} orientation index. Assuming that dominant and interfering speakers are in different ranges, the output layer works as a softmax classifier estimating the probability of the signal coming from a specific direction for each frequency block.

3.3. Training method

The training of the DNN is divided into pre-training and fine-tuning stages similar to [23, 24]. In the pre-training phase, we use the greedy layer-wise training [25, 26] and the limited memory BFGS (L-BFGS) optimization algorithm to minimize the cost function. Each of the two sparse AEs and the softmax classifier are trained individually by using each output layer as input for the next layer, fixing the weights of the previous layers at each stage. The two AEs are trained using the unlabeled data of observed signals of the individual dominant speaker convolved with the corresponding RIR. Then, the outputs of the 2^{nd} AE are used to train the softmax classifier, which activates the j^{th} neuron giving the highest probability that the current set $U_{(t,n)}$ is oriented to the j^{th} direction. A cross-entropy loss function is therefore used to optimize the softmax classifier. Finally, the fine-tuning stage consists in stacking the softmax classifier and the AEs together, then training the overall DNN. This stage employs L-BFGS optimization to minimize the difference between

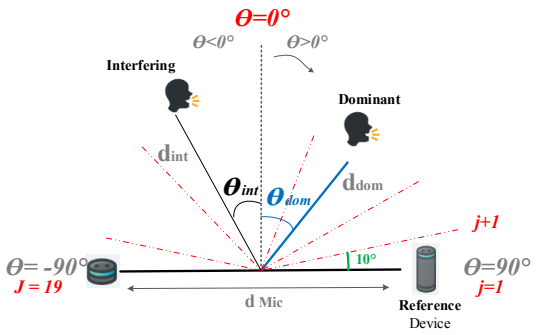

Figure 3: *The experimental setup. The index $j = [0, J]$, corresponding to a DoA ($\theta_{dom/int}$) ranging from −90° to +90°.*

the output of the DNN and the label of the training dataset.

For the different training stages the following parameters were chosen. The 1^{st} AE has a weight decay $\lambda = 9 \times 10^{-4}$, which is then set to $\lambda = 1 \times 10^{-4}$ for the 2^{nd} AE, softmax classifier and fine tuning. Both AEs use a sparsity parameter $\rho = 0.3$ with sparsity penalty $\beta = 2$ and trained over 300 epochs. However, the softmax classifier and the fine tuning stages are trained over 200 epochs.

3.4. Time-frequency masking

The softmax function outputs a source occupation probability of each block as noticed in the Fig. 1.Each T-F unit in the same block is assigned to the same source's occupation probability of the mixture. This probability mask is followed by an ungroup operation where the dominant target source is recovered by applying the softmask to the mixture's spectrogram and reconstructed using the inverse STFT (ISTFT). We assume that the nearest sensor to the dominant speaker is the reference channel in the CPS computation such that its corresponding azimuth orientation is positive, as shown in Fig. 3. In contrast, the competing source is assigned to the negative orientations. The same strategy will be applied for the interferer as it will be considered the dominant source in its closest reference device, and will be assigned to the positive orientation. It should be noticed that the flowchart is applied separately for each device unlike free-field source separation methods usually separating interfering contributions from the same mixture.

4. Experiments and discussion

A telecommunication scenario involving a single, static speech source, is first simulated to train our model for speech localization. Second, the robustness of the proposed method is investigated against overlapping speech from two sources. Therefore, we can study the impact of the source orientation on the CTC performance.

4.1. Experimental Setting

The training and test samples were generated using Pyroomacoustics [27]. The room size is $8 \times 6 \times 3\,\mathrm{m}^3$ with a reverberation time of about RT_{60} =0.2 s. Two sensors were then placed in the center of the room providing a horizontal aperture with inter-sensor distance $d_{mic} = [1, 1.5, 2]$ m.

Anechoic utterances from the LibriSpeech dataset [28] were randomly extracted for both train and test sets. For the train dataset, mixtures were generated for each of the J consid-

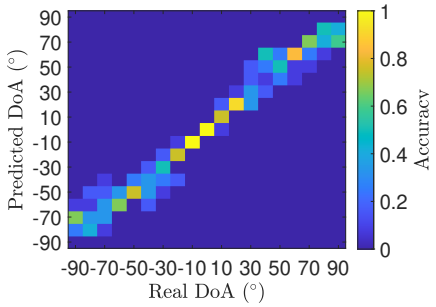

Figure 4: *The localization accuracy of the DNNs: Similarity between the ground truth (real) DOAs and predicted DOAs ranging from −90° to +90°.*

ered ranges between −90° to +90° in steps of 10° and the source was located at distances $d_{dom} = [1, 2, 3]$ m. 10 audio segments of duration 15 s were generated for each configuration, resulting in 7 h of speech. The test set is designed in a similar manner, limiting the orientation of the dominant source to [0°,90°] to ensure its proximity to the dominant recording channel. We then introduce an interfering source, equidistant to the center of the array.The interfering source's orientation is set to [−90°,0°], with a minimum spacing of 40° (i.e. 4 classes) with respect to the dominant source. The interfering signal is scaled to provide source-to-interference ratio ($iSIR$) levels of 0, 5 and 10 dB with respect to the dominant source in each mixture. The audio recordings were performed with a sampling rate of 16 kHz and synchronized with the fixed ground truth positional data relative to the inter-sensors center. Each data sample is labeled using a 19-dimensional one-hot encoded vector corresponding to the set azimuth for each generated mixture.

4.2. Experimental results

As the proposed CTC solution is closely depending on the DNNs-based DOA estimator, its overall performance can be then evaluated in terms of source localization accuracy and CTC quality. Hence, in this section, we analyze the DoA estimator accuracy, as well as the CTC output quality.

4.2.1. The localization accuracy of the DNNs

A good indication of the quality of the DNN output is the plot of the ground-truth DOAs in which a target source is positioned versus the DNN output predicted DoA. By averaging on the frequency dimension and on the number of consecutive frames within the same test set J, we obtain a confusion matrix representing the DoA accuracy as shown in Fig. 4. It reflects that the used DNN shows a high similarity between the ground-truth DoA and the predicted labels. High accuracy is noticed at small azimuth angles, however slightly smaller accuracy is shown for higher azimuth orientation. Nevertheless, DoA mismatch corresponds approximately to 2 ranges, or an azimuth error of about 20°. This error is due to the windowing effect on CPS representation introducing some deviation in the detected highest evidence referring to the ground-truth DoA, which subsequently affects the activated range (neuron) in the softmax layer. Nevertheless, the effect of this bias on T-F Mask is expected to be significantly small since the CTC scenario is assuming more than 4 ranges separating the dominant speaker from the interfering one. Hence, our DNN-based DoA estimator is sufficiently accurate in the near-field telecommunication scenario. It should be noted that our DoA resolution is closely related to

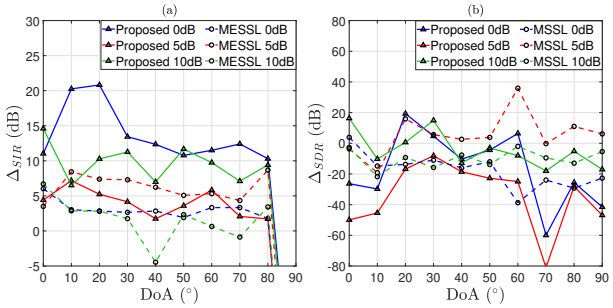

Figure 5: *CTC performance comparison between the proposed workflow and MESSL method [29] for Signal-to interference ratio ($iSIR$) ranging from 0 to 10 (dB) when dominant source is located in all considered DoA ranges in degrees: (a) the source- to-interference ratio improvement (Δ_{SIR}), (b) the source-to-distortion ratio improvement (Δ_{SDR}).*

the choice of frequency bins, set to $K = 64$. A smaller value allows for higher resolution in the localization of the dominant source, however, it increases the training time. Therefore, we need to find a compromise in terms of DoA accuracy resolution and training time.

4.2.2. CTC evaluation

The CTC performance of our proposed system is evaluated under different $iSIRs$ and different azimuth ranges, as described in Section 4.1.We present in Fig. 5 (a) and (b) the improvements in terms of source- to-interference ratio (Δ_{SIR}) and source-to-distorsion ratio (Δ_{SDR}) respectively, using the museval library [30]. The proposed results are compared to a model-based expectation-maximization BSS (MESSL) method [29], which also relies on a localization estimator. We can clearly notice that the proposed method achieves a better interference rejection with a higher Δ_{SIR} reaching up to 20 dB for lower azimuth values, for which our DNN is most accurate. However, increasing the source's orientation affects the T-F mask resolution leading to a lower interference rejection, especially when the source is perfectly aligned with the microphone pair. We also notice that interference rejection is performed with a higher distortion constraint, as shown in Fig. 5 (b).The main source of distortion comes from the resolution of the T-F mask affected by DoA estimation errors under low $iSIR$ levels and high azimuth angles. Nevertheless, the proposed method, outperforms the MESSL approach in terms of CTC with a higher Δ_{SIR}. By establishing a compromise with the distortion level, our method is suitable for protecting the privacy of external speakers in the same acoustic space.

5. Conclusion

When multiple users are simultaneously speaking to their devices, then the competing speakers will leak into all devices. Such cross-talk reduces the audio quality and can lead to breaches in privacy. Inspired by the recent advances in blind source separation, we propose to adapt a T-F masking algorithm inferred by a deep direction estimation for the CTC task. Our proposed CTC system uses frequency-dependent features for an unsupervised learning approach, mapping between the inter-channels cues and contributing sources' locations using a DNNs based model. The proposed method can thus be used to improve user experience in multi-user scenarios in terms of privacy by canceling interfering speakers in a near-field context.

6. References

[1] T. Pfau, D. P. W. Ellis, and A. Stolcke, "Multispeaker speech activity detection for the ICSI meeting recorder," in *in Proc. of ASRU IEEE, Madonna di Campiglio*, 2001, pp. 107–110.

[2] T. Prätzlich, R. M. Bittner, A. Liutkus, and M. Müller, "Kernel additive modeling for interference reduction in multi-channel music recordings," in *Proc ICASSP*. IEEE, 2015, pp. 584–588.

[3] E. K. Kokkinis, J. D. Reiss, and J. Mourjopoulos, "Wiener filter approach to microphone leakage reduction in close-microphone applications," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 3, pp. 767–779, 2012.

[4] S. Wrigley, G. J. Brown, V. Wan, and S. Renals, "Speech and crosstalk detection in multichannel audio," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 13, no. 1, pp. 84–91, 2005.

[5] C. Uhle and J. Reiss, "Determined source separation for microphone recordings using IIR filters," in *Pro AES Convention*, 2010, pp. 1–4.

[6] A. Lombard and W. Kellermann, "Multichannel cross-talk cancellation in a call-center scenario using frequency-domain adaptive filtering," in *Proc of IWAENC*, 2008, pp. 14–17.

[7] T. Matheja, M. Buck, and T. Fingscheidt, "A dynamic multichannel speech enhancement system for distributed microphones in a car environment," *EURASIP J. Adv. Signal Process*, vol. 191, pp. 1–21, 2013.

[8] J. J. Carabias-Orti, M. Cobos, P. Vera-Candeas, and F. J. Rodríguez-Serrano, "Nonnegative signal factorization with learnt instrument models for sound source separation in close-microphone recordings," *EURASIP J. Adv. Signal Process.*, vol. 184, pp. 1–16, 2013.

[9] D. D. Carlo, K. Déguernel, and A. Liutkus, "Gaussian framework for interference reduction in live recordings," in *Proc AES International Conference on Semantic Audio*, 2017, pp. 1–8.

[10] P. Meyer, S. Elshamy, and T. Fingscheidt, "A multichannel Kalman-based Wiener filter approach for speaker interference reduction in meetings," in *Proc ICASSP*. IEEE, 2020, p. 451–455.

[11] ——, "Multichannel speaker interference reduction using frequency domain adaptive filtering," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2020, no. 1, pp. 1–17, 2020.

[12] K. Ochi, N. Ono, S. Miyabe, and S. Makino, "Multi-talker speech recognition based on blind source separation with ad-hoc microphone array using smartphones and cloud storage," in *Proc Interspeech*. ISCA, 2016, p. 3369–3373.

[13] J. P. Dmochowski, Z. Liu, and P. A. Chou, "Blind source separation in a distributed microphone meeting environment for improved teleconferencing," in *Proc ICASSP*. IEEE, 2008, p. 89–92.

[14] Y. Yu, W. Wang, and P. Han, "Localization based stereo speech source separation using probabilistic time-frequency masking and deep neural networks," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2016, no. 1, pp. 1–18, 2016.

[15] A. Alinaghi, P. J. Jackson, Q. Liu, and W. Wang, "Joint mixing vector and binaural model based stereo source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 9, pp. 1434–1448, 2014.

[16] S. Rickard and O. Yilmaz, "On the approximate w-disjoint orthogonality of speech," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, 2002, pp. I–529–I–532.

[17] P. P. Zarazaga, T. Bäckström, and S. Sigg, "Acoustic fingerprints for access management in ad-hoc sensor networks," *IEEE Access*, vol. 8, pp. 166 083–166 094, 2020.

[18] E. K. Kokkinis, J. D. Reiss, and J. Mourjopoulos, "A Wiener filter approach to microphone leakage reduction in close-microphone applications," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 3, pp. 767–779, 2012.

[19] C. Zheng, A. Schwarz, W. Kellermann, and X. Li, "Binaural coherent-to-diffuse ratio estimation for dereverberation using an ITD model," in *European Signal Processing Conf.*, 2015, pp. 1048–1052.

[20] M. Bouafif and Z. Lachiri, "TDOA estimation for multiple speakers in underdetermined case," in *Interspeech*, Portland, Oregon, USA, 2012, pp. 1748–1751.

[21] Y. Yu, W. Wang, J. Luo, and F. Pengming, "Localization based stereo speech separation using deep networks," in *Proc. ICASSP*. IEEE, 2015.

[22] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 08, pp. 1798–1828, 2013.

[23] D. Yu and L. Deng, "Deep learning and its applications to signal and information processing," *IEEE Signal Process. Mag.*, vol. 28, no. 1, pp. 145–154, 2011.

[24] J. Dean, R. M. G. Corrado, K. Chen, and M. D. M. Mao, "Large scale distributed deep networks," *Advances in Neural Information Processing Systems.*, pp. 1223–1231, 2012.

[25] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layerwise training of deep networks," *Advances In Neural Information Processing Systems.*, vol. 19, p. 153, 2007.

[26] D. C. Liu and J. Nocedal, "On the limited memory bfgs method for large scale optimization," *Mathematical Programming.*, vol. 45, no. 1-3, pp. 503–528, 1989.

[27] R. Scheibler, E. Bezzam, and I. Dokmanic, "Pyroomacoustics: A python package for audio room simulations and array processing algorithms," *CoRR*, vol. abs/1710.04196, 2017. [Online]. Available: http://arxiv.org/abs/1710.04196

[28] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *Proc ICASSP*, IEEE. IEEE, 2015, pp. 5206–5210.

[29] M. Mandel, R. Weiss, and D. Ellis, "Model-based expectation-maximization source separation and localization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 2, pp. 382–394, 2010.

[30] F. R. Stöter, A. Liutkus, and N. Ito, "The 2018 signal separation evaluation campaign," in *Latent Variable Analysis and Signal Separation: 14th International Conference, LVA/ICA 2018*, Surrey, UK, 2018, pp. 293–305.