MIMO-Speech: End-to-End Multi-Channel Multi-Speaker Speech Recognition

Chang, Xuankai; Zhang, Wangyou; Qian, Yanmin; Le Roux, Jonathan; Watanabe, Shinji

TR2019-157   December 19, 2019

Abstract

Recently, the end-to-end approach has proven its efficacy in monaural multi-speaker speech recognition. However, high word error rates (WERs) still prevent these systems from being used in practical applications. On the other hand, the spatial information in multi-channel signals has proven helpful in far-field speech recognition tasks. In this work, we propose a novel neural sequence-tosequence (seq2seq) architecture, MIMO-Speech, which extends the original seq2seq to deal with multi-channel input and multi-channel output so that it can fully model multi-channel multi-speaker speech separation and recognition. MIMO-Speech is a fully neural end-toend framework, which is optimized only via an ASR criterion. It is comprised of: 1) a monaural masking network, 2) a multi-source neural beamformer, and 3) a multi-output speech recognition model. With this processing, the input overlapped speech is directly mapped to text sequences. We further adopted a curriculum learning strategy, making the best use of the training set to improve the performance. The experiments on the spatialized wsj1-2mix corpus show that our model can achieve more than 60% WER reduction compared to the single-channel system with high quality enhanced signals (SI-SDR = 23.1 dB) obtained by the above separation function

IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)

MIMO-SPEECH: END-TO-END MULTI-CHANNEL MULTI-SPEAKER SPEECH RECOGNITION

Xuankai Chang[1,2], Wangyou Zhang[2], Yanmin Qian[2†], Jonathan Le Roux[3], Shinji Watanabe[1†]

[1]Center for Language and Speech Processing, Johns Hopkins University, USA
[2]SpeechLab, Department of Computer Science and Engineering, Shanghai Jiao Tong University, China
[3]Mitsubishi Electric Research Laboratories (MERL), USA

ABSTRACT

Recently, the end-to-end approach has proven its efficacy in monaural multi-speaker speech recognition. However, high word error rates (WERs) still prevent these systems from being used in practical applications. On the other hand, the spatial information in multi-channel signals has proven helpful in far-field speech recognition tasks. In this work, we propose a novel neural sequence-to-sequence (seq2seq) architecture, MIMO-Speech, which extends the original seq2seq to deal with multi-channel input and multi-channel output so that it can fully model multi-channel multi-speaker speech separation and recognition. MIMO-Speech is a fully neural end-to-end framework, which is optimized only via an ASR criterion. It is comprised of: 1) a monaural masking network, 2) a multi-source neural beamformer, and 3) a multi-output speech recognition model. With this processing, the input overlapped speech is directly mapped to text sequences. We further adopted a curriculum learning strategy, making the best use of the training set to improve the performance. The experiments on the spatialized wsj1-2mix corpus show that our model can achieve more than 60% WER reduction compared to the single-channel system with high quality enhanced signals (SI-SDR = 23.1 dB) obtained by the above separation function.

Index Terms— Overlapped speech recognition, end-to-end, neural beamforming, speech separation, curriculum learning.

1. INTRODUCTION

The cocktail party problem, where the speech of a target speaker is entangled with noise or speech of interfering speakers, has been a challenging problem in speech processing for more than 60 years [1]. In recent years, there have been many research efforts based on deep learning addressing the multi-speaker speech separation and recognition problems. These works can be categorized into two classes depending on the type of input signals, namely single-channel and multi-channel.

In the single-channel multi-speaker speech separation and recognition tasks, several techniques have been proposed, achieving significant progress. One such technique is deep clustering (DPCL) [2–4]. In DPCL, a neural network is trained to map each time-frequency unit to an embedding vector, which is used to assign each unit to a source by a clustering algorithm afterwards. DPCL was then integrated into a joint training framework with end-to-end speech recognition in [5], showing promising performance. Another approach called permutation-free training [2, 3] or permutation-invariant training (PIT) [6, 7] relies on training a neural network to

estimate a mask for every speaker with a permutation-free objective function that minimizes the reconstruction loss. PIT was later applied to multi-speaker automatic speech recognition (ASR) by directly optimizing a speech recognition loss [8, 9] within a DNN-HMM hybrid ASR framework. In recent years, end-to-end models have drawn a lot of attention in single-speaker ASR systems and shown great success [10–13]. These models have simplified the ASR paradigm by unifying acoustic, language, and phonetic models into a single neural network. In [14, 15], joint CTC/attention-based encoder-decoder [12] end-to-end models were developed to solve the single-channel multi-speaker speech recognition problem, where the encoder separates the mixed speech features and the attention-based decoder generates the output sequences. Although significant performance improvements have been achieved in the monaural case, there is still a large performance gap compared with that of single-speaker speech recognition systems, making such models not yet ready for widespread application in real scenarios.

The other important case is that of multi-channel multi-speaker speech separation and recognition, where the input signals are collected by microphone arrays. Acquiring multi-channel data is not so limiting nowadays, where microphone arrays are widely deployed in many devices. When multi-channel data is available, the spatial information can be exploited to determine the speaker location and to separate the speech with higher accuracy. Yoshioka et al [16] proposed a method for performing multi-channel speech separation under the PIT framework. A mask-based beamformer called the unmixing transducer was used to separate the overlapped speech. Another method proposed by Wang et al [17] leverages the inter-channel differences as spatial features combined with the single-channel spectral features as the input, to separate the multi-channel data using the DPCL technique.

Previous works based on multi-channel multi-speaker input mainly focus on separation. In this paper, we propose an end-to-end multi-channel multi-speaker speech recognition system. Such a sequence-to-sequence model is trained to directly map multi-channel input (MI) speech signals where multiple speakers speak simultaneously, to multiple output (MO) text sequences, one for each speaker. We refer to this system as MIMO-Speech. The recent research on single-speaker far-field speech recognition has shown that neural beamforming techniques for denoising [18, 19] can achieve state-of-the art results in robust ASR tasks [20–22]. Several works have shown that it is feasible to design a totally differentiable end-to-end model by integrating the neural beamforming mechanism and the sequence-to-sequence speech recognition together [23–26]. [27] further shows that the neural beamforming function in a multi-channel end-to-end system can enhance the signals. In light of this success, we redesigned the neural-beamformer front-end to allow it to attend

---

to multiple beams at different directions. After getting the separated signals, the log filter bank features are extracted inside the neural network. Finally, a joint CTC/attention-based encoder-decoder recognizes each feature stream. With this framework, the outputs of the beamformer in the middle of the model can also be used as speech separation signals. During the training, a data scheduling strategy using curriculum learning is specially designed and leads to an additional performance boost. To prove the basic concept of our method, we first evaluated our proposed method in the anechoic scenario. From the results, we find that even without explicitly optimizing for separation, the intermediate signals after the beamformer still show very good quality in terms of audibility. Then we also tested the model on the reverberant case to give a preliminary result.

2. END-TO-END MULTI-CHANNEL MULTI-SPEAKER ASR

In this section, we first present the proposed end-to-end multi-channel multi-speaker speech recognition model, which is shown in Fig. 1. We then describe the techniques applied in scheduling the training data, which have an important role in improving the performance.
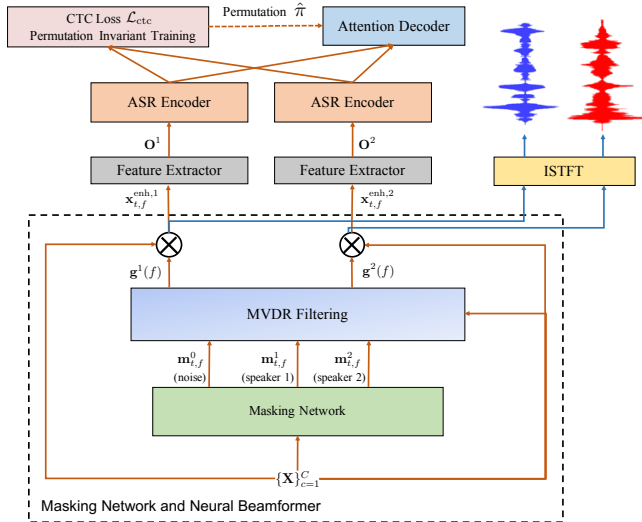

Fig. 1. End-to-End Multi-channel Multi-speaker Model

2.1. Model Architecture

By using the differences in the signals recorded at each sensor, distributed sensors can exploit spatial information. They are thus particularly useful for separating sources that are spatially partitioned. In this work, we present a sequence-to-sequence architecture with multi-channel input and multi-channel output to model the multi-channel multi-speaker speech recognition, shown in Fig. 1 for the case of two speakers. The proposed end-to-end multi-channel multi-speaker ASR model can be divided into three stages. The first stage is a single-channel masking network to perform pre-separation by predicting multiple speaker and noise masks for each channel. Then a multi-source neural beamformer is used to spatially separate multiple speaker sources. In the last stage, an end-to-end ASR module with permutation-free training is used to perform the multi-output speech recognition.

We used a similar architecture as in [23], where the masking network and the neural beamformer are integrated into an attention-based encoder-decoder neural network, and the whole model is jointly optimized solely via a speech recognition objective. The input of the model can consist of an arbitrary number of channels C, and its output is the text sequence for each speaker directly. We denote by J the number of speakers in the mixed utterances, and for simplicity of notation, we shall consider the noise component as the 0-th source.

2.1.1. Monaural Masking Network

The monaural masking network, shown at the bottom of Fig. 1, estimates the masks of each channel for every speaker and an extra noise component. Let us denote by $\mathbf{X}_c = (x_{t,f,c})_{t,f} \in \mathbb{C}^{T \times F}$ the complex STFT of the c-th channel of the observed multi-channel multi-speaker speech, where $1 \le t \le T$, $1 \le f \le F$, $1 \le c \le C$ denote time, frequency, and channel indices, respectively. The mask estimation module produces time-frequency masks $\mathbf{M}_c^i = (m_{t,f,c}^i)_{t,f} \in [0,1]^{T \times F}$, with $i \in \{1, \ldots, J\}$ for each of the J speakers, and $i = 0$ for the noise, using the complex STFT of the c-th channel of the observed multi-channel multi-speaker speech as input. The computation is performed independently on each of the input channels:

$$\mathbf{M}_c = \text{MaskNet}(\mathbf{X}_c), \qquad (1)$$

where $\mathbf{M}_c = (\mathbf{M}_c^i)_i \in [0,1]^{T \times F \times J}$ is the set of estimated masks for the c-th channel.

2.1.2. Multi-source Neural Beamformer

The multi-source neural beamformer is a key component in the proposed model, which produces the separated speech of each speaker. The masks obtained on each channel for each speaker and the noise are used in the computation of the power spectral density (PSD) matrices of each source as follows [18, 28]:

$$\mathbf{\Phi}^i(f) = \frac{1}{\sum_{t=1}^{T} \mathbf{m}_{t,f}^i} \sum_{t=1}^{T} \mathbf{m}_{t,f}^i \mathbf{x}_{t,f} \mathbf{x}_{t,f}^H \in \mathbb{C}^{C \times C}, \qquad (2)$$

where $i \in \{0, \ldots, J\}$, $\mathbf{x}_{t,f} = \{x_{t,f,c}\}_{c=1}^{C}$, $\mathbf{m}_{t,f}^i = \{m_{t,f,c}^i\}_{c=1}^{C}$, and ^{H} represents the conjugate transpose.

After getting the PSD matrices of every speaker and the noise, we estimate the beamformer's time-invariant filter coefficients $\mathbf{g}^i(f)$ at frequency f for each speaker $i \in \{1, \cdots, J\}$ via the MVDR formalization [29] as follows:

$$\mathbf{g}^i(f) = \frac{(\sum_{j \ne i} \mathbf{\Phi}^j(f))^{-1} \mathbf{\Phi}^i(f)}{\text{Tr}((\sum_{j \ne i} \mathbf{\Phi}^j(f))^{-1} \mathbf{\Phi}^i(f))} \mathbf{u} \in \mathbb{C}^C, \qquad (3)$$

where $\mathbf{u} \in \mathbb{R}^C$ is a vector representing the reference microphone that is derived from an attention mechanism [23], and $\text{Tr}(\cdot)$ denotes the trace operation. Notice that in Eq. 3, the formula to derive the filter coefficient is different from that in [23] in the way that the noise PSD is replaced by $\sum_{j \ne i} \mathbf{\Phi}^j(f)$. This is because both noise and other speakers are considered as interference when focusing on a given speaker. This is akin to the speech-speech-noise (SSN) model in [16]. Such a method is employed to make more accurate estimations of the PSD matrices, in which the traditional PSD matrix is expressed using the PSD matrix of interfering speaker and that of the background noise.

Finally, the beamforming filters $\mathbf{g}^i(f)$ obtained in Eq. 3 are used to separate and denoise the input overlapped multi-channel signals

$\mathbf{x}_{t,f} \in \mathbb{C}^C$ to obtain a single-channel estimate of the enhanced STFT $\hat{s}_{t,f}^i$ for speaker i:

$$\hat{s}_{t,f}^i = (\mathbf{g}^i(f))^H \mathbf{x}_{t,f} \in \mathbb{C}. \tag{4}$$

Each separated speech signal waveform can be obtained by inverse STFT for listening, as $\mathrm{iSTFT}(\hat{\mathbf{S}}^i)$, $i = 1, \ldots, J$.

2.1.3. End-to-End Speech Recognition

The outputs of the neural beamformer are estimates of the separated speech signals for each speaker. Before feeding these streams to the end-to-end speech recognition submodule, we need to convert the STFT features to normalized log filterbank features. A log mel filterbank transformation is first applied on the magnitude of the beamformed STFT signal $\hat{\mathbf{S}}^i = (\hat{S}_{t,f}^i)_{t,f}$ for each speaker i, and a global mean-variance normalization is then performed on the log-filterbank feature to produce a proper input $\mathbf{O}^i$ for the speech recognition submodule:

$$\mathbf{FBank}^i = \mathrm{MelFilterBank}(|\hat{\mathbf{S}}^i|), \tag{5}$$

$$\mathbf{O}^i = \mathrm{GlobalMVN}(\log(\mathbf{FBank}^i)). \tag{6}$$

We briefly introduce the end-to-end speech recognition submodule used here, which is similar to the joint CTC/attention-based encoder-decoder architecture [12]. The feature vectors $\mathbf{O}^i$ are first transformed to a hidden representation $\mathbf{H}^i$ by an encoder network. A decoder then generates the output token sequences based on the history information $\mathbf{y}$ and a weighted sum vector $\mathbf{c}$ obtained with an attention mechanism. The end-to-end speech recognition is computed as follows:

$$\mathbf{H}^i = \mathrm{Encoder}(\mathbf{O}^i) \tag{7}$$

$$\mathbf{c}_n^i, \alpha_n^i = \mathrm{Attention}(\alpha_{n-1}^i, \mathbf{e}_{n-1}^i, \mathbf{H}^i) \tag{8}$$

$$\mathbf{e}_n^i = \mathrm{Update}(\mathbf{e}_{n-1}^i, \mathbf{c}_{n-1}^i, \mathbf{y}_{n-1}^i) \tag{9}$$

$$\mathbf{y}_n^i \sim \mathrm{Decoder}(\mathbf{c}_n^i, \mathbf{y}_{n-1}^i), \tag{10}$$

where i denotes the index of the source stream and n an output label sequence index.

Typically, the history information $\mathbf{y}$ is replaced by the reference labels $\mathbf{R} = (r_1, \cdots, r_N)$ in a teacher-forcing fashion at training time. However, since there are multiple possible assignments between the inputs and the references, it is necessary to used permutation invariant training (PIT) in the end-to-end speech recognition [14, 15]. The best permutation of the input sequences and the references is determined by the connectionist temporal classification (CTC) loss $\mathrm{Loss}_{\mathrm{ctc}}$:

$$\hat{\pi} = \operatorname*{argmin}_{\pi \in \mathcal{P}} \sum_i \mathrm{Loss}_{\mathrm{ctc}}(\mathbf{Z}^i, \mathbf{R}^{\pi(i)}), \; i = 1, \ldots, J, \tag{11}$$

where $\mathbf{Z}^i$ denotes the output sequence computed from the encoder output $\mathbf{H}^i$ for the CTC loss, $\mathcal{P}$ is the set of all permutations on $\{1, \ldots, J\}$, and $\pi(i)$ is the i-th element for permutation π.

The final ASR loss of the model is obtained as:

$$\mathcal{L} = \lambda \mathcal{L}_{\mathrm{ctc}} + (1 - \lambda)\mathcal{L}_{\mathrm{att}}, \tag{12}$$

$$\mathcal{L}_{\mathrm{ctc}} = \sum_i \mathrm{Loss}_{\mathrm{ctc}}(\mathbf{Z}^i, \mathbf{R}^{\hat{\pi}(i)}), \tag{13}$$

$$\mathcal{L}_{\mathrm{att}} = \sum_i \mathrm{Loss}_{\mathrm{att}}(\mathbf{Y}^i, \mathbf{R}^{\hat{\pi}(i)}), \tag{14}$$

where $0 \leq \lambda \leq 1$ is an interpolation factor, and $\mathrm{Loss}_{\mathrm{att}}$ is the cross-entropy loss to train the attention-based encoder-decoder networks.

2.2. Data Scheduling and Curriculum Learning

From preliminary empirical results, we find that it is relatively difficult to perform straightforward end-to-end training of such a multi-stage model, especially without an intermediate criterion to guide the training. In our model, the speech recognition submodule has the same architecture as the typical end-to-end speech recognition model, and the input is expected to be similar to the log filterbank of single-speaker speech. Thus, in order to train the model properly, we did not only use the spatialized utterances of the multi-speaker corpus but also the single-speaker utterances from the original WSJ training set. During training, every batch is randomly chosen either from the multi-channel multi-speaker set or from the single-channel single-speaker set. For single-speaker batches, the masking network and neural beamformer stages are bypassed, and the input is directly fed to the recognition submodule. Furthermore, the loss is calculated without considering permutations, as there is only a single speaker per input.

With this data scheduling scheme, the model can achieve a decent performance from random initialization. For multi-channel multi-speaker data batches, the loss of the ASR objective function is back-propagated down through the model to the masking network. For data batches consisting of single-speaker utterances, only the speech recognition part is optimized, which leads to more accurate loss computation in the future. The single-speaker data batches rectify the behavior of the ASR model as it performs regularization during the training.

According to previous researches, starting from easier subtasks can lead the model to learn better, an approach called curriculum learning [30, 31]. To further exploit the data scheduling scheme, we introduce more constraints on the order of the data batches of the training set. As was observed in prior research by [9], the signal-to-noise ratio (SNR, the energy ratio between the target speech and the interfering sources) has a great influence on the final recognition performance. When the speech energy levels of the target speaker and the interfering sources are obviously different, the recognition accuracy of the interfering source speech is very poor. Thus, we sort the multi-speaker data in ascending order of SNR between the loudest and quietest speaker, thus starting with mixtures where both speakers are at similar levels. Furthermore, we sort the single-speaker data from short to long, as short sequences tend to be easier to learn in seq2seq learning. The strategy is formally depicted in Algorithm 1. We applied such a curriculum learning strategy in order to make the model learn step by step and expect it to improve the training.

3. EXPERIMENT

To check the effectiveness of our proposed end-to-end model, we evaluated it on the remixed WSJ1 data used in [14], which we here refer to as the wsj1-2mix dataset. The multi-speaker speech training set was generated by randomly selecting two utterances from the WSJ SI284 corpus, resulting in a 98.5 h dataset. The signal-to-noise ratio (SNR) of one source against the other was randomly chosen from a uniform distribution in the range of $[-5, 5]$ dB. The validation and evaluation sets were generated in a similar way by selecting source utterances from the WSJ Dev93 and Eval92 respectively, and the durations are 1.3 h and 0.8 h. We then create a new spatialized version of the wsj1-2mix dataset following the process applied to the wsj0-2mix dataset in [17], using a room impulse response (RIR) generator[1], where the characteristics of each two-speaker mixture

---

[1]Available online at `https://github.com/ehabets/RIR-Generator`

Algorithm 1: Curriculum learning strategy

---

1 Load the training dataset $\mathbf{X}$;
2 Categorize the training data $\mathbf{X}$ into single-channel single-speaker data $\mathbf{X}_{\text{clean}}$ and multi-channel multi-speaker data $\mathbf{X}_{\text{noisy}}$;
3 Sort the single-channel single-speaker training data in $\mathbf{X}_{\text{clean}}$ in ascending order of the utterance lengths, leading to $\mathbf{X}'_{\text{clean}}$;
4 Sort the multi-channel multi-speaker training data in $\mathbf{X}_{\text{noisy}}$ in ascending order of the SNR level, leading to $\mathbf{X}'_{\text{noisy}}$;
5 Divide $\mathbf{X}'_{\text{clean}}$ and $\mathbf{X}'_{\text{noisy}}$ into minibatch sets $\mathcal{B}_{\text{clean}}$ and $\mathcal{B}_{\text{noisy}}$;
6 Sort batches to alternate between batches from $\mathcal{B}_{\text{clean}}$ and $\mathcal{B}_{\text{noisy}}$;
7 **while** *model is not converged* **do**
8   **for** *each b in all minibatches* **do**
9   Feed minibatch b into the model, update the model;
10   **end**
11 **end**
12 **while** *model is not converged* **do**
13   Shuffle the training data in $\mathbf{X}_{\text{clean}}$ and $\mathbf{X}_{\text{noisy}}$ randomly and divide them into minibatch sets $\mathcal{B}'_{\text{clean}}$ and $\mathcal{B}'_{\text{noisy}}$;
14   Select each minibatch randomly from $\mathcal{B}'_{\text{clean}}$ and $\mathcal{B}'_{\text{noisy}}$ and feed it into the model iteratively to update the model;
15 **end**

---

are randomly generated including room dimensions, speaker locations, and microphone geometry[2].

To train the model, we used the spatialized wsj1-2mix data with $J = 2$ speakers as well as the train_si284 training set from the WSJ1 dataset to regularize the training procedure. All input data are raw waveform audio signals. The STFT was computed using 25 ms-width Hann window with 10 ms shift, with zero-padding resulting in a spectral dimension $F = 257$. In our experiments, we only report results in the case of $C = 2$ channels, but our model is flexible and can be used with an arbitrary number of channels. We first report recognition and separation results in the anechoic scenario in Sections 3.2 and 3.3. Then we show preliminary results in the reverberant scenario in Section 3.4.

3.1. Configurations

Our end-to-end multi-channel multi-speaker model is completely built based on the ESPnet framework [32] with Pytorch backend. All the network parameters were initialized randomly from uniform distribution in the range $[-0.1, 0.1]$. We used AdaDelta with $\rho = 0.95$ and $\epsilon = 1e^{-8}$ as optimization method. The maximum number of epochs for training is set to 15 but the training process is stopped early if performance does not increase for 3 consecutive epochs. For decoding, a word-based language model [33] was trained on the transcripts of the WSJ corpus.

3.1.1. Neural Beamformer

The mask estimation network is a 3-layer bidirectional long-short term memory with projection (BLSTMP) network with 512 cells in

each direction. The computation of the reference microphone vector has the same parameters as in [23] except the vector dimension which is here set to 512. In the MVDR formula of Eq. 3, a small value ϵ is added to the PSD matrix to guarantee that an inverse exists.

3.1.2. Encoder-Decoder Network

The encoder network consists of two VGG-motivated CNN blocks and three BLSTMP layers. The CNN layers have a kernel size of 3×3 and the number of feature maps is 64 and 128 in the first and second block, respectively. Every BLSTMP layer has 1024 memory cells in each direction with projection size 1024. 80 dimensional log filterbank features are extracted for each separated speech signals and global mean-variance normalization is applied, using the statistics of the single-speaker WSJ1 training set. In the decoder network, there is only a single layer of unidirectional long-short term memory network (LSTM) and the number of cells is 300. The interpolation factor λ of the loss function in Eq. 12 is set to 0.2.

3.2. Performance of Multi-Speaker Speech Recognition

In this subsection, we describe the speech recognition performance on the spatialized anechoic wsj1-2mix data, which only modifies the signals via delays and decays due to the propagation. Note that although beamforming algorithms can address the anechoic case without too much effort, it is still necessary to show that our proposed end-to-end method can address the multi-channel multi-speaker speech recognition problem and both the speech recognition submodule and the neural beamforming separation submodule perform well as they are designed. We shall also note that the whole system is trained solely through an ASR objective, and it is thus not trivial for the system to learn how to properly separate the signals even in the anechoic case.

The multi-speaker speech recognition performance is shown in Table 1. There are three single-channel end-to-end speech recognition baseline systems. The first one is a single-channel multi-speaker ASR model trained on the first channel of the spatialized corpus, where the model is the same as the one proposed in [15]. The second is a single-channel multi-speaker ASR model trained with speech that is enhanced by BeamformIt [34], which is a well-known delay-and-sum beamformer. And the third one is to use BeamformIt to first separate the speech by choosing its best and second-best output streams, and then to recognize them with a normal single-speaker end-to-end ASR model. The spatialization of the corpus results in a degradation of the performance: the multi-speaker ASR model trained with the 1st channel has a word error rate (WER) of 29.43% on the evaluation set, compared to only 20.43% obtained on the original unspatialized wsj1-2mix data in [15]. Using the BeamformIt tool to enhance the spatialized signal can improve the recognition accuracy of a multi-speaker model, leading to a WER of 21.75% on the evaluation set. However, traditional beamforming algorithms such as BeamformIt can not perfectly separate the overlapped speech signals, and the performance of the single-speaker model in terms of WER is very poor, 98.00%.

The performance of our proposed end-to-end multi-channel multi-speaker model (MIMO-Speech) is shown at the bottom of the table. The curriculum learning strategy described in Section 2.2 is used to further improve performance. From the table, it can be observed that MIMO-Speech is significantly better than traditional methods, achieving 4.51% character error rate (CER) and 8.62% word error rate (WER). Compared against the best baseline model,

the relative improvement is over 60% in terms of both CER and WER. When applying our data scheduling scheme by sorting the multi-speaker speech in ascending order of SNRs, an additional performance boost can be realized. The final CER and WER on the evaluation set are 3.75% and 7.55% respectively, with over 12% relative improvement against MIMO-Speech without curriculum learning. Overall, our proposed MIMO-Speech network can achieve good recognition performance on the spatialized anechoic wsj1-2mix corpus.

Table 1. Performance terms of average CER and WER [%] on the spatialized anechoic wsj1-2mix corpus.

Model	dev CER	eval CER
2-spkr ASR (1st channel)	22.65	19.07
BeamformIt Enhancement (2-spkr ASR)	15.23	12.45
BeamformIt Separation (1-spkr ASR)	77.30	77.10
MIMO-Speech	7.29	4.51
+ Curriculum Learning (SNRs)	**6.34**	**3.75**

Model	dev WER	eval WER
2-spkr ASR (1st channel)	34.98	29.43
BeamformIt Enhancement (2-spkr ASR)	26.61	21.75
BeamformIt Separation (1-spkr ASR)	98.60	98.00
MIMO-Speech	13.54	8.62
+ Curriculum Learning (SNRs)	**12.59**	**7.55**

3.3. Performance of Multi-Speaker Speech Separation

One question regarding our model is whether the front-end of MIMO-Speech, the neural beamformer, learns a proper beamforming behavior as other algorithms do since there is no explicit speech separation criterion to optimize the network. To investigate the role of the neural beamformer, we consider the masks $\mathbf{m}^i$ that are used to compute the PSD matrices and the enhanced separated STFT signals $\hat{\mathbf{s}}^i, i = 1, \ldots, J$ obtained at the output of the beamformer. Example results are shown in Fig. 2. Note that in our model, the masks are not constrained to sum to 1 at each time-frequency unit, resulting in a scaling indeterminacy within each frequency. For better readability in the figures, we here renormalize each mask using its median within each frequency. In the figure, the difference between the masks from each speaker is clear. And from the spectrogram, it is also observed that for each separated stream, the signals are less overlapped compared with the input multi-speaker speech signal. The mask and spectrogram examples suggest that MIMO-Speech can separate the speech to some level.

To evaluate the separation quality, we reconstruct the separated waveforms for each speaker from the outputs of the beamformer via inverse STFT, and compare them with the reference signals in terms of PESQ and scale-invariant signal-to-distortion ratio (SI-SDR) [35]. The results are shown in Table 2. As we can see, the separated audios have very good quality. The separated signals from the MIMO-Speech model [3] have an average PESQ value of 3.6 and an average SI-SDR of 23.1 dB. When using curriculum learning, PESQ and SI-SDR degrade slightly, but the quality is still very high. This result suggests that our proposed MIMO-Speech model is capable of learning to separate overlapped speech via beamforming, based solely on an ASR objective.

---

[3]Audio samples are available online at `https://simpleoier.github.io/MIMO-Speech/index.html`

(a) Mask for Speaker 1   (b) Mask for Speaker 2

(c) Separated Speech for Speaker 1   (d) Separated Speech for Speaker 2
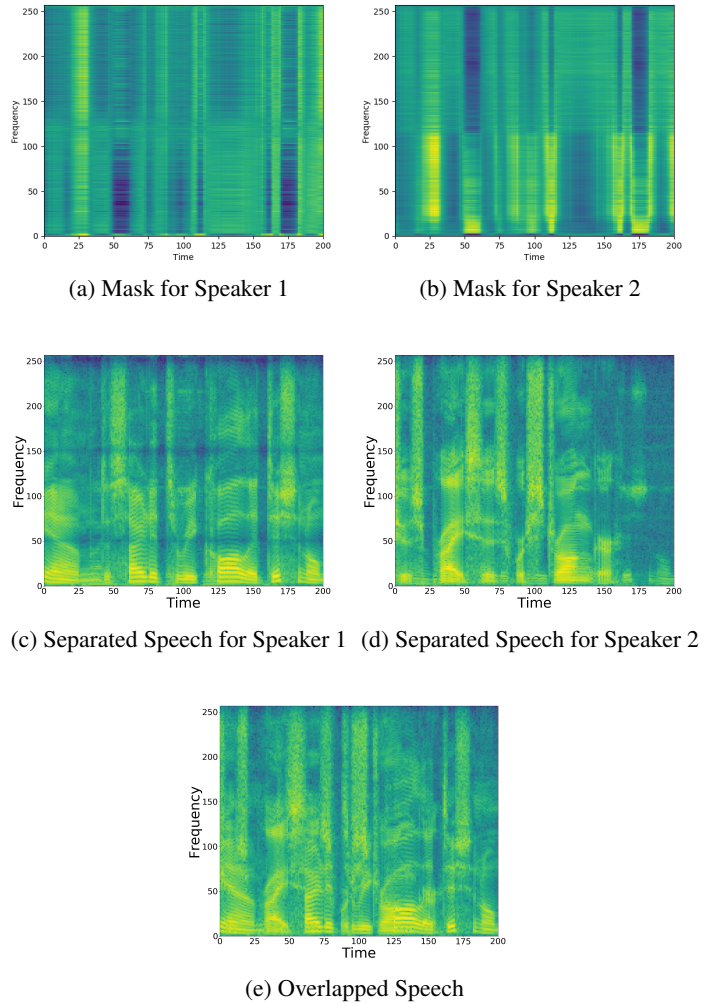
(e) Overlapped Speech

Fig. 2. Example of masks output by the masking network and separated speech log spectrograms output by the MVDR beamformer.

In order to further explore the neural beamformer's effect, we show an example of estimated beam pattern [36] for the separated sources. Figure 3 shows the beam pattern of two separated signals at frequencies $\{500 \text{ Hz}, 1000 \text{ Hz}, 2000 \text{ Hz}, 4000 \text{ Hz}\}$. The value of the beam at different degrees quantifies the reduction of the speech signals received. As we can see from the figures, the crests and troughs of the beams are different for the two speakers, which shows the neural beamformer is trained properly and can tell the difference between the sources.

### 3.4. Evaluation on the spatialized reverberant data

To give a comprehensive analysis of the MIMO-Speech model, we investigated how the model performs in a more realistic case, using the spatialized reverberant wsj1-2mix data. As a comparison, we first trained a normal single-speaker end-to-end speech recognition system. The model is trained with the spatialized reverberant speech from each single speaker. The performance is shown in Table 3. For the MIMO-Speech model, the spatialized reverberant wsj1-2mix training dataset was added to the training set for

Table 2. Performance in terms of average PESQ and SI-SDR [dB] on the spatialized anechoic wsj1-2mix corpus.

Model	dev PESQ	eval PESQ
MIMO-Speech	3.6	3.6
+ Curriculum Learning (SNRs)	**3.7**	**3.6**
Model	dev SI-SDR	eval SI-SDR
MIMO-Speech	**22.1**	**23.1**
+ Curriculum Learning (SNRs)	21.1	21.8
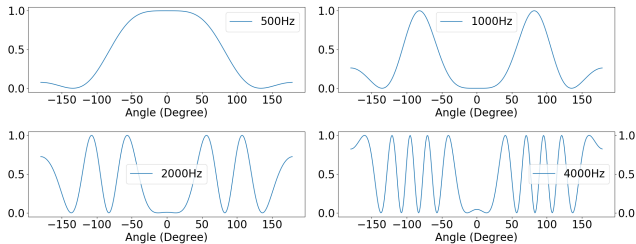
(a)Speaker 1
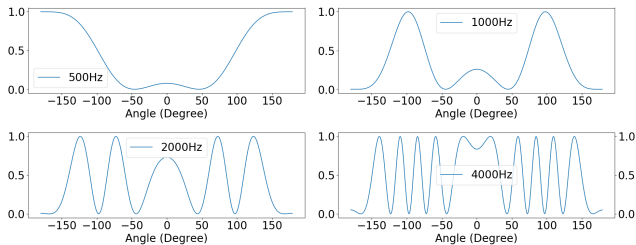
(b) Speaker 2

Fig. 3. Example of beam patterns of the separated speech.

the multi-conditioned training. The results on the speech recognition task are shown in Table 4. The reverberant speech is difficult to recognize as the performance shows severe degradation when we tried to infer the reverberant speech using the anechoic multi-speaker model. The multi-conditioned training can release such degradation, improving the WER by over 60%. The results suggest that the proposed MIMO-Speech also has potential for application in complex scenarios. As a complementary experiment, we used Nara-WPE [37] to perform speech dereverberation only for the development and evaluation data. The speech recognition results are shown in Table.5 which suggests that the speech dereverberation techniques only in the inference stage can lead to further improvement. Note that the results here are just a preliminary study. The main drawback here is that we did not consider any dereverberation techniques in designing our model.

Table 3. Performance in terms of average CER and WER [%] of the baseline single-speaker end-to-end speech recognition model trained on reverberant (R) single-speaker speech and evaluated on reverberant (R) multi-speaker speech.

Model	dev CER (R)	eval CER (R)
End-to-End Model (R)	81.6	82.7
Model	dev WER (R)	eval WER (R)
End-to-End Model (R)	103.9	104.2

Table 4. Performance in terms of average CER and WER [%] on the spatialized wsj1-2mix corpus of MIMO-Speech trained on either anechoic (A) or reverberant (R) and evaluated on either the anechoic (A) or reverberant (R) evaluation set.

Model	eval CER (A)	eval CER (R)
MIMO-Speech (A)	4.51	62.32
MIMO-Speech (R)	4.08	18.15
Model	eval WER (A)	eval WER (R)
MIMO-Speech (A)	8.62	81.30
MIMO-Speech (R)	8.72	29.99

Table 5. Performance in terms of average CER and WER [%] on the spatialized wsj1-2mix corpus of MIMO-Speech trained on either anechoic (A) or reverberant (R) and evaluated on the reverberant data after Nara-WPE dereverberation (D).

Model	dev CER (D)	eval CER (D)
MIMO-Speech (A)	51.00	52.02
MIMO-Speech (R)	20.09	15.04
Model	dev WER (D)	dev WER (D)
MIMO-Speech (A)	69.08	69.42
MIMO-Speech (R)	33.09	25.28

4. CONCLUSION

In this paper, we have proposed an end-to-end multi-channel multi-speaker speech recognition model called MIMO-Speech. More specifically, the model takes multi-speaker speech recorded by a microphone array as input and outputs text sequences for each speaker. Furthermore, the front-end of the model, involving a neural beamformer, learns to perform speech separation even though no explicit signal reconstruction criterion is used. The main advantage of the proposed approach is that the whole model is differentiable and can be optimized with an ASR loss as target. In order to make the training easier, we utilized single-channel single-speaker speech as well. We also designed an effective curriculum learning strategy to improve the performance. Experiments on a spatialized version of the wsj1-2mix corpus show that the proposed framework has fairly good performance. However, performance on reverberant data still suffers from a large gap against the anechoic case. In future work, we will continue to improve this system by incorporating dereverberation strategies as well as by integrating further the masking and beamforming approaches.

5. ACKNOWLEDGEMENT

6. REFERENCES

[1] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *The Journal of the Acoustical Society of America*, vol. 25, no. 5, 1953.

[2] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2016.

[3] Y. Isik, J. Le Roux, Z. Chen, S. Watanabe, and J. R. Hershey, "Single-channel multi-speaker separation using deep clustering," in *Proc. ISCA Interspeech*, Sep. 2016.

[4] T. Menne, I. Sklyar, R. Schlüter, and H. Ney, "Analysis of deep clustering as preprocessing for automatic speech recognition of sparsely overlapping speech," *arXiv preprint arXiv:1905.03500*, 2019.

[5] S. Settle, J. Le Roux, T. Hori, S. Watanabe, and J. R. Hershey, "End-to-end multi-speaker speech recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2018.

[6] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2017.

[7] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 25, no. 10, 2017.

[8] D. Yu, X. Chang, and Y. Qian, "Recognizing multi-talker speech with permutation invariant training," in *Proc. ISCA Interspeech*, Aug. 2017.

[9] Y. Qian, X. Chang, and D. Yu, "Single-channel multi-talker speech recognition with permutation invariant training," *Speech Communication*, vol. 104, 2018.

[10] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *Proc. International Conference on Machine Learning (ICML)*, Jun. 2014.

[11] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2016.

[12] S. Kim, T. Hori, and S. Watanabe, "Joint CTC-attention based end-to-end speech recognition using multi-task learning," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2017.

[13] T. Hori, S. Watanabe, and J. Hershey, "Joint CTC/attention decoding for end-to-end speech recognition," in *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, vol. 1, Jul. 2017.

[14] H. Seki, T. Hori, S. Watanabe, J. Le Roux, and J. R. Hershey, "A purely end-to-end system for multi-speaker speech recognition," in *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, Jul. 2018.

[15] X. Chang, Y. Qian, K. Yu, and S. Watanabe, "End-to-end monaural multi-speaker ASR system without pretraining," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019.

[16] T. Yoshioka, H. Erdogan, Z. Chen, X. Xiao, and F. Alleva, "Recognizing overlapped speech in meetings: A multichannel separation approach using neural networks," in *Proc. ISCA Interspeech*, Sep. 2018.

[17] Z.-Q. Wang, J. Le Roux, and J. R. Hershey, "Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2018.

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

[19] H. Erdogan, J. R. Hershey, S. Watanabe, M. I. Mandel, and J. Le Roux, "Improved MVDR beamforming using single-channel mask prediction networks," in *Proc. ISCA Interspeech*, Sep. 2016.

[20] T. Menne, J. Heymann, A. Alexandridis, K. Irie, A. Zeyer, M. Kitza, P. Golik, I. Kulikov, L. Drude, R. Schlüter *et al.*, "The RWTH/UPB/FORTH system combination for the 4th CHiME challenge evaluation," in *Proc. CHiME workshop*, Sep. 2016.

[21] J. Heymann, L. Drude, C. Boeddeker, P. Hanebrink, and R. Haeb-Umbach, "Beamnet: End-to-end training of a beamformer-supported multi-channel ASR system," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2017.

[22] W. Minhua, K. Kumatani, S. Sundaram, N. Ström, and B. Hoffmeister, "Frequency domain multi-channel acoustic modeling for distant speech recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019.

[23] T. Ochiai, S. Watanabe, T. Hori, and J. R. Hershey, "Multi-channel end-to-end speech recognition," in *Proc. International Conference on Machine Learning (ICML)*, 2017.

[24] S. Braun, D. Neil, J. Anumula, E. Ceolini, and S.-C. Liu, "Multi-channel attention for end-to-end speech recognition." in *Proc. ISCA Interspeech*, Sep. 2018.

[25] X. Wang, R. Li, S. H. Mallidi, T. Hori, S. Watanabe, and H. Hermansky, "Stream attention-based multi-array end-to-end speech recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019.

[26] A. Shanmugam Subramanian, X. Wang, S. Watanabe, T. Taniguchi, D. Tran, and Y. Fujita, "An investigation of end-to-end multichannel speech recognition for reverberant and mismatch conditions," *arXiv preprint arXiv:1904.09049*, Apr. 2019.

[27] T. Ochiai, S. Watanabe, and S. Katagiri, "Does speech enhancement work with end-to-end ASR objectives?: Experimental analysis of multichannel end-to-end ASR," in *Proc. International Workshop on Machine Learning for Signal Processing (MLSP)*, Sep. 2017.

[28] T. Yoshioka, N. Ito, M. Delcroix, A. Ogawa, K. Kinoshita, M. Fujimoto, C. Yu, W. J. Fabian, M. Espi, T. Higuchi *et al.*, "The NTT CHiME-3 system: Advances in speech enhancement and recognition for mobile multi-microphone devices," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Dec. 2015.

[29] M. Souden, J. Benesty, and S. Affes, "On optimal frequency-domain multichannel linear filtering for noise reduction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, 2009.

[30] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proc. International Conference on Machine Learning (ICML)*, Jun. 2009.

[31] D. Amodei *et al.*, "Deep Speech 2: End-to-end speech recognition in English and Mandarin," in *Proc. International Conference on Machine Learning (ICML)*, Jun. 2016.

[32] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. E. Y. Soplin, J. Heymann, M. Wiesner, N. Chen *et al.*, "ESPnet: End-to-end speech processing toolkit," *arXiv preprint arXiv:1804.00015*, 2018.

[33] T. Hori, J. Cho, and S. Watanabe, "End-to-end speech recognition with word-based RNN language models," in *Proc. IEEE Spoken Language Technology Workshop (SLT)*, Dec. 2018.

[34] 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, Sep. 2007.

[35] J. Le Roux, S. T. Wisdom, H. Erdogan, and J. R. Hershey, "SDR – half-baked or well done?" in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019.

[36] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 25, no. 4, 2017.

[37] L. Drude, J. Heymann, C. Boeddeker, and R. Haeb-Umbach, "NARA-WPE: A Python package for weighted prediction error dereverberation in Numpy and Tensorflow for online and offline processing," in *ITG Fachtagung Sprachkommunikation (ITG)*, Oct. 2018.