Narrow-band Deep Filtering for Multichannel Speech Enhancement

Xiaofei Li, Radu Horaud

Narrow-band Deep Filtering for Multichannel Speech Enhancement

Xiaofei Li and Radu Horaud

***Abstract*—In this paper we address the problem of multichannel speech enhancement in the short-time Fourier transform (STFT) domain and in the framework of sequence-to-sequence deep learning. A long short-time memory (LSTM) network takes as input a sequence of STFT coefficients associated with a frequency bin of multichannel noisy-speech signals. The network's output is a sequence of single-channel cleaned speech at the same frequency bin. We propose several clean-speech network targets, namely, the magnitude ratio mask, the complex ideal ratio mask, the STFT coefficients and spatial filtering. A prominent feature of the proposed model is that the same LSTM architecture, with identical parameters, is trained across frequency bins. The proposed method is referred to as narrow-band deep filtering. This choice stays in contrast with traditional wide-band speech enhancement methods. The proposed deep filter is able to discriminate between speech and noise by exploiting their different temporal and spatial characteristics: speech is non-stationary and spatially coherent while noise is relatively stationary and weakly correlated across channels. This is similar in spirit with unsupervised techniques, such as spectral subtraction and beamforming. We describe extensive experiments with both mixed signals (noise is added to clean speech) and real signals (live recordings). We empirically evaluate the proposed architecture variants using speech enhancement and speech recognition metrics, and we compare our results with the results obtained with several state of the art methods. In the light of these experiments we conclude that narrow-band deep filtering has very good performance, and excellent generalization capabilities in terms of speaker variability and noise type.**

***Index Terms*—Speech enhancement, speech denoising, deep fitlering, recurrent neural networks, LSTM.**

I. Introduction

This paper addresses the problem of multichannel speech enhancement/denoising using deep learning. In recent years, speech enhancement based on deep neural networks has been thoroughly and successfully investigated, see [1] for an overview. These methods are often conducted in the time-frequency (TF) domain, and can be broadly categorized into either monaural or multichannel techniques. The monaural techniques use a neural network to map noisy-speech spectral feature onto clean speech targets. The input features, e.g. (logarithm) signal spectra, cepstral coefficients, or linear prediction based features, generally represent the frame-wise full-band spectral structure associated with noisy speech. The target consists of either clean speech spectral features or of ideal binary/ratio masks (IBM/IRM) which are subsequently applied to the noisy-speech input sequence. There is only a handful of methods that process frequency bands separately, e.g. [2], [3], namely a neural network is trained for each subband: these subband spectral features are mapped onto subband targets. Widely used neural architectures for speech enhancement include feed-forward neural networks (FNNs), convolutional neural networks (CNNs) and recurrent neural networks (RNNs). The temporal dynamics of speech can be modeled by stacking context frames in the FNN input, or by dilated CNN [4], while it is automatically modeled by RNNs. In [5], [6], the memory-enhanced RNN, i.e. LSTM, is used to learn the long-term dependencies of signals.

As for multichannel speech enhancement, it is popular to combine supervised monaural approaches with unsupervised beamforming methods, e.g. [7], [8]. The output of the former, i.e. a TF mask, is used to discriminate between speech and noisy TF units, based on which the steering vector of desired speech and noise covariance are computed by the latter. These approaches don't learn the spatial information. To exploit the spatial information, interchannel features (sometimes combined with spectral features), e.g. interaural time-, phase-, and level-difference (ITD, IPD and ILD) and the cross-correlation function (CCF), provide input to a neural network either for full-band TF mask prediction, e.g. [9], [10], [11], or for subband TF mask prediction, e.g. [3], [12]. Due to the use of the interchannel features, these methods are sensitive to the position of the speech source. Therefore, either they consider the position of the speech source to be fixed or to be known, or they are able to discriminate between speech sources. In [13], the magnitude and phase of the short-time Fourier transform (STFT) coefficients of all frequency bands and microphones are directly input to a convolutional recurrent neural network (CRNN), and predict the monaural full-band TF masks, where the convolutional layers extract the inter-channel information and the recurrent layers learn the temporal dynamics. This method is designed to discriminate between the spatial characteristics of directional speech sources and diffuse or uncorrelated sources, i.e. noise, and it is not sensitive to the position of the speech source. In the above multichannel techniques, TF masks serve as a preliminary of a beamformer-based estimator. Even though TF masking is able to improve the speech perceptual quality, it is widely accepted that the signal artifacts created by masking, more specifically by the nonlinear operation of masking, is harmful for automatic speech recognition (ASR). Therefore, beamforming is generally used as an interface between the speech enhancement/separation front-end and the ASR back-end. In [14], skipping the masking step, an FNN is designed to learn the beamformer directly form the time-domain multichannel CCF. In [15], the raw waveform is used as input and

X. Li and R. Horaud are with Inria Grenoble Rhône-Alpes and with Univ. Grenoble Alpes, France.

a number of multichannel convolutional kernels are learned to perform the spatial filtering.

In this work, we propose an LSTM-based multichannel speech denoising method. Unlike the vast majority of existing approaches that perform wide-band speech enhancement, the proposed method processes each STFT frequency bin separately: this is referred to as narrow-band (or frequency-wise) deep filtering. The proposed LSTM training is performed with input and target sequences of noisy- and clean-speech, respectively. Each input is a sequence of multichannel STFT coefficients associated with a single frequency bin. Correspondingly, the target is a sequence of clean speech taken at the same frequency for the reference channel. We propose to train four architecture variants using the following clean-speech targets: the STFT magnitude mask, the STFT complex mask, the STFT coefficients and the spatial filter. Importantly, the network weights are shared across frequency bins, which encourages the network to learn common information accross frequency bins, and also leads to a dramatic reduction in the complexity and computational burden of the training process. Our approach is grounded by the fact that a large number of unsupervised speech enhancement methods exploit frequency-wise narrow-band information. More precisely, the proposed method is motivated on the following grounds:

- The frequency-wise temporal evolution of the STFT magnitude is informative due to the non-stationary nature of speech against the stationarity of noise, which stands at the foundation of unsupervised singlechannel noise power estimation, e.g. [16], [17], as well as multichannel relative transfer function (RTF) estimation [18], [19]. Recently it was demonstrated that a LSTM network is able to accomplish monaural frequency-wise noise power estimation [20];
- The frequency-wise spatial characteristics of the STFT coefficients fully reflect both the directionality of speech and the diffusion of noise. This is the foundation of speech enhancement methods such as the coherent-to-diffuse power ratio method [21] and beamforming techniques [18]. Moreover, the temporal dynamics of frequency-wise spatial correlation contain motion information associated with a speech source;
- The frequency-wise representation is informative for clean-speech estimation; Indeed, singlechannel spectral subtraction (Bayesian filtering) [22], [23], multichannel spatial filtering, e.g. beamforming [18], and multichannel Wiener filtering [24], are performed frequency-wise.

Overall, the proposed LSTM architecture is expected to fully exploit the frequency-wise information, not only by learning a regression from the input sequence to the output sequence, but also by learning a group of functions for clean speech estimation. By sharing the network weights across frequencies, the network is encouraged not to learn the subband spectral structure of siganls, but to learn the narrow-band information mentioned above, and to perform narrow-band deep filtering. The proposed method is similar to [13] in that the network learns how to discriminate between the spatial characteristics of directional speech sources and the diffuse/uncorrelated nature of noise, hence the method is agnostic to the position of the speech source.

Compared to full-band techniques [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], the proposed method ignores cross-band information, and focuses on learning narrow-band information. This has the following advantages: (i) it is questionable whether full-band models are able to learn the narrow-band information mentioned above. As shown below, by focusing on the narrow-band signal representations, the proposed method is able to learn long-term temporal dependencies, e.g. on the order of 150 STFT frames; (ii) due to the reduced dimension of both the input and the output, the proposed network has a smaller number of parameters than full-band models, and hence it requires much less training data and both training and prediction have a lower computational cost; (iii) the proposed method is not sensitive to the wide-band spectral pattern of signals, since it only exploits the narrow-band information. As a result, the proposed network has a very good generalization capability in terms of speaker variability and noise type, and (iv) experiments demonstrate that the enhanced speech obtained with the proposed method can be directly used for ASR, which means the signal artifacts caused by the prediction error of the proposed narrow-band network are not detrimental for ASR. The reason for this will be explained in Section III.

This paper is an extended version of a recently published conference paper [25], in which we proposed a narrow-band LSTM architecture for speech enhancement and we demonstrated its effectiveness when using the magnitude ratio mask as a network target. In this paper we extend this approach by using other possible targets, namely the complex ideal ratio mask and the STFT complex coefficient, as well as a spatial-filtering target. In addition to the experimental validation using mixed data (multichannel noise recordings of various kinds are mixed with multichannel clean speech recordings), we describe and discuss experiments performed with real data (live recordings). We empirically evaluate the proposed architecture variants and compare their performance with several state of the art methods, based on speech enhancement and speech recognition scores.

The remainder of this paper is organized as follows. Section II describes the proposed narrow-band deep filtering model and the proposed LSTM architectures. Section III describes the experimental setup, the LSTM network training characteristics, the speech enhancement and speech recognition metrics that we used, and describes the experiments performed with the mixed and real datasets. Section IV concludes the paper. Supplemental material (examples of processed noisy speech utterances) are available online.[1]

II. Narrow-band Speech Enhancement Networks

Let the multichannel signals be represented in the STFT domain:

$$x_i(k,t) = s_i(k,t) + u_i(k,t), \tag{1}$$

[1]https://team.inria.fr/perception/research/mse-lstm/

where $x_i(k,t)$, $s_i(k,t)$ and $u_i(k,t)$ are the complex-valued STFT coefficients of the microphone, speech and noise signals, respectively, and where $i \in \{1 \dots I\}$, $k \in \{0 \dots K-1\}$ and $t \in \{1 \dots T\}$ denote the channel (microphone), frequency-bin and frame indices, respectively. In this paper the focus is on signal denoising task and hence the reverberation effect is not addressed. Therefore, the speech signals are assumed to be reverberation free, even though we experiment with real-recorded multichannel data that may include some reverberation. The objective is to recover a monaural speech signal, e.g. $s_r(k,t)$, where r denotes the reference channel. In the proposed method and as already mentioned, a single network is trained using the narrow-band sequences over all frequency bins, and the trained network is then used to predict a sequence at each frequency bin. Thence, for the sake of clarity, the frequency-bin index k will be omitted hereafter.

A. Input Features

For each TF bin, the real and imaginary parts, $\mathcal{R}(\cdot)$, $\mathcal{I}(\cdot)$ of the multichannel STFT coefficients are concatenated into the vector:

$$\mathbf{x}(t) = \big(\mathcal{R}(x_1(t)), \mathcal{I}(x_1(t)), \dots, \mathcal{R}(x_I(t)), \mathcal{I}(x_I(t))\big)^\top, \quad (2)$$

where $^\top$ denotes vector transpose. $\mathbf{x}(t) \in \mathbb{R}^{2I}$ contains information associated with one TF bin. The input sequence of LSTM is a temporal sequence of such vectors at each frequency bin, namely:

$$\tilde{\mathbf{X}} = \big(\mathbf{x}(1), \dots, \mathbf{x}(t), \dots, \mathbf{x}(T)\big), \quad (3)$$

where T denotes the number of time steps of the LSTM network. To facilitate network training, the input sequence has to be normalized to equalize the input levels across channels and across time. We empirically set to 1 the STFT magnitude of the reference channel, namely:

$$\begin{cases} \mathbf{X} = \tilde{\mathbf{X}}/\mu \\ \text{with}: \quad \mu = \frac{1}{T}\sum_{t=1}^{T} |x_r(t)|. \end{cases} \quad (4)$$

B. Output Target and Training Loss

As already mentioned, we want to recover the clean speech signal of the reference channel, e.g. $s_r(t)$. To this aim, we test the following network targets.

1) Magnitude Ratio Mask (MRM): For each TF bin, the rectified STFT magnitude ratio mask

$$M(t) = \min\left(\frac{|s_r(t)|}{|x_r(t)|}, 1\right) \quad (5)$$

is the target, where the function $\min(\cdot)$ rectifies the mask to fall in the range $[0, 1]$. For each frequency bin, the target sequence is

$$\mathbf{M} = \big(M(1), \dots, M(t) \dots, M(T)\big). \quad (6)$$

The mean squared error (MSE) of MRM, i.e. $(M(t) - \hat{M}(t))^2$, is taken as the training loss, where $\hat{M}(t)$ denotes the MRM network prediction. At test, the MRM prediction $\hat{M}(t)$ is used to estimate the module of the STFT coefficient while its phase is the phase of the reference channel:

$$|\hat{s}(t)| = \hat{M}(t)|x_r(t)|, \quad (7)$$

$$\arg(\hat{s}(t)) = \arg(x_r(t)) \quad (8)$$

It was demonstrated in [26] that, in the framework of monaural full-band masking, the MRM achieves the best performance among various magnitude-based masks, such as IBM or IRM. Our preliminary experiments within the present framework also demonstrate that this target performs slightly better than IRM.

2) Complex Ideal Ratio Mask: In order to estimate the phase of clean speech, [27] proposed the complex ideal ratio mask (cIRM), defined as the ratio of the STFT coefficients between the expected clean speech and the signal associated with the reference microphone. Indeed, it was shown in [27] that a better performance is achieved by exploiting the phase of the expected clean-speech signal. We tested this target in the framework of our model by exactly following the protocol presented in [27].

3) STFT Complex Coefficient: It was mentioned in [27] that the monaural full-band masking network is not able to directly estimate the clean phase. In this work, we test the capability of the proposed network to directly estimate the STFT complex coefficient (CC) of clean speech. For one TF bin, the real and imaginary parts of $s_r(t)$, i.e.

$$\mathbf{s}(t) = (\mathcal{R}(s_r(t)), \mathcal{I}(s_r(t))) \in \mathbb{R}^2 \quad (9)$$

are directly used as the network target. For each frequency bin, the target sequence is

$$\tilde{\mathbf{S}} = \big(\mathbf{s}(1), \dots, \mathbf{s}(t) \dots, \mathbf{s}(T)\big). \quad (10)$$

According to the input sequence normalization, i.e. (4), the target sequence is also normalized with μ:

$$\mathbf{S} = \tilde{\mathbf{S}}/\mu \quad (11)$$

The training loss is the MSE between the normalized STFT coefficient of clean speech and the STFT predicted by the network, i.e. $\|\mathbf{s}(t)/\mu - \hat{\mathbf{s}}(t)\|^2$. At test, $\mu\hat{\mathbf{s}}(t)$ corresponds the predicted enhanced signal.

4) Spatial Filtering: The combination of TF masking and beamforming techniques often achieve state-of-the-art ASR performance. Beamforming, or spatial filtering (SF), is performed in narrow-band wise, including parameters estimation and filter derivation, which is naturally consistent to the present framework. In this work, we propose to estimate a spatial filter to mimic beamforming-like techniques. Formally, for each TF bin, let the output of the multichannel spatial filter network, $\mathbf{w}(t) \in \mathbb{R}^{2I}$ be defined by:

$$\mathbf{w}(t) = \big(\mathcal{R}(w_1(t)), \mathcal{I}(w_1(t)), \dots, \mathcal{R}(w_I(t)), \mathcal{I}(w_I(t))\big)^\top. \quad (12)$$

For each frequency bin, the output sequence is

$$\mathbf{W} = \big(\mathbf{w}(1), \dots, \mathbf{w}(t) \dots, \mathbf{w}(T)\big). \quad (13)$$

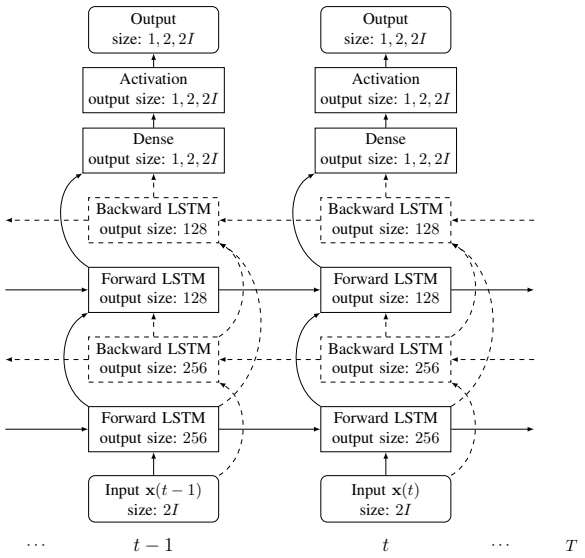


Fig. 1: Diagram of the proposed architecture. The unidirectional (forward) LSTM is represented with solid-lines blocks and arrows, while the additional blocks and arrows needed for BLSTM are represented with dashed lines.

The output is then used to estimate the clean speech,

$$\hat{\mathbf{s}}_{\text{sf}}(t) = (\mathcal{R}(\hat{s}_{\text{sf}}(t)), \mathcal{I}(\hat{s}_{\text{sf}}(t)))^T, \tag{14}$$

by applying the following complex-valued spatial filtering to the input:

$$\mathcal{R}(\hat{s}_{\text{sf}}(t)) = \sum_{i=1}^{I} \big(\mathcal{R}(w_i(t))\mathcal{R}(x_i(t)) - \mathcal{I}(w_i(t))\mathcal{I}(x_i(t))\big),$$

$$\mathcal{I}(\hat{s}_{\text{sf}}(t)) = \sum_{i=1}^{I} \big(\mathcal{R}(w_i(t))\mathcal{I}(x_i(t)) + \mathcal{I}(w_i(t))\mathcal{R}(x_i(t))\big).$$

Rather than imposing a specific beamformer as the training target, we let the network learn to predict an output that minimizes the error between the clean speech and the estimated speech, which is in the same spirit as the multichannel Wiener filter. In practice, the estimated speech (14) is explicitly computed from the network output (12) and input (2). Then, the training loss is the MSE between the estimated speech (14) and the clean-speech target (normalized with μ), namely $|(\mathbf{s}(t) - \hat{\mathbf{s}}_{\text{sf}}(t))/\mu|^2$.

C. Network Architectures

The architectures of the proposed LSTM and bidirectional LSTM (BLSTM) networks are shown on Fig. 1. The sequence-to-sequence scheme is adopted to map the input sequence onto the output sequence. Two LSTM layers are stacked. Through a dense layer, the output vector of the second LSTM layer is mapped onto the output vector. Then an activation is applied to obtain the network output. The output size of LSTM layers are set based on preliminary experiments. Notice that this figure summarizes three networks with three different targets and associated outputs, namely, MRM, CC, and SF. While the input sequence at frequency bin k is the same for all three networks, namely $\mathbf{X}(k)$ defined in (4), the network outputs and the output dimensions are different. The output sequences $\mathbf{M}(k)$, $\mathbf{S}(k)$ and $\mathbf{W}(k)$, defined by (6), (11) and (13), are of dimension 1, 2, and $2I$, respectively.

Moreover, we chose different activation functions for each one of these networks, namely *sigmoid*, *identity* and *tanh*, respectively. We remind that the same network (same parameters) is trained for all the frequency bins $k \in \{0 \dots K-1\}$. The number of parameters to be learned slightly varies with the number of microphones and with the dimension of the output. On an average, the LSTM and BLSTM networks have $470,000$ and $1,200,000$ parameters, respectively.

III. Experiments

A. Experimental Setup

1) Data Generation: We use the CHiME4 dataset [28], which was recorded with six microphones embedded in a tablet device. CHiME4 toolkit provides a method to simulate the multichannel data. However, instead of using the multichannel frequency responses, this method only simulates the multichannel time delays. Our preliminary experiments show that training the network with this type of simulated data performs poorly with real test data. Therefore, we use real data both for training and and for testing purposes. The noise-free multichannel speech data were recorded in a booth (BTH) and the training, development and evaluation data were recorded by three different groups of four speakers. The multichannel background noise were recorded with four noisy environments, namely bus (BUS), cafe (CAF), pedestrian area (PED), and street junction (STR). For each type of noise, four to five sessions were recorded at different times, with a duration of about 0.5 hours per session.

The four speakers in BTH training set (399 utterances) are used for network training, and the eight speakers in BTH development (410 utterances) and evaluation (330 utterances) sets are used for test. Each noise session is split into two sub-sessions used for training (60%) and for test (40%), respectively, which means that different noise instances are used for training and for test. To generate the training data, noise segments randomly extracted from the training sub-sessions are mixed with BTH training utterances, with signal-to-noise-ratios (SNRs) randomly selected from the interval$[-5, 10]$ dB. Each training utterance is mixed with fifteen different randomly selected noise segments, and a total of about 11.3 hours of training data are generated.

Two groups of data are tested, (i) MIXED data: background noise segments randomly extracted from the test sub-sessions are mixed with BTH test utterances, with SNRs in $\{-4, 0, 4, 8\}$ dB. For each noise type and SNR, about 200 test utterances are generated; (ii) REAL data: the development (Dev) and evaluation (Eval) sets from CHiME4 real data were recorded in the four noisy locations by the same speakers in both development and evaluation BTH sets.

The signals are transformed to the STFT domain using a 512-sample (32 ms) Hamming window with a frame step of 256 samples. The sequence length for training is set to $T = 192$ frames (about 3 s), which means the LSTM network is trained to learn 192 time steps of memory. The training sequences are picked out from the utterance-level signals with 50% overlap for two adjacent sequences. In total, about 6.55 million training sequences are generated. For test, the utterances are not cut into sequences with length of 192 frames but, instead, the entire utterances are directly used for sequence-to-sequence prediction.

2) Training Configuration: We found that the microphone #1 recording in the evaluation set has a much larger volume than the volume used in other recording sets. The issue of microphone array mismatch is beyond the scope of this work, thus microphone #1 is not used. Microphone #2 is not used as well, due to its low availability. We conducted experiments with two microphone configurations, i.e. microphones #3, #4, #5 and #6 (4CH), and microphones #5 and #6 (2CH). Microphone #6 is taken as the reference channel. The network variants are named based on the network type, i.e. LSTM or BLSTM, on the output target, i.e. MRM, cIRM, CC or SF, and on the microphone configuration, i.e. 2CH or 4CH. For example, BLSTM-SF-4CH refers to BLSTM with spatial filtering as target and with four microphones as input. All these network variants are trained individually from scratch.

We use the Keras environment [29] to implement the proposed architectures and associated methods. The Adam optimizer [30] is used with a learning rate of 0.001. The batch size is set to 512. The training sequences were shuffled. Based on some preliminary experiments, the BLSTM-CC and -SF are trained with ten epochs, while all the other networks are trained with five epochs.

3) Performance Metrics: To evaluate and benchmark the performance of the proposed speech enhancement methods, three metrics are used, including two intrusive metrics, (i) the perceptual evaluation of speech quality (PESQ) [31] which evaluates the quality of the enhanced signal in terms of both noise reduction and speech distortion, (ii) the short-time objective intelligibility (STOI) [32], a metric that highly correlates with noisy speech intelligibility; and a non-intrusive metric, (iii) the normalized speech-to-reverberation modulation energy ratio (SRMR) [33], which measures the amount of noise, and also reflects the speech intelligibility. For all the metrics, the larger the better. For MIXED data, in order to measure PESQ and STOI, the BTH clean signal is taken as the reference signal. PESQ and STOI are not used for REAL data because the close-talk signals provided in the CHiME4 dataset are not reliable.

For REAL data, in addition to speech enhancement performance, we tested the performance of automatic speech recognition (ASR) obtained with the enhanced signals. The ASR of [34], with already-trained ASR models and decoding recipe provided in CHiME4 is taken as the baseline system.[2] This system uses mel-frequency cepstral coefficients (MFCC), a DNN-HMM acoustic model and an RNN language model. The DNN-HMM acoustic model is trained using the single-channel noisy multi-condition CHiME4 training data. The ASR performance is measured with the word error rate (WER), the lower the better.

4) Comparison with the State of the Art: We compare the proposed methods with three methods, (i) BeamformIt [35], based on an unsupervised filter-and-sum beamforming technique; (ii) the neural network based generalized eigenvalue beamformer (NN-GEV) [7] that uses an BLSTM network to estimate a spectral mask, based on which a generalized-eigenvalue beamformer is computed and applied to speech denoising. We use the toolkit provided by the authors of [7],[3] in which the BLSTM parameters had already been trained using the CHiME4 training dataset, and (iii) the multichannel CRNN method [13] which takes as input multichannel full-band STFT coefficients and which predicts single-channel TF MRMs, i.e. (5). Several CNN layers are employed for each STFT frame to extract the inter-channel information, then followed by one LSTM layer to learn the inter-frame information, where two past frames and two future frames are taken as the context for each frame. Since the authors' implementation is not publicly available, we implemented the method and used the CHiME4 dataset to train, test and evaluate [13]. We use the twelve BTH different speakers from which we generated 9.14 million samples for training this CRNN-based model. We did not evaluate the speaker generalization capability since this was demonstrated in [13].

B. Evaluation of Generalization Capability

The default training setup presented in Section III-A1 is *speaker independent and noise-type dependent* (SID-ND): even though training and test use different noise instances, they both use all four noise types. To evaluate the generalization capability of the proposed network in terms of speaker identity and of noise type, two extra training setups are also tested : (i) *speaker independent and noise-type independent* (SID-NID): four speakers are used for training and the other eight speakers are used for test, and three noise types are used for training and the other noise type is used for test, and (ii) *speaker dependent and noise-type dependent* (SD-ND): all twelve speakers and all four noise types are used to generate training data. In both these cases, 6.55 million sequences were generated.

Fig. 2 shows the speech enhancement results obtained with the MIXED data for these three training configurations: interestingly, they yield comparable PESQ and SRMR scores and slightly different STOI scores. Fig. 3 shows the ASR results obtained with the REAL data for these three training setups. The associated WER scores are also quite similar, except for the Eval PED data for which SD-ND training noticeably outperform SID-NID training. Note that the SRMR

[2]http://spandh.dcs.shef.ac.uk/chime_challenge/chime2016/download.html

[3]https://github.com/fgnt/nn-gev

scores obtained with the REAL data and with the three training setups (not included in the paper) are also quite similar.

These results empirically show that the proposed methods are not too sensitive with respect to speaker identity and noise type, therefore they have good generalization capabilities. The reason for which this doesn't extend to the Eval PED data is not clear. The networks are trained using narrow frequency bands, hence the wide-band spectral-pattern differences between the training and test samples, of both speech and noise, are not taken into account and hence they shouldn't have an impact on the generalization capabilities of the proposed model. The networks are actually trained to learn some functions based on the temporal and spatial characteristics of speech and noise, which are independent with respect to their spectral content. In addition, in the CHiME4 data, the microphone-to-speaker relative positions are time-varying for both the training and test data, which means that the proposed method also generalizes well in terms of moving speakers. Overall, the proposed models learn features that are suitable across frequency bins, as well as for unseen speakers and noise types.

C. Unidirectional versus Bidirectional LSTM

As presented in Section III-A, we perform sequence-to-sequence network training using fixed-length sequences with $T = 192$ frames, which means the back propagation (through time) of gradients is truncated at 192 time steps. In other words, the network is trained to learn 192 time steps of memory. At test, the network predicts length-varying utterances. Utterances with different lengths have different memory lengths, moreover, different time steps in one utterance have

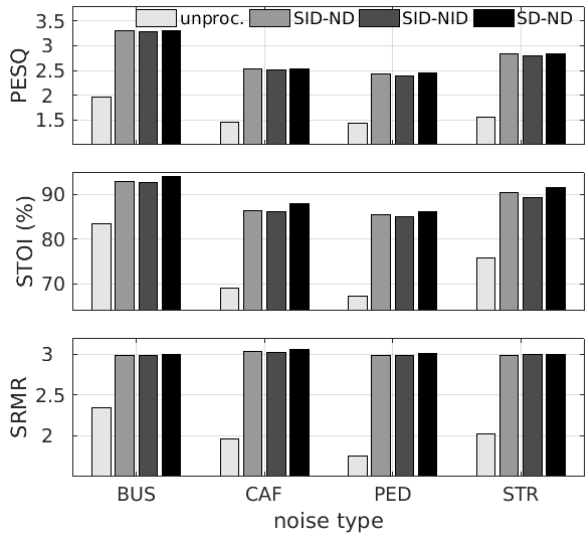


Fig. 2: Speech enhancement results obtained with the MIXED data for the proposed BLSTM-SF-4CH method with three different training setups, speaker independent and noise-type dependent (SID-ND), speaker independent and noise-type independent (SID-NID), and speaker dependent and noise-type dependent (SD-ND). The SNR is of 0 dB.

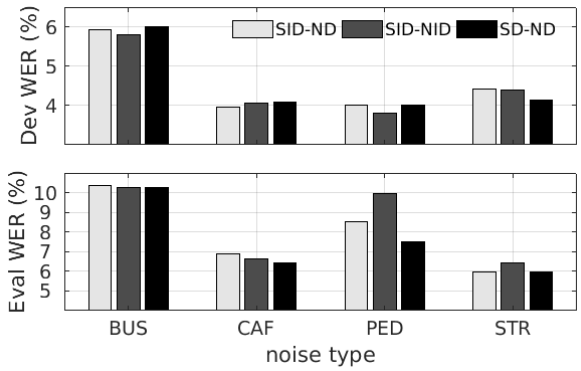


Fig. 3: Speech recognition results obtained with the REAL data for the proposed BLSTM-SF-4CH method and with three different training setups: SID-ND, SID-NID, and SD-ND.

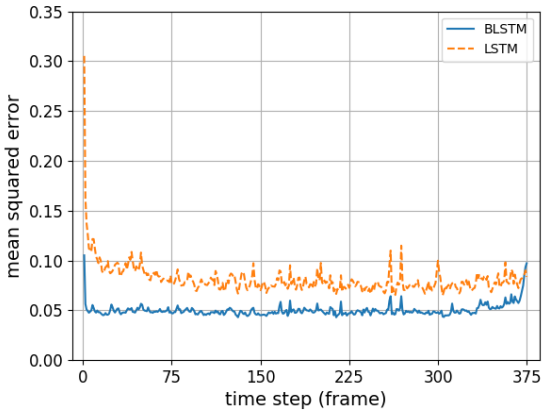


Fig. 4: The loss evolution, i.e. MSE, as a function of time step, for the proposed BLSTM-SF-4CH (blue) and LSTM-SF-4CH (orange) methods.

differrent forward/backward memories. To analyze how the memories work, and how many time steps could be memorized in the proposed narrow-band LSTM framework, Fig. 4 shows the MSEs as a function of time step. To obtain this plot, we generated one extra group of data (we used the same data generation protocol as with the MIXED test data), which includes 1.3 million sequences with a fixed length of $T = 375$ frames (six seconds). The MSEs averaged over all the sequences are shown in Fig. 4. The MSE of LSTM quickly drops from 0.3 to 0.1 in a few time steps, which means a few past frames are already very effective to reduce the loss. The MSE of LSTM then slowly converges to 0.077 in about 150 time steps, which means that, for one time step, the frames earlier than about 150 time steps do not contribute anymore. This is due to one of the following reasons (or the combination of them): (i) the LSTM network is only able to learn the memory of about 150 time steps, and (ii) about 150 time steps already provide enough context information in terms of the temporal and spatial properties of the signal.

When future frames are used, the MSE drops from 0.077 for LSTM to about 0.05 for BLSTM. At the two ends,

TABLE I: Speech enhancement results obtained with the MIXED data. SNR is of 0 dB.

		PESQ ↑					STOI (%) ↑					SRMR ↑				
		BUS	CAF	PED	STR	Average	BUS	CAF	PED	STR	Average	BUS	CAF	PED	STR	Average
	unproc.	1.97	1.47	1.43	1.55	1.61	83.4	69.1	67.2	75.7	73.9	2.34	1.97	1.75	2.03	2.02
	BeamformIt [35]	2.06	1.54	1.51	1.65	1.69	84.6	71.1	70.5	77.0	75.8	2.39	2.03	1.91	2.13	2.16
	NN-GEV [7]	2.15	1.56	1.61	1.73	1.76	87.3	74.6	75.3	81.4	79.7	2.54	2.25	2.15	2.40	2.34
	CRNN [13]	2.59	1.86	1.78	2.07	2.07	89.9	**80.5**	79.4	85.3	83.8	2.91	2.85	2.81	2.87	2.86
2CH	BLSTM-MRM	2.86	2.08	2.05	2.36	2.34	89.9	79.6	78.8	84.7	83.3	2.94	2.93	2.88	2.91	2.91
	BLSTM-cIRM	3.00	2.12	2.10	2.48	2.42	90.6	79.2	79.0	85.0	83.4	**2.99**	3.00	2.97	2.98	2.99
	BLSTM-CC	3.02	**2.16**	2.12	**2.49**	2.44	**91.0**	80.2	**79.6**	**85.5**	**84.1**	**2.99**	**3.03**	**2.99**	2.99	3.00
	BLSTM-SF	**3.03**	**2.16**	**2.13**	**2.49**	**2.45**	90.8	80.2	79.5	**85.5**	84.0	**2.99**	**3.03**	**2.99**	**3.00**	**3.01**
	BeamformIt [35]	2.09	1.59	1.56	1.67	1.73	85.5	73.7	73.1	78.0	77.6	2.37	2.05	1.98	2.10	2.12
	NN-GEV [7]	2.40	1.76	1.78	1.98	1.98	91.0	82.9	83.4	88.9	86.6	2.78	2.62	2.56	2.75	2.68
	CRNN [13]	2.78	2.05	1.95	2.29	2.27	91.7	85.8	84.5	88.8	87.7	2.92	2.91	2.86	2.90	2.90
4CH	BLSTM-MRM	3.10	2.38	2.27	2.63	2.59	91.6	85.2	83.8	88.8	87.3	2.95	2.97	2.91	2.92	2.94
	BLSTM-cIRM	3.29	2.47	2.37	2.83	2.74	92.3	85.3	84.3	89.6	87.9	**2.99**	**3.04**	**2.99**	**2.99**	**3.00**
	BLSTM-CC	3.28	**2.53**	**2.43**	2.84	2.77	**93.0**	86.3	85.3	**90.6**	**88.8**	2.98	**3.04**	**2.99**	2.98	**3.00**
	LSTM-SF	3.07	2.29	2.19	2.64	2.55	91.2	83.0	81.7	88.1	86.0	2.93	2.87	2.82	2.90	2.88
	BLSTM-SF	**3.31**	**2.53**	**2.43**	**2.85**	**2.78**	92.8	**86.4**	**85.4**	90.4	88.7	2.98	**3.04**	**2.99**	**2.99**	**3.00**

BLSTM has a larger MSE due to the insufficient past or future context. At the end part, BLSTM has enough past context. The MSE is reduced from 0.097 at the 375-th frame to 0.06 at the 369-th frame, and to 0.05 at the 350-th frame. This indicates that, when enough past context is being used, about six future frames are already very effective to reduce the loss, and about 25 future frames provide sufficient information to further reduce the loss to a satisfactory value. For an online application, past information is always available. The amount of future frames to be used can be chosen as a trade-off between performance and processing latency: (i) 25 future frames can be used to have the best prediction performance that BLSTM can achieve, which however leads to a 400 ms latency, (ii) 6 future frames can be chosen to have a good performance with 96 ms latency, which is not a problem from a practical point of view.

Tables I and II show the experimental results obtained with the MIXED and REAL data, respectively. Comparing the results of LSTM-SF-4CH and of BLSTM-SF-4CH (the last two rows), one can see that BLSTM achieves, indeed, noticeably better than LSTM in terms of both speech enhancement and speech recognition. A larger error obtained with LSTM than with BLSTM would lead to a larger speech distortion and to less noise reduction. The difference in performance between LSTM and BLSTM can easily be perceived by listening to the enhanced signals.[4] The comparison between LSTM and BLSTM, based on the performance of LSTM-SF-4CH and BLSTM-SF-4CH, also holds for other proposed targets and numbers of channels. Thence, in the following, we will only analyze the performance of BLSTM networks.

D. Results with MIXED Data

Table I shows the speech enhancement results obtained with the MIXED data and with an SNR of 0 dB. It is not surprising that, except for some SRMR scores, the 4CH cases perform better than the 2CH ones, since richer spatial information is available. We will explain the SRMR exceptions later. In the following, we compare the 4CH performance scores (the comparison is equally valid for the 2CH cases).

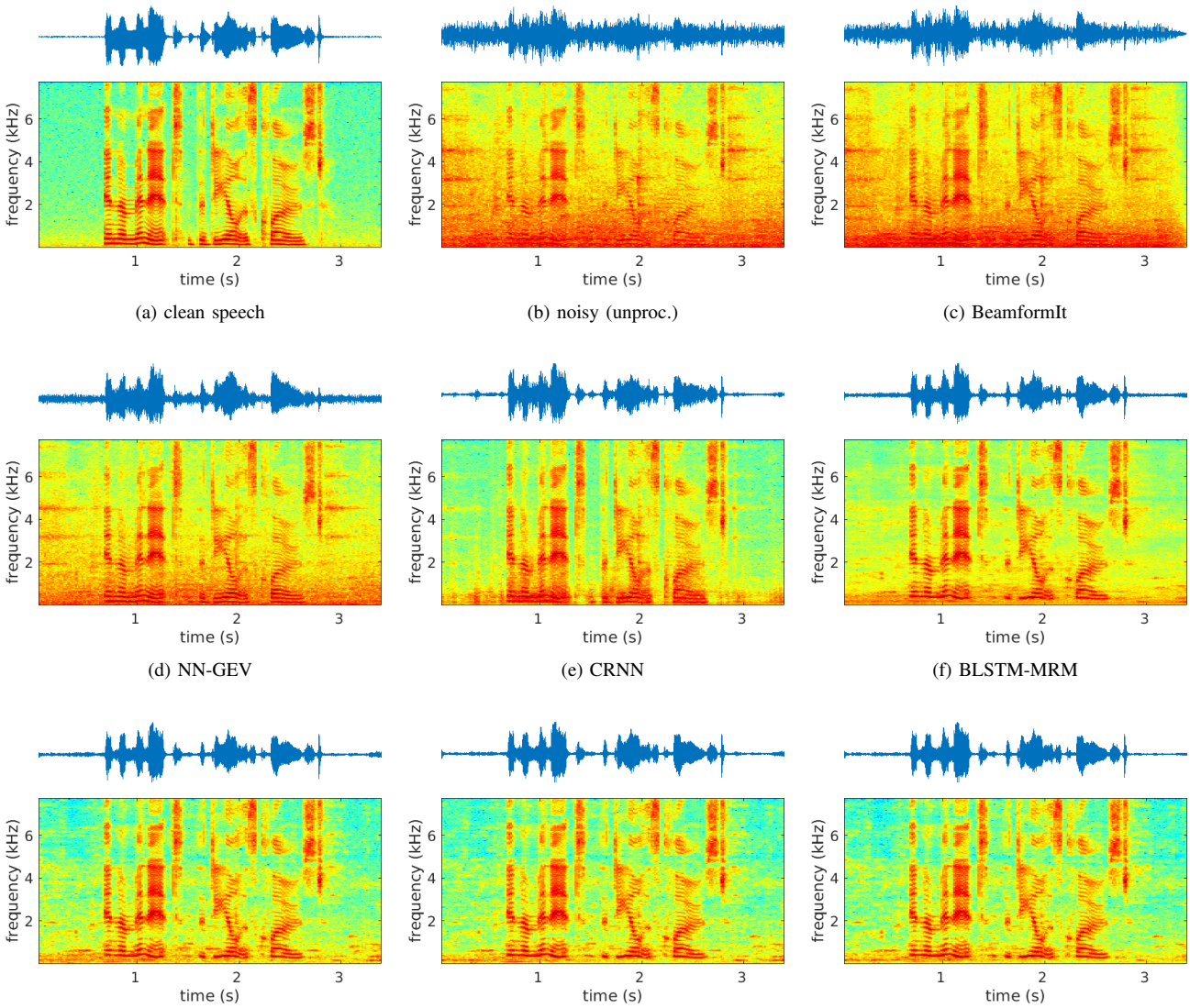
Over the unprocessed signals, BeamformIt improves the three scores to a certain extent. NN-GEV, which uses a deep neural network to classify the speech and noise TF bins, performs much better than Beamformit. It was demonstrated in [7] that the speech enhancement performance of NN-GEV is quite close to the performance of an oracle beamformer. [5] CRNN yields much higher PESQ and SRMR scores and slightly higher STOI scores than NN-GEV. This indicates that, for speech denoising, masking-based methods have a better potential (oracle) than beamforming methods.

BLSTM-MRM and CRNN both predict magnitude ratio masks (MRMs). Compared to CRNN, BLSTM-MRM achieves similar STOI and SRMR scores and better PESQ scores. Better PESQ scores indicate better speech quality. By listening to the enhanced signals, one may notice the dramatic noise reduction of both CRNN and BLSTM-MRM. However, some speech distortions can be heard in the CRNN outputs, while the BLSTM-MRM outputs are less distorted and more natural. The possible reason for this is that when TF masks are predicted for all the frequencies together, some structured prediction errors may lead to audible distortions. By structured prediction errors it is meant that the prediction errors are correlated between frequencies.

By recovering the phase of clean speech, BLSTM-cIRM, BLSTM-CC and BLSTM-SF improve the performance over BLSTM-MRM. Listening to the enhanced signals, more remaining noise can be perceived in the BLSTM-MRM output than in the methods that predict the phase. These results demonstrate that the proposed narrow-band LSTM network is able to predict the complex STFT coefficients of clean speech, since (i) the complex STFT coefficients of multichannel noisy speech are taken as the network input, (ii) the network is able to find the cues to recover the clean phase. We believe that the spatial coherence of directional speech is an important cue that is taken into account by the network. BLSTM-CC

[4]https://team.inria.fr/perception/research/mse-lstm/

[5]The oracle uses the true speech/noise classification for the estimation of the beamforming parameters.



(a) clean speech (b) noisy (unproc.) (c) BeamformIt

(d) NN-GEV (e) CRNN (f) BLSTM-MRM

(g) BLSTM-cIRM (h) BLSTM-CC (i) BLSTM-SF

Fig. 5: Waveforms and spectrograms of the clean-speech input, of the added noise and of the results obtained with state of the art methods and with the proposed BLSTM models, associated with one utterance from the MIXED dataset using four channels (4CH). In this example, CAF noise is added to the clean speech signal and the SNR is of 0 DB.

and BLSTM-SF yield comparable performance, which means that the spatial filtering scheme used in BLSTM-SF does not help to improve the speech enhancement performance. BLSTM-cIRM performs slightly worse than BLSTM-CC and BLSTM-SF, and the performance loss may be caused by the nonlinear transformation of the original cIRM, which is applied to compress the range of the original cIRM. The difference between these three methods are not audible by listening to the enhanced signals. For both the 2CH and 4CH cases, the SRMR scores of BLSTM-cIRM, BLSTM-CC and BLSTM-SF reach the value of clean speech, i.e. about 3.0, which means the remaining noise in the enhanced signals of these methods can not be well measured with SRMR anymore.

Fig. 5 shows waveforms and spectrograms associated with one example. It can be seen that two beamformers (Fig. 5 (c) and (d)), the speech spectra are well preserved, while a large amount of noise still remain, which corresponds to the low speech scores presented in Table I. CRNN (Fig. 5 (e)) largely removes the noise and recovers the speech structure. However, the recovered spectral pattern look somewhat blurred along the frequency axis, which reflects the structured prediction errors mentioned above. This phenomenon can be widely observed with other full-band techniques, e.g. [4]-[13], which indicates that the networks are not fully capable of recovering the details of the (full-band) high-dimensional output vector. In contrast, the proposed narrow-band methods (Fig. 5 (f)-(i)) are able to recover frequency details, due to the untied frequencies and the reliable network prediction. This is consistent to the results of Table I, namely that BLSTM-cIRM, BLSTM-CC and BLSTM-SF perform similarly, and remove more noise than BLSTM-MRM by exploiting the predicted phase. In the very low

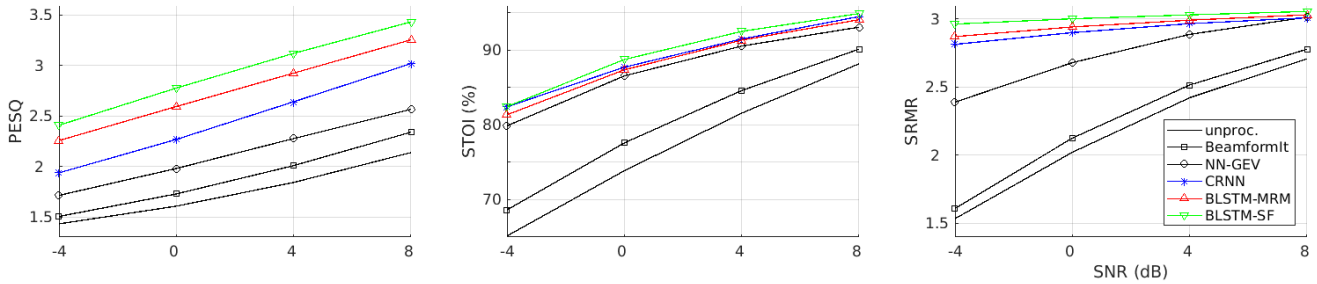


Fig. 6: Speech enhancement results obtained with the MIXED data, averaged over all noise types, as a function of SNR.

TABLE II: Speech enhancement and ASR results obtained with the REAL data, where the SRMR scores are averaged over the developement and evaluation datasets.

		SRMR ↑					WER ↓ (%) Dev					WER ↓ (%) Eval				
		BUS	CAF	PED	STR	Average	BUS	CAF	PED	STR	Average	BUS	CAF	PED	STR	Average
	unproc.	1.75	2.00	2.18	1.97	1.98	14.35	10.55	6.56	10.57	10.51	35.43	23.10	17.71	14.90	22.78
	BeamformIt [35]	1.75	2.11	2.25	2.05	2.04	13.75	7.24	5.07	8.42	8.62	25.56	16.06	13.57	12.25	16.86
	NN-GEV [7]	2.04	2.26	2.38	2.24	2.23	10.84	6.45	**4.45**	7.11	7.21	20.15	12.68	9.79	9.17	12.95
	CRNN [13]	2.57	2.78	2.72	2.70	2.69	-	-	-	-	-	-	-	-	-	-
2CH	BLSTM-MRM	2.82	2.84	2.85	2.80	2.83	**10.56**	**5.43**	4.71	**6.06**	**6.69**	**17.82**	10.14	**9.47**	**7.55**	**11.24**
	BLSTM-cIRM	2.88	2.91	2.91	2.86	2.89	10.86	5.71	5.02	6.55	7.03	19.72	10.70	10.52	7.92	12.21
	BLSTM-CC	2.89	2.93	2.93	**2.88**	2.91	11.93	5.77	4.66	6.15	7.13	20.42	9.84	10.41	7.75	12.10
	BLSTM-SF	**2.92**	**2.95**	**2.94**	**2.88**	**2.92**	11.85	5.93	4.71	6.25	7.18	20.86	**9.77**	9.73	7.84	12.05
	BeamformIt [35]	1.78	2.20	2.32	2.11	2.10	8.95	6.12	4.06	6.81	6.49	18.70	11.49	11.23	10.12	12.88
	NN-GEV [7]	2.36	2.56	2.61	2.52	2.51	**5.12**	**3.91**	**3.44**	**4.16**	**4.16**	10.92	**6.39**	**7.10**	6.71	**7.78**
	CRNN [13]	2.64	2.81	2.76	2.74	2.74	-	-	-	-	-	-	-	-	-	-
4CH	BLSTM-MRM	2.81	2.87	2.81	2.80	2.82	7.42	4.16	4.12	4.65	5.08	**10.23**	6.50	11.08	7.96	8.94
	BLSTM-cIRM	2.85	2.92	2.87	2.84	2.87	6.40	4.19	4.07	4.84	4.87	10.66	6.99	8.26	6.69	8.15
	BLSTM-CC	2.88	**2.93**	2.87	2.85	2.88	7.21	4.01	4.09	4.70	5.00	11.50	7.13	10.74	6.78	9.04
	LSTM-SF	2.71	2.78	2.76	2.74	2.75	6.42	4.20	4.32	5.06	5.00	11.31	6.95	9.83	6.91	8.75
	BLSTM-SF	**2.89**	**2.93**	**2.89**	**2.86**	**2.89**	5.92	3.97	4.01	4.42	4.58	10.38	6.89	8.54	**5.96**	7.94

frequency region, the proposed methods failed to properly predict the speech spectra due to the very low SNR in this region. For this case, CRNN works well by predicting all the frequencies together. Results obtained with other SNR values are shown in Fig. 6. For the sake of clarity of illustration, the curves of BLSTM-cIRM and of BLSTM-CC are not shown as they are very close to the BLSTM-SF curves. It can be seen that the conclusions drawn above hold for a wide range of SNR values.

E. Results with REAL Data

Table II shows the speech enhancement and speech recognition scores obtained with the REAL data. The proposed methods largely improve the SRMR scores over the unprocessed signals, which means that the proposed networks that are trained with mixed signals generalize well to signals recorded in real situations and for which the ground-truth clean-speech signals are not available.

BeamformIt considerably reduces the WER scores obtained with unprocessed signals. For beamforming techniques, the beam pattern of a microphone array is highly corrrelated to the number of microphones, i.e. larger the better. Thence, with only two microphones, NN-GEV does not yield a good beam pattern, and performs worse than the proposed methods. For the 2CH case, the WER scores of BLSTM-cIRM, BLSTM-CC and BLSTM-SF are comparable, and it is surprising that the ASR performance of BLSTM-MRM is considerably higher. It is not clear why BLSTM-MRM yields better (smaller) WER scores than the other proposed variants.

For the 4CH case, NN-GEV performs the best. The WER scores obtained with BLSTM-SF are close to the ones obtained with NN-GEV, and are better than the ones obtained with BLSTM-MRM, BLSTM-cIRM and BLSTM-CC. These results testify that, when a large number of microphones are used, linear spatial filtering, e.g. NN-GEV and BLSTM-SF, is the method of choice in conjunction with speech recognition.

The results and experiments that were just described provide empirical evidence that the proposed narrow-band TF masking methods are well suited to enhance the speech signals prior to speech recognition tasks. The belief that the nonlinear masking process is harmful in the case of speech recognition may be due to the presence of structured signal artifacts associated with full-band speech enhancement methods.

In this work, the reverberation effect is not taken into account. The training speech, i.e. BTH speech, is inconsistent to the REAL test speech in which reverberation presents, especially in the BUS environments. However, the network still preforms quite well. In this experiment, it is infeasible to evaluate how the network treats reverberation due to the lack of reference clean speech. The dereverberation topic will be studied in the future.

IV. Conclusions

In this paper we proposed a narrow-band deep filtering method to address the problem of multichannel speech enhancement. Unsupervised methods, such as spectral subtraction or spatial filtering, have shown some advantages of narrow-band processing for discriminating between speech and noise. The proposed LSTM-based method is able to exploit rich narrow-band features and it outperforms the methods mentioned above. Interestingly, narrow-band LSTM preserves one of the most prominent merits of unsupervised models, namely it is agnostic to speaker identity and to noise type.

Four targets were used for training: the magnitude ratio mask (MRM), the complex ideal ratio mask (cIRM), the complex coefficients (CC) and the spatial filter (SF). We empirically evaluated the merits of these targets and their corresponding architecture variants, using both speech enhancement and speech recognition scores, namely STOI, PESQ, SRMR, and WER. In terms of speech enhancement, cIRM-, CC- and SF-based networks outperform the MRM-based network in terms of speech enhancement. Using only two microphones (2CH) MRM yields the best speech recognition scores, while in the case of four microphones (4CH) MRM is outperformed by the other models. The best WER scores are obtained with four microphones and with the spatial-filtering BLSTM network. Compared with the state-of-the-art methods, the proposed architectures yield the best speech enhancement results, while the speech recognition results are comparable with the results obtained by NN-GEV [7].

It is interesting to note that by ignoring wide-band spectral and spatial patterns, the proposed model has several merits: there is a dramatic reduction in both the number of network parameters and in the size of the training dataset, thus considerably reducing the computational burden, it has excellent generalization capabilities, and it avoids structured signal artifacts. It is however true that wide-band patterns contain interesting features that are not used with narrow-band models and which are worth to be included in order to further improve the performance of the proposed model. Therefore, it would be interesting to investigate new architectures that can incorporate wide-band spectral and spatial features while preserving the advantages of narrow-band models, most notably their excellent generalization capabilities and their robustness against signal artifacts that correlate across the spectrum.

References

[1] D. Wang and J. 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.

[2] Y. Wang, K. Han, and D. Wang, "Exploring monaural features for classification-based speech segregation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 2, pp. 270–279, 2013.

[3] Y. Jiang, D. Wang, R. Liu, and Z. Feng, "Binaural classification for reverberant speech segregation using deep neural networks," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 22, no. 12, pp. 2112–2121, 2014.

[4] K. Tan, J. Chen, and D. Wang, "Gated residual networks with dilated convolutions for monaural speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 1, pp. 189–198, 2019.

[5] F. Weninger, F. Eyben, and B. Schuller, "Single-channel speech separation with memory-enhanced recurrent neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 3709–3713.

[6] J. Chen and D. Wang, "Long short-term memory for speaker generalization in supervised speech separation," *The Journal of the Acoustical Society of America*, vol. 141, no. 6, pp. 4705–4714, 2017.

[7] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 196–200.

[8] C. Boeddeker, H. Erdogan, T. Yoshioka, and R. Haeb-Umbach, "Exploring practical aspects of neural mask-based beamforming for far-field speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 6697–6701.

[9] X. Zhang and D. Wang, "Deep learning based binaural speech separation in reverberant environments," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 25, no. 5, pp. 1075–1084, 2017.

[10] Z.-Q. Wang and D. Wang, "Combining spectral and spatial features for deep learning based blind speaker separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 2, pp. 457–468, 2019.

[11] T. Yoshioka, Z. Chen, C. Liu, X. Xiao, H. Erdogan, and D. Dimitriadis, "Low-latency speaker-independent continuous speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6980–6984.

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

[13] S. Chakrabarty and E. A. Habets, "Time-frequency masking based online multi-channel speech enhancement with convolutional recurrent neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 787–799, 2019.

[14] X. Xiao, S. Watanabe, H. Erdogan, L. Lu, J. Hershey, M. L. Seltzer, G. Chen, Y. Zhang, M. Mandel, and D. Yu, "Deep beamforming networks for multi-channel speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5745–5749.

[15] T. N. Sainath, R. J. Weiss, K. W. Wilson, B. Li, A. Narayanan, E. Variani, M. Bacchiani, I. Shafran, A. Senior, K. Chin *et al.*, "Multichannel signal processing with deep neural networks for automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 5, pp. 965–979, 2017.

[16] X. Li, L. Girin, S. Gannot, and R. Horaud, "Non-stationary noise power spectral density estimation based on regional statistics," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016, pp. 181–185.

[17] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1383–1393, 2012.

[18] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *Signal Processing, IEEE Transactions on*, vol. 49, no. 8, pp. 1614–1626, 2001.

[19] X. Li, L. Girin, R. Horaud, and S. Gannot, "Estimation of relative transfer function in the presence of stationary noise based on segmental power spectral density matrix subtraction," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015, pp. 320–324.

[20] X. Li, S. Leglaive, L. Girin, and R. Horaud, "Audio-noise power spectral density estimation using long short-term memory," *IEEE Signal Processing Letters*, 2019.

[21] A. Schwarz and W. Kellermann, "Coherent-to-diffuse power ratio estimation for dereverberation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 6, pp. 1006–1018, 2015.

[22] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[23] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal processing*, vol. 81, no. 11, pp. 2403–2418, 2001.

[24] M. Brandstein and D. Ward, *Microphone arrays: signal processing techniques and applications*. Springer Science & Business Media, 2013.

[25] X. Li and R. Horaud, "Multichannel speech enhancement based on time-frequency masking using subband long short-term memory," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2019.

[26] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1849–1858, 2014.

[27] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 3, pp. 483–492, 2016.

[28] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015, pp. 504–511.

[29] F. Chollet *et al.*, "Keras," https://github.com/fchollet/keras, 2015.

[30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations*, 2015.

[31] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, 2001, pp. 749–752.

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

[33] J. F. Santos, M. Senoussaoui, and T. H. Falk, "An improved non-intrusive intelligibility metric for noisy and reverberant speech," in *International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE, 2014, pp. 55–59.

[34] T. Hori, Z. Chen, J. R. Hershey, J. Le Roux, V. Mitra, and S. Watanabe, "The MERL/SRI system for the 3RD CHiME challenge using beamforming, robust feature extraction, and advanced speech recognition," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 475–481.

[35] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2011–2022, 2007.