ENHANCEMENT OF SPATIAL CLUSTERING-BASED TIME-FREQUENCY MASKS USING LSTM NEURAL NETWORKS

Felix Grezes$^1$, Zhaoheng Ni$^1$, Viet Anh Trinh$^1$, Michael Mandel$^{12}$

$^1$The Graduate Center, City University of New York
$^2$Brooklyn College, City University of New York
{fgrezes,zni,vtrinh}@gradcenter.cuny.edu, mim@sci.brooklyn.cuny.edu

ABSTRACT
Recent works have shown that Deep Recurrent Neural Networks using the LSTM architecture can achieve strong single-channel speech enhancement by estimating time-frequency masks. However, these models do not naturally generalize to multi-channel inputs from varying microphone configurations. In contrast, spatial clustering techniques can achieve such generalization but lack a strong signal model. Our work proposes a combination of the two approaches. By using LSTMs to enhance spatial clustering based time-frequency masks, we achieve both the signal modeling performance of multiple single-channel LSTM-DNN speech enhancers and the signal separation performance and generality of multi-channel spatial clustering. We compare our proposed system to several baselines on the CHiME-3 dataset. We evaluate the quality of the audio from each system using SDR from the BSS Eval toolkit and PESQ. We evaluate the intelligibility of the output of each system using word error rate from a Kaldi automatic speech recognizer.

Index Terms— Speech Enhancement, Microphone Array, LSTM, Spatial Clustering, Beamforming

1. INTRODUCTION
With speech recognition techniques approaching human performance on noise-free audio with a close-talking microphone [1], recent research has focused on the more difficult task of speech recognition in far-field, noisy environments. This task requires robust speech enhancement capabilities.

One approach to speech enhancement is spatial clustering, which groups together spectrogram points coming from the same spatial location [2]. This information can be used to drive beamforming, which linearly combines multiple microphone channels into an estimate of the original signal that is optimal under some test-time criterion [3]. This optimality is typically based on properties of the signals or the spatial configuration of the recordings at test time, with no training ahead of time.

Another approach is to use a signal models trained using neural networks. Recent work on deep recurrent neural networks using the LSTM architecture [4] can achieve significant single-channel noise reduction [5, 6], and so there is interest in using trainable deep-learning models to perform beamforming. This is especially useful for optimizing beamformers directly for automatic speech recognition [7, 8], although such optimization must happen at training time on a large corpus of training data. Such models have difficulty generalizing across microphone arrays, including differences in number of microphones and array geometries, such as occurs between the AMI corpus [9, 10] and the CHIME challenge [11]. In contrast to deep learning-based beamforming, spatial clustering is an unsupervised method for performing source separation, so it easily adapts across microphone arrays [12, 13, 14]. Such methods group spectrogram points based on similarities in spatial properties, but are typically not able to take advantage of signal models, such as models of speech or noise.

Developed by Mandel et al. [12], Model-based EM Source Separation and Localization (MESSL) is a system that computes time-frequency spectrogram masks for source separation as a byproduct of estimating the spatial location of the sources. It does so using the expectation maximization (EM) algorithm, iteratively refining the estimates of the spatial parameters of the audio sources and the spectrogram regions dominated by each source.

While MESSL utilizes spatial information to separate multiple sources, it does not model the content of the original signals. This is an advantage when separating unknown sources, but performance can be improved when a model of the target source is available. The goal of this paper is to augment the capabilities of MESSL by adding a speech signal model based on neural networks trained to enhance the masks produced by MESSL.

In this paper we describe a novel method of combining single-channel LSTM-RNN-based speech enhancement into MESSL. We train a distinct LSTM model that uses the single-channel noisy audio to enhance the masks produced by MESSL. To show how these methods enhance the speech of the CHiME-3 outdoor 6-channel audio, compared to baselines, we report the enhancement performance measured by PESQ score [15], the SDR, SIR, and SAR scores from BSS Eval toolkit [16], as well as the WER as reported by the Kaldi toolkit [17] trained on a separate corpus, the indoor 8-channel AMI corpus.

2. RELATED WORK
Recently, Nugraha et al. [18] also studied multi-channel source separation using deep feedforward neural networks, using a multi-channel Gaussian model to combine the source spectrograms, to take advantage of the spatial information present in the microphone array. They explore the efficacy of different loss functions and other model hyper-parameters. One of their findings is that the standard mean-square error loss function performed close to the best. In contrast to our work they do not use spatial information that beamforming can give.

Pfeifenberger et al. [19] proposed an optimal multi-channel filter which relies solely on speech presence probability. This speech-noise mask is predicted using a 2-layer feedforward neural network using features based on the leading eigenvector of the spatial covariance matrix of short time segments. Using a single eigenvector
Table 1. Training Targets and their Associated Loss Function

<table>
<thead>
<tr>
<th>Training Targets</th>
<th>Loss Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ideal Amplitude (IA) Masks</td>
<td>$m_{ia}(\omega, t) =</td>
</tr>
<tr>
<td>Phase Sensitive (PA) Masks</td>
<td>$m_{ps}(\omega, t) = \cos(\theta_{\omega, t})$</td>
</tr>
<tr>
<td>Magnitude Spectrum (MS) Approximation</td>
<td>$m_{ma}(\omega, t) =</td>
</tr>
<tr>
<td>Phase-sensitive Spectrum (PS) Approximation</td>
<td>$m_{ps}(\omega, t) = \cos(\theta_{\omega, t})s(\omega, t)$</td>
</tr>
</tbody>
</table>

Fig. 1. Multi-channel Spatial Clustering Based Time-Frequency Mask Enhancement System

Heymann et al [20, 21] also study the combination of multi-channel beamforming with single-channel neural network model. Similar to ours, the proposed model consists of a bidirectional LSTM layer, followed by feedforward layers, in their case three. Of particular note is the companion paper by Boeddeker et al. [22], which derives the derivative of an eigenvalue problem involving complex-valued eigenvectors, allowing their system to propagate errors in the final SNR through the beamforming and back to the single-channel DNNs. While we do not optimize our system in this end-to-end manner, the combination of MESSL with the per-channel DNNs may provide advantages in modeling the spatial information.

This paper builds upon previous work by the authors in [23] and [24], in which we propose two other methods of improving MESSL: a naive combination of the MESSL mask with the masks produced by a LSTM network trained to enhance noisy spectrograms, and using the LSTM-based masks to initialize the EM algorithm of MESSL. This previous work also describes a novel supervised-mvdr beamforming technique to obtain cleaner references for the CHiME-3 dataset.

3. METHODS

3.1. Training the Networks to Enhance the MESSL Masks

To improve the quality of the binary masks produced by MESSL, we trained LSTM neural networks to enhance a MESSL mask when passing this mask and its associated noisy spectrogram through the network. We tested four different training targets: ideal amplitude (IA) masks, phase sensitive (PA) masks, magnitude spectrum (MS) approximations and phase-sensitive (PS) spectrum approximations , based on work by Erdogan et al. [25], as shown in Table 1.

The LSTM operates on single-channel recordings. Each channel in the multi-channel recording is processed independently and in parallel by the LSTM, following [26]. In the single-channel setting, the short-time Fourier transform of the recorded noisy signal, $y(\omega, t)$ is assumed to be

$$y(\omega, t) = s(\omega, t) + n(\omega, t)$$  \hspace{1cm} (1)

where $s(\omega, t)$ is the (possibly reverberant) target speech and $n(\omega, t)$ makes the input to the DNN independent of the number of microphones, and thus adaptable to new microphone configurations. It is trained on the simulated noisy data portion of CHIME-3. They show that this filter improves the PESQ score of the audio. This approach uses an early fusion of the microphone channels before they are processed by the DNN, as opposed to our late fusion after the DNN processes each channel.

This previous work also describes a novel supervised-mvdr beamforming technique to obtain cleaner references for the CHiME-3 dataset.
is non-stationary additive noise. For the purposes of defining the targets and cost functions in Table 1, let
\[
\theta_{\omega,t} = \angle s(\omega,t) - \angle g(\omega,t)
\]
i.e., the phase difference between the target clean spectrogram and the input noisy spectrogram. In each case the network was configured to output a [0, 1] valued mask \(\hat{m}(\omega,t)\) for each frame of the input noisy spectrogram.

For the masks targets, the network was trained to minimize the binary cross-entropy loss, while for the spectrum approximations targets the network was trained to minimize the mean-squared error. While in theory phase-sensitive masks may have negative values, causing problems with the cross-entropy loss function, in practice these were rare enough that we simply clipped those values to be 0.

For each training target type, we explored various hyper-parameter combinations: single or double bi-directional LSTM layers of size 256, 512, 1024 or 2048; merging of the bi-directional forward and backward outputs by summing, multiplying, averaging or concatenating; using a sigmoid or hard sigmoid (a piece-wise linear approximation of sigmoid that is faster to compute) for the output layer activation function. The exploration was done by randomly generating a network from the above 64 combinations and training it until the loss on the development set no-longer improved. For each training type, we report our best configuration in Table 2.

The spectrogram inputs were converted from a linear to decibel scale, and normalized to mean 0 and variance 1 at each frequency bin. The MESSL binary masks were passed through the logit function. To perform the computation and training of our LSTM neural networks, we used the KERAS python library [27], built upon the Tensorflow library [28].

Figure 2 gives an example of how one of our networks has learned how to use the noisy spectrogram to refine a mask produced by MESSL.

### 4. EXPERIMENTS AND RESULTS

#### 4.1. The CHiME-3 Corpus

The CHiME-3 corpus features both live and simulated, 6-channel single speaker recordings from 12 different speakers (6 male, 6 female), in 4 different noisy environments: café, street junction, public transport and pedestrian area. In our work, we used the official data split, with 1600 real noisy utterances in the training set for training, 1640 real noisy utterances in the development set for validation. We did not use the simulated data to train our models. We tested our models on the proposed 2640 utterances in the test set, which contains audio both from real noisy recordings and simulated noisy recordings. In order to perform speech recognition, we used the Kaldi toolkit trained on the AMI corpus, which features 8 microphones, recording overlapping speech in meeting rooms. These differences provide an additional challenge, but are essential to evaluating the generalization abilities of our model.

#### 4.2. Supervised MVDR Speech Reference

Because the real subset of the CHiME-3 recordings were spoken in a noisy environment, it is not possible to provide a true clean reference signal for them. Instead, an additional microphone was placed close to the talker’s mouth to serve as a reference. While this reference has a higher signal-to-noise ratio than the main microphones, it is not noise free. In addition, because it is mounted close to the mouth, it contains sounds that are not desired in a clean output and actually could hurt ASR performance, namely pops, lip smacks, and other mouth noises. In order to obtain a cleaner reference signal, we use the close mic signal as a frequency-dependent voice activity detector to control the MVDR beamforming of the signals from the array microphones, as described in [24].

#### 4.3. Evaluation Metrics

We evaluate the performance of our enhancement system in terms of both speech quality and intelligibility to a speech recognizer. For quality, we use the Signal-to-Distortion Ratio (SDR) from the BSS_EVAL toolkit [16] and the Perceptual Evaluation of Speech Quality (PESQ) score [29]. PESQ is measured in units of mean opinion score (MOS) between 0 and 5, higher being better. For SDR, we used the source-based (not spatial image-based) scoring. For the simulated data, the reference signals were given by the booth.
In this paper we propose a novel method to adapt parallel single-channel LSTM-based enhancement to multi-channel audio, combining the speech-signal modeling power of the LSTM neural network with the spatial clustering power of MESSL, further enhancing the audio. We show that this method can help MESSL improve the quality of audio, with similar intelligibility.

Our future work will continue to explore different ways of integrating the LSTM speech-signal model with MESSL. Preliminary results show that the spatial information is more valuable than the single-channel speech information with respect to WER. To further test this hypothesis, the next step is to integrate a mask cleaning LSTM model in each loop of MESSL’s EM algorithm, i.e., use the mask enhancer model to clean the MESSL masks before the estimation of the spatial parameters.

5. CONCLUSION AND FUTURE WORK

6. ACKNOWLEDGEMENTS

This material is based upon work supported by the Alfred P Sloan foundation and the National Science Foundation (NSF) under Grant No. IIS-1409431. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF.
7. REFERENCES


