Neural Networks Using Full-Band and Subband Spatial Features for Mask Based Source Separation

Alexander Bohlender[1], Ann Spriet[2], Wouter Tirry[2], and Nilesh Madhu[1]
[1] IDLab, Department of Electronics and Information Systems, Ghent University - imec, Ghent, Belgium
[2] Goodix Technology (Belgium) B.V., Leuven, Belgium
Email: alexander.bohlender@ugent.be; aspriet@goodix.com; wtirry@goodix.com; nilesh.madhu@ugent.be

***Abstract*—With a microphone array, spatial diversity can be exploited to estimate time-frequency masks that effectively suppress interfering speakers as well as noise. Here, we propose a deep learning approach where the signal components are distinguished based on the associated directions of arrival. To capture the target signal spectrogram more accurately, the estimation can be performed for each subband separately. In order to also take advantage of cross-band dependencies, we additionally consider a combined subband and full-band architecture. Our evaluation indicates that this combination consistently improves the performance in terms of instrumental quality metrics as compared to a pure subband or full-band method. Further, the comparison with two baseline approaches demonstrates the effectiveness of the location based deep learning approach.**

***Index Terms*—source separation, speech enhancement, time-frequency masking, neural networks, direction-of-arrival**

I. Introduction

Time-frequency (TF) masks identify short-time Fourier transform (STFT) bins that are dominated by a signal of interest, which makes them an effective tool for tasks like speech enhancement and source separation [1]. Whereas they can also be acquired with unsupervised clustering approaches like [2], [3], the focus of this paper is on deep neural network (DNN) based mask estimation. When multiple microphones are available, spatial information can be exploited for this purpose. Because source locations are not specific to one signal type, e. g., speech, they are highly useful to distinguish components in a mixture of contributions from multiple sources. For compact arrays, it is primarily the *phase* component which contains the spatial information. This is exploited, e. g., in [4], [5]. Whereas [4] considers only phase, a combination of magnitude and phase is used in [5]. In [6], where this aspect is studied more closely, no significant improvement is observed when magnitude is incorporated in addition to the spatial information.

Approaches like [4]–[6] mix the information from all frequencies. This mixing can, however, cause a loss of the TF fine structure of the resulting mask. A recent single-speaker approach for denoising that considers each frequency independently is [7]. With a long short-term memory (LSTM) network, the authors report a performance that is at least comparable to other state-of-the-art approaches, although cross-frequency information is neglected entirely.

Thus, the aforementioned approaches can be classified as being either full-band (FB) ([4]–[6]), or subband (SB) ([7]) methods. In this work, we propose convolutional neural network (CNN)-based architectures of both classes and, since there are arguments in favor of both, consider a combination of the two. In contrast to [5], [7], where noise suppression is the primary focus, our goal is the separation of (localized) sources. An LSTM is used to exploit temporal context. Nevertheless, in contrast to [6] where models are trained for specific scenarios, we do not require assumptions regarding, e. g., the movement of sources. This is achieved by using the source directions of arrival (DOAs) to distinguish between the components.

Following the problem formulation in Sec. II, we present three variants of the proposed DNN in Sec. III: the FB, SB, and mixed mask estimators. The evaluation results in Sec. IV show that whereas the SB approach best captures the fine structure of the target signal, the mixed approach yields the highest scores in terms of instrumental metrics. Sec. V concludes the paper.

This work is partially supported by the Research Foundation - Flanders (FWO) under grant numbers 11G0721N and G081420N.

II. Mask Based Speech Separation

A. Signal Model and Problem Statement

In the STFT domain, the signal at the n-th microphone is denoted by $Y_n(\mu,\lambda)$, where $\mu=0,\ldots,M-1$ is the frequency index, and λ the frame index. The N microphones pick up a mixture of filtered versions of the J dry source signals $S_j(\mu,\lambda)$, where $1\leq j\leq J$, and an additive noise $V_n(\mu,\lambda)$. Denoting the direct path components by $S^{\mathrm{dir}}_{j,n}(\mu,\lambda)$, and the reverberation components by $S^{\mathrm{rev}}_{j,n}(\mu,\lambda)$, we obtain the signal model

$$Y_n(\mu,\lambda) = \textstyle\sum_j \left(S^{\mathrm{dir}}_{j,n}(\mu,\lambda) + S^{\mathrm{rev}}_{j,n}(\mu,\lambda)\right) + V_n(\mu,\lambda). \quad (1)$$

The direct path component differs from the dry source signal only in terms of a time delay, and an attenuation factor. Therefore, our aim is to extract all direct path components $S^{\mathrm{dir}}_{j,1}(\mu,\lambda)$ at the reference microphone $n=1$ (this selection is arbitrary).

B. Time-Frequency Masking

We define a mask $\mathcal{M}_j(\mu,\lambda)$ to quantify the activity of the j-th source at the corresponding TF point. In this work, we employ the most straightforward approach to obtain a target speech estimate: the mask is multiplied with the reference microphone signal directly, i. e., $\widehat{S}^{\mathrm{dir}}_{j,1}(\mu,\lambda)=\mathcal{M}_j(\mu,\lambda)\cdot Y_1(\mu,\lambda)$.

An overview of various suitable masks can be found, e. g., in [1]. Without loss of generality, we consider, here, the (bounded) squared spectral magnitude mask (SSMM)

$$\mathcal{M}_j(\mu,\lambda) = \min\left\{\left|\gamma_j S^{\mathrm{dir}}_{j,1}(\mu,\lambda)\right|^2 \Big/ \left|Y_1(\mu,\lambda)\right|^2; 1\right\}. \quad (2)$$

By defining the mask w. r. t. the squared magnitudes, unwanted components are suppressed more vigorously, albeit at the cost

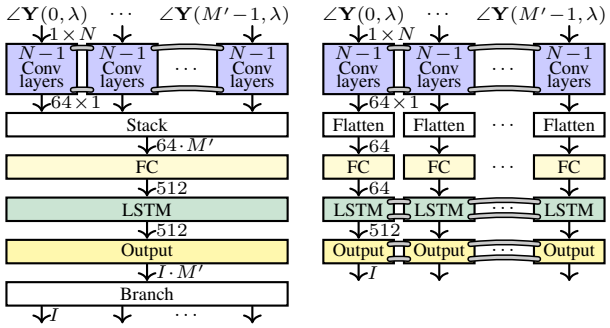

Fig. 1: FB-MEst and SB-MEst architectures for TF mask estimation (MEst). Connections (⬭) indicate layers that share the same trainable parameters.

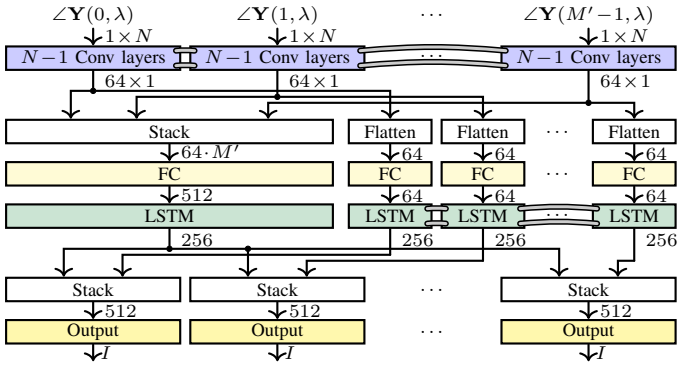

Fig. 2: Mix-MEst: one half of the features based on which the output masks are computed are obtained following, respectively, the FB and SB approaches.

of more speech distortion compared to, e. g., the SMM, or the ideal ratio mask (IRM). Additionally, for a more consistent distribution of the target mask values across different source-array distances and reverberation levels, we use the normalization

$$\gamma_j = \sqrt{\frac{\sum_{\mu,\lambda,n}\left|S^{\mathrm{dir}}_{j,n}(\mu,\lambda)+S^{\mathrm{rev}}_{j,n}(\mu,\lambda)\right|^2}{\sum_{\mu,\lambda,n}\left|S^{\mathrm{dir}}_{j,n}(\mu,\lambda)\right|^2}} \tag{3}$$

on the direct path component. This permits us to set upper and lower mask bounds without compromising the suppression of unwanted components in adverse conditions (see Sec. III-E).

III. CNN for Multisource TF Mask Estimation

The angular space around the array is partitioned into a set of I discrete DOAs. For each direction, one TF mask will be generated to recover the impinging direct path sound, while suppressing noise, and interference from other directions. The J localized sources may be uniquely identified by their DOAs. Using a broadband DOA estimator, such as [8], the required subset of the I masks can be found by selecting, for each source, the nearest discrete DOA for which a mask is available.

A. Full-band TF Mask Estimator

The CNN shown in Fig. 1a will be referred to as the full-band mask estimator (FB-MEst). The vector of microphone signal phases $\angle\mathbf{Y}(\mu,\lambda)=[\angle Y_1(\mu,\lambda),\ldots,\angle Y_N(\mu,\lambda)]$ serves as input. Convolutions are applied across the channel dimension, in the form of 64 frequency independent filters per input map, each of length 2. Without zero-padding or pooling, the channel dimension is thus reduced to 1 after $N-1$ layers. The features from all frequencies are stacked to one vector of length $64\cdot M'$, where $M'=M/2+1$ is the number of discrete frequencies up to the Nyquist frequency. Empirically, we find that increasing the size of the fully connected (FC) and LSTM layers does not improve the performance. The output consists of $I\cdot M'$ elements, which are divided into the masks for all I DOAs.

Relation to prior work: Essentially, the architecture is identical to the CNN/LSTM used in [8], which is an extension of the CNN proposed in [9]. Only the tasks differ: whereas [8], [9] focus on *broadband* DOA estimation, FB-MEst takes advantage of the same DOA framework to acquire masks, i. e., *narrowband* information, that is conditioned on the DOAs.

In [5], the authors of [9] use a similar architecture for TF masking as well. However, only one mask for separating speech from noise is estimated. In this regard, FB-MEst can be seen as an extension of [5] to the multi-speaker case.

B. Subband TF Mask Estimator

For acquiring narrowband information, the input for each *individual* TF bin is of particular interest. We therefore propose to leave the frequency structure intact after the convolutional layers. Thus, in the subband MEst (SB-MEst) architecture depicted in Fig. 1b, the information from different frequencies is not mixed, i. e., cross-band dependencies are neglected. Nevertheless, the task remains the same for each frequency. This is exploited by *tying* the trainable parameters across all subbands, thereby enforcing a more abstract (frequency independent) feature representation, and reducing the total number of parameters. Here, a frequency independent LSTM, for example, is reasonable as a similar temporal evolution is expected for all frequencies. Only in the FC layer, we untie the parameters to account for the frequency dependence of the input.

Without frequency mixing, the FC layer receives 64 (rather than $64\cdot M'$) features. Therefore, we also set the output size of this layer to 64 (rather than 512). So as to not constrain the mask construction, we do not reduce the size of the LSTM.

Relation to prior work: An LSTM network that uses only subband information is also proposed in [7]. The separation of multiple speakers, however, is not addressed. In this case, the trainable parameters are shared across frequency in *all* layers. This is practicable under the assumption of a single localized target component (speech in noise). Because the interchannel phase differences are dependent on the frequency as well as the DOA, frequency dependent processing *is* important, however, to distinguish between components from different directions.

The distinctness of the DOAs is, on the other hand, exploited for speaker separation with a binaural setup in [10], where small frequency blocks are processed with independently trained DNNs. Unlike SB-MEst, however, a narrowband DOA classification is performed, rather an estimation of masks for DOAs that have been identified in advance.

C. Mixed Full-Band and Subband TF Mask Estimator

By disregarding cross-band dependencies, the SB-MEst architecture does not fully exploit all available information. To address this limitation, we propose a combination of FB-MEst and SB-MEst, as illustrated in Fig. 2. This will be referred to as Mix-MEst. As the figure shows, half of the features provided to the output layer are generated according to the FB-MEst approach, the other half according to the SB-MEst approach.

(a) 3-mic subarray (b) 9-mic subarray

Fig. 3: Two different subarrays (●) (3 and 9 microphones, respectively) of the the miniDSP UMA-16 array [15] 16-microphone URA will be considered.

Further, in contrast to SB-MEst, the output layer parameters are untied so that the cross-band information can be used differently at each frequency. At high frequencies, for example, SB information may be less reliable due to spatial aliasing.

Relation to prior work: A different FB and SB combination is used in the very recently proposed FullSubNet [11]. Instead of a parallel structure, a cascade is employed: FB *output* and SB *input* are concatenated before the final layers. Since the approach is used exclusively for single-channel speech enhancement, the input consists only of magnitude information.

D. Complexity

The number of multiply-accumulate (MAC) operations per frame is lowest for FB-MEst (about $10.8{\times}10^6$ for $M'{=}257$, $I{=}37$, and $J{=}2$, disregarding convolutional layers), and highest for SB-MEst ($305{\times}10^6$). Mix-MEst lies in between the two ($94.7{\times}10^6$). Because the trainable parameters are the same for all subbands, except in the FC layer, the total number of parameters, in contrast, is lowest for SB-MEst ($2.3{\times}10^6$).

E. Training

We generate training data as proposed in [8]. A dynamic setting is considered, where each talker can be active at different times and at different locations. To model this, a Markov chain $A_j(\lambda)$ is used to decide when the j-th source is active. The two states $A_j(\lambda){=}\{1,0\}$, respectively, indicate activity and inactivity of source j for time frame λ. During inactivity, the contribution of this source is set to 0. Here, the probability for a transition between the two states is chosen so that there is an average of one transition per $1.5\,\mathrm{s}$.

The TIMIT [12] and PTDB-TUG [13] speech databases are used for the source signals. The location of a source remains fixed while it is active. A new location (i. e., DOA and source-array distance) is selected once a previously inactive source becomes active again ($A_j(\lambda){=}1$, $A_j(\lambda{-}1){=}0$). The training set includes sequences with both $J{=}1$, and $J{=}2$ sources.

The source signals are convolved with room impulse responses (RIRs) that we simulate using [14]. In the simulation, the source is placed at one of $I{=}37$ different azimuth angles $\varphi{=}0°, 5°, \ldots, 180°$, such that array and sources are coplanar. Two different array geometries are used (see Fig. 3). To cover a wide range of acoustic conditions, we consider $R{=}10$ different rooms with reverberation times ranging from $T_{60}{=}0.2\,\mathrm{s}$ to $0.8\,\mathrm{s}$. Further, the number of positions of the array per room is $P{=}7$, and the number of source-array distances per array position $D{=}4$. For the validation, a different set of RIRs is used. Finally, a spherically isotropic (diffuse), but temporally uncorrelated noise field is simulated as described in [16]. For the additive mixing, the sources-to-noise ratio (SNR) is selected randomly between $0\,\mathrm{dB}$ and $30\,\mathrm{dB}$ for each mixture.

To permit a satisfactory suppression, it is important to also capture low mask values accurately. For example, the difference between 0.01 and 0.10 ($20\,\mathrm{dB}$) is more significant than between 0.9 and 1.0 (about $1\,\mathrm{dB}$). Therefore, we employ a dB representation for the target output, based on which the mean squared error loss is computed as well. However, accurately estimating *very* low mask values is neither feasible nor beneficial. Rather, preserving a certain noise floor can help masking artifacts such as musical tones. Along with the upper bound of $0\,\mathrm{dB}$ imposed by (2), we therefore lower bound the mask values by $-40\,\mathrm{dB}$. To reflect this, we use an otherwise linear output activation function that clips values outside this interval.

Although I masks may be obtained in total, only those output masks that correspond to the true source DOAs are used to compute the loss. Based on the validation loss, the weight decay parameter of the AdamW optimizer [17] is set to 0.002 for the 9-mic SB-MEst architecture, and 0.001 for all others. All training sequences have a length of $2\,\mathrm{s}$. We make use of dropout with rate 0.5, and batch normalization. The ReLU activation function is used in the hidden layers.

IV. Evaluation

A. Setup

Microphone signals for the evaluation are generated by additively mixing $J{=}2$ source contributions, and recorded diffuse noise. The source signals consist of 5 concatenated utterances of the TSP speech database [18], each of which has been convolved with a RIR at sampling rate $f_s = 48\,\mathrm{kHz}$. A new (unique) DOA, and thus a new RIR, is selected with probability 50% after each utterance. The resulting mixture is downsampled to $16\,\mathrm{kHz}$, and transformed into the STFT domain. The frames of length 512 samples are windowed with a square-root Hann window ($M{=}512$). The frame shift is 160 samples.

The RIRs for azimuth angles $\varphi{=}0°, 20°, \ldots, 180°$ were recorded in a meeting room ($T_{60}{=}660\,\mathrm{ms}$) using the miniDSP UMA-16 array [15]. The source-array distance was $2\,\mathrm{m}$. To obtain diffuse noise, the pub noise signal from the ETSI background noise database [19] was simultaneously played back by four loudspeakers located at the corners of a room with $T_{60}{\approx}1\,\mathrm{s}$, and recorded using the same array.

To estimate the DOAs, we make use of the CNN/LSTM broadband DOA estimator of [8]. Being a data-driven approach, it is relatively robust to reverberation and noise: the DOA estimation error $|\varphi{-}\widehat{\varphi}|$ does not exceed $5°$ in 85% of the frames at the considered noisy conditions ($\mathrm{SNR}{=}0\,\mathrm{dB}$) for both arrays. The architecture is equivalent to that of FB-MEst, except that a DOA classification is performed instead of a TF mask estimation. The J classes with the highest probability are used as the DOA estimates, where the number of sources J is assumed known a-priori. Note that the same DOA estimates are used for *all* approaches considered in the following.

B. Baselines

For the comparison with the proposed approaches, i. e., FB-MEst, which can be seen as a variation of [5], SB-MEst, and Mix-MEst, we consider the ideal ("oracle") mask, as well as two baselines that take advantage of classical DOA estimation

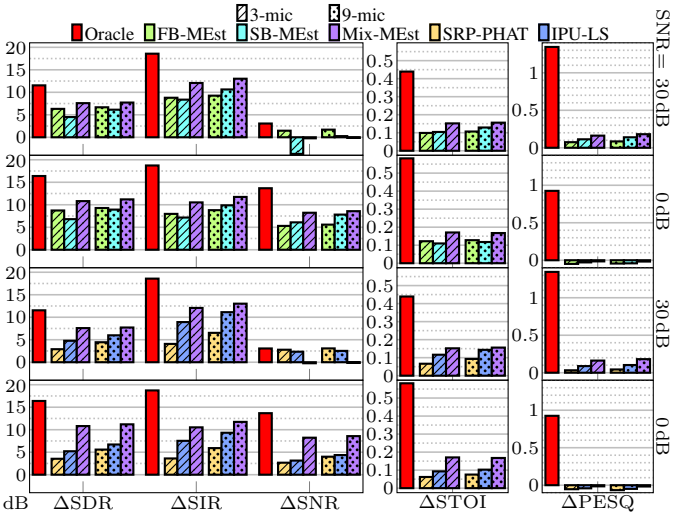

Fig. 4: The architectures are compared in the first two rows, followed by the comparison of Mix-MEst with the baseline methods in the bottom two rows.

methods. Thereby, we aim to provide an overview of the respective strengths and weaknesses of these methods.

One approach to acquire TF masks with narrowband DOA estimates is to use the estimates to derive a Gaussian mixture model from which posterior probabilities that serve as the masks can subsequently be extracted [20]. The proposed approach, however, incorporates only the *broadband* DOAs. For comparability, we therefore instead exploit that various classical DOA estimation methods, such as narrowband realizations of SRP-PHAT [21], and IPU-LS [22], are based on the maximization of a function $\mathcal{J}(\mu,\lambda,\varphi)$ over all angles φ. By evaluating this function *only* at the broadband DOAs φ_j, a TF mask for source separation is straightforwardly given by

$$\mathcal{M}_j^{\text{sep}}(\mu,\lambda) = \frac{\mathcal{J}(\mu,\lambda,\varphi_j)}{\sum_{j'} \mathcal{J}(\mu,\lambda,\varphi_{j'})}. \tag{4}$$

However simple, our experiments indicate that this approach is relatively robust, which makes it suitable for assessing the benefit of using deep learning in DOA based source separation based on the comparison with the proposed approach. To account for *diffuse* noise, we combine (4) with a postfilter [23] $\mathcal{M}_j^{\text{noi}}(\mu,\lambda)$ that is applied independently to each initial source estimate $\mathcal{M}_j^{\text{sep}}(\mu,\lambda)\cdot Y_1(\mu,\lambda)$. Empirically, we set lower bounds $-40\,\text{dB}$ for $\mathcal{M}_j^{\text{sep}}(\mu,\lambda)$, and $-12\,\text{dB}$ for $\mathcal{M}_j^{\text{noi}}(\mu,\lambda)$ to reduce musical noise. The required power spectral density matrices are estimated by recursive averaging with time constant $40\,\text{ms}$. For conciseness, we simply refer to the combination $\mathcal{M}_j^{\text{sep}}\cdot\mathcal{M}_j^{\text{noi}}$ as the SRP-PHAT or IPU-LS mask.

C. Results

Fig. 4 shows numerical results for two different SNRs: $30\,\text{dB}$ (first row), and $0\,\text{dB}$ (second row). For all metrics, based on 25 independently generated sets of microphone signals, the average improvement (Δ) compared to the unprocessed reference microphone signal is presented. Specifically, we consider the *segmental* source-to-distortion ratio (SDR), source-to-interferences ratio (SIR), and SNR [24], as well as STOI [25], and wideband PESQ [26] on a MOS-LQO scale.

To better understand the results, it is instructive to take a closer look at one particular example, as depicted in Fig. 5.

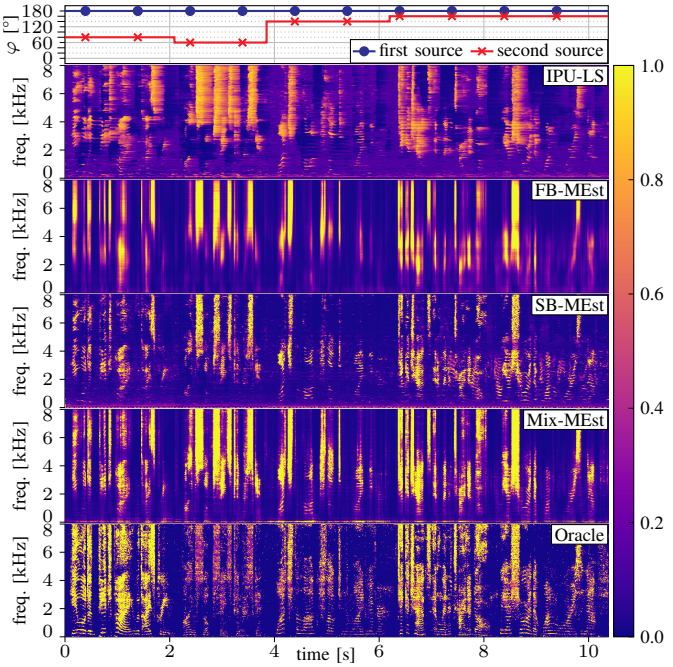

Fig. 5: TF masks for the *second* (-×-) source (9-mic array, $\text{SNR}=0\,\text{dB}$). Out of the two baselines, only the best performing one (IPU-LS) is displayed.

Only in this case, we assumed the availability of the true DOAs. Despite being undesirably coarse, the comparison with the oracle SSMM indicates that TF regions of high desired signal energy are identified correctly by FB-MEst. As Fig. 4 shows, this may lead to a relatively good performance of FB-MEst (▯) in terms of the considered metrics (for $\text{SNR}=0\,\text{dB}$: about $\Delta\text{SDR}=9\,\text{dB}$ with both arrays). However, Fig. 5 demonstrates that the SB-MEst mask captures the details significantly better: the harmonic structure of the target signal can clearly be recognized in the mask. This also becomes apparent upon listening to the resulting audio files[1]: although FB-MEst attains a better *overall* suppression of interference and noise, the signal is not enhanced *locally*, e. g., between harmonics. The lack of suppression in regions with a strong presence of the desired signal gives the auditory impression of a considerable target speech distortion. Consequently, Fig. 4 shows a better ΔPESQ for SB-MEst (0.14 for the 9-mic array) than for FB-MEst (0.08) at $\text{SNR}=30\,\text{dB}$. In the presence of strong noise ($\text{SNR}\lesssim 0\,\text{dB}$), however, we find that the given conditions are too adverse for PESQ to be a reliable measure.

Despite capturing the fine structure of the speech well, by neglecting cross-band information SB-MEst cannot suppress unwanted components satisfactorily. As Fig. 5 indicates, along with the audio files, Mix-MEst inherits the advantages of both FB-MEst and SB-MEst. The perceived speech distortion is clearly reduced compared to FB-MEst, but remains noticeable. In terms of the numerical results in Fig. 4, the improvement of Mix-MEst (▯) compared to FB-MEst and SB-MEst is evident (e. g., for $\text{SNR}=0\,\text{dB}$ with the 3-mic array: $\Delta\text{STOI}=0.17$ with Mix-MEst, 0.12 or less with the other approaches).

In the last two rows of Fig. 4, we compare Mix-MEst with the baselines introduced in Sec. IV-B. It appears that IPU-

[1] https://users.ugent.be/~abohlend/EUSIPCO2021/

LS is better suited for the direct mask computation than SRP-PHAT. Upon closer examination, we find that this is due to the sharper peaks produced by the IPU-LS cost function. Therefore, we will focus on this baseline method in the following. Considering speaker separation only, IPU-LS performs favorably ($\Delta\mathrm{SIR} = 11\,\mathrm{dB}$ with the 9-mic array for $\mathrm{SNR} = 30\,\mathrm{dB}$). The audio example, in which the sources come relatively close ($\Delta\varphi = 20°$) in the final seconds, shows that Mix-MEst, in contrast to the IPU-LS mask, then no longer suppresses the interfering speaker very well. The impression of a distorted target signal can also be avoided because, like SB-MEst, the IPU-LS based method only enhances the signal locally as cross-band dependencies are not exploited. The noise suppression, however, which is addressed solely by the postfilter, is inferior to Mix-MEst. This is reflected in the audio example, where musical noise is present in the IPU-LS output, and the instrumental metrics in Fig. 4 (for the 9-mic array at noisy conditions: $\Delta\mathrm{SNR} = 9\,\mathrm{dB}$ with Mix-MEst, only $4\,\mathrm{dB}$ with IPU-LS).

V. Conclusions

We proposed a deep learning approach for mask based source separation, where signal components are distinguished based on the DOAs. Three related CNN architectures were considered: FB-MEst mixes information from all frequencies, whereas SB-MEst processes each subband independently. To fully exploit the information contained in each individual TF bin without neglecting cross-band dependencies, Mix-MEst combines both. The evaluation based on speech signals showed that masks produced by Mix-MEst capture the coarse as well as the fine structure of the ideal mask fairly well. However, there is only a limited suppression in TF regions where the target signal is dominant, which can give the impression of the speech being distorted. Therefore, under certain conditions, SB-MEst may still be preferred despite the metrics indicating an inferior performance. Finally, comparing Mix-MEst with two baseline approaches based on classical narrowband DOA estimation demonstrates that incorporating deep learning is beneficial particularly in adverse conditions, but reveals that the separation of closely spaced sources can still be improved.

To better capture target speech while exploiting cross-band information, the further improvement of the local suppression could be studied in future work. Moreover, whereas it was observed that powerful TF masks can be derived from spatial information with just 3 microphones, the performance was only slightly better for the 9-microphone array. To benefit more from additional microphones, the integration of the masks into an adaptive beamforming framework may be considered.

References

[1] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 10, pp. 1702–1726, 2018.

[2] A. Jourjine, S. Rickard, and O. Yilmaz, "Blind separation of disjoint orthogonal signals: demixing n sources from 2 mixtures," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, vol. 5, 2000, pp. 2985–2988.

[3] N. Ito, S. Araki, and T. Nakatani, "Complex angular central gaussian mixture model for directional statistics in mask-based microphone array signal processing," in *Proc. 24th Eur. Signal Process. Conf.*, 2016, pp. 1153–1157.

[4] P. Pertilä and J. Nikunen, "Distant speech separation using predicted time–frequency masks from spatial features," *Speech Communication*, vol. 68, pp. 97–106, 2015.

[5] S. Chakrabarty and E. A. P. Habets, "Time–frequency masking based online multi-channel speech enhancement with convolutional recurrent neural networks," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 4, pp. 787–799, 2019.

[6] L. Pfeifenberger, M. Zöhrer, and F. Pernkopf, "Eigenvector-based speech mask estimation for multi-channel speech enhancement," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 12, pp. 2162–2172, 2019.

[7] X. Li and R. Horaud, "Multichannel speech enhancement based on time-frequency masking using subband long short-term memory," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2019, pp. 298–302.

[8] A. Bohlender, A. Spriet, W. Tirry, and N. Madhu, "Exploiting temporal context in cnn based multisource DOA estimation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 1594–1608, 2021.

[9] S. Chakrabarty and E. A. P. Habets, "Multi-speaker DOA estimation using deep convolutional networks trained with noise signals," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 1, pp. 8–21, 2019.

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

[11] X. Hao, X. Su, R. Horaud, and X. Li, "Fullsubnet: A full-band and sub-band fusion model for real-time single-channel speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 6633–6637.

[12] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, D. N. L., and Z. V., "TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1," *Linguistic Data Consortium*, 1993.

[13] G. Pirker, M. Wohlmayr, S. Petrik, and F. Pernkopf, "A pitch tracking corpus with evaluation on multipitch tracking scenario," in *Proc. 12th Annu. Conf. Int. Speech Commun. Assoc.*, 2011, pp. 1509–1512.

[14] E. A. P. Habets, "RIR generator," https://github.com/ehabets/RIR-Generator, accessed: May 31, 2021.

[15] miniDSP, "UMA-16 USB microphone array," https://www.minidsp.com/products/usb-audio-interface/uma-16-microphone-array, accessed: May 31, 2021.

[16] E. A. P. Habets and S. Gannot, "Generating sensor signals in isotropic noise fields," *J. Acoust. Soc. Amer.*, vol. 122, no. 6, pp. 3464–3470, 2007.

[17] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Representat.*, 2019, pp. 1–19.

[18] P. Kabal, "TSP speech database," McGill University, Montreal, Quebec, Canada, Tech. Rep., 2002.

[19] European Telecommunications Standards Institute, "Speech processing, transmission and quality aspects (STQ); speech quality performance in the presence of background noise; part 1: Background noise simulation technique and background noise database," ETSI EG 202 396-1, 2008.

[20] M. Taseska and E. A. P. Habets, "MMSE-based source extraction using position-based posterior probabilities," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2013, pp. 664–668.

[21] J. H. DiBiase, "A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays," Ph.D. dissertation, Brown University, Providence, RI, USA, May 2000.

[22] A. Bohlender, A. Spriet, W. Tirry, and N. Madhu, "Least-squares DOA estimation with an informed phase unwrapping and full bandwidth robustness," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 4841–4845.

[23] I. A. McCowan and H. Bourlard, "Microphone array post-filter based on noise field coherence," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 709–716, 2003.

[24] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 4, pp. 1462–1469, 2006.

[25] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 7, pp. 2125–2136, 2011.

[26] International Telecommunication Union, "Wideband extension to recommendation P.862 for the assessment of wideband telephone networks and speech codecs," ITU-T Recommendation P.862.2, 2007.