Multichannel End-to-end Speech Recognition

Tsubasa Ochiai [1] **Shinji Watanabe** [2] **Takaaki Hori** [2] **John R. Hershey** [2]

Abstract

The field of speech recognition is in the midst of a paradigm shift: end-to-end neural networks are challenging the dominance of hidden Markov models as a core technology. Using an attention mechanism in a recurrent encoder-decoder architecture solves the dynamic time alignment problem, allowing joint end-to-end training of the acoustic and language modeling components. In this paper we extend the end-to-end framework to encompass microphone array signal processing for noise suppression and speech enhancement within the acoustic encoding network. This allows the beamforming components to be optimized jointly within the recognition architecture to improve the end-to-end speech recognition objective. Experiments on the noisy speech benchmarks (CHiME-4 and AMI) show that our multichannel end-to-end system outperformed the attention-based baseline with input from a conventional adaptive beamformer.

1. Introduction

Existing automatic speech recognition (ASR) systems are based on a complicated hybrid of separate components, including acoustic, phonetic, and language models (Jelinek, 1976). Such systems are typically based on deep neural network acoustic models combined with hidden Markov models to represent the language and phonetic context-dependent state and their temporal alignment with the acoustic signal (DNN-HMM) (Bourlard & Morgan, 1994; Hinton et al., 2012). As a simpler alternative, end-to-end speech recognition paradigm has attracted great research interest (Chorowski et al., 2014; 2015; Chan et al., 2016; Graves & Jaitly, 2014; Miao et al., 2015). This paradigm simplifies the above hybrid architecture by subsuming it into a single neural network. Specifically, an attention-based encoder-decoder framework (Chorowski et al., 2014) integrates all of those components using a set of recurrent neural networks (RNN), which map from acoustic feature sequences to character label sequences.

However, existing end-to-end frameworks have focused on clean speech, and do not include speech enhancement, which is essential to good performance in noisy environments. For example, recent industrial applications (e.g., Amazon echo) and benchmark studies (Barker et al., 2016; Kinoshita et al., 2016) show that multichannel speech enhancement techniques, using beamforming methods, produce substantial improvements as a pre-processor for conventional hybrid systems, in the presence of strong background noise. In light of the above trends, this paper extends the existing attention-based encoder-decoder framework by integrating multichannel speech enhancement. Our proposed *multichannel end-to-end speech recognition* framework is trained to directly translate from multichannel acoustic signals to text.

A key concept of the multichannel end-to-end framework is to optimize the entire inference procedure, including the beamforming, based on the final ASR objectives, such as word/character error rate (WER/CER). Traditionally, beamforming techniques such as delay-and-sum and filter-and-sum are optimized based on a signal-level loss function, independently of speech recognition task (Benesty et al., 2008; Van Veen & Buckley, 1988). Their use in ASR requires ad-hoc modifications such as Wiener post-filtering or distortionless constraints, as well as steering mechanisms determine a look direction to focus the beamformer on the target speech (Wölfel & McDonough, 2009). In contrast, our framework incorporates recently proposed neural beamforming mechanisms as a differentiable component to allow joint optimization of the multichannel speech enhancement within the end-to-end system to improve the ASR objective.

Recent studies on neural beamformers can be categorized into two types: (1) beamformers with a filter estimation network (Li et al., 2016; Xiao et al., 2016a; Meng et al., 2017) and (2) beamformers with a mask estimation network (Heymann et al., 2016; Erdogan et al., 2016). Both methods obtain an enhanced signal based on the formal-

[1]Doshisha University, Kyoto, Japan [2]Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA, USA. Correspondence to: Tsubasa Ochiai <eup1105@mail4.doshisha.ac.jp>, Shinji Watanabe <watanabe@merl.com>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017.

ization of the conventional filter-and-sum beamformer in the time-frequency domain. The main difference between them is how the multichannel filters are produced by the neural network. In the former approach, the multichannel filter coefficients are direct outputs of the network. In the latter approach, a network first estimates time-frequency masks, which are used to compute expected speech and noise statistics. Then, using these statistics, the filter coefficients are computed based on the well-known MVDR (minimum variance distortionless response) formalization (Capon, 1969). In both approaches, the estimated filter coefficients are then applied to the multichannel noisy signal to enhance the speech signal. Note that the mask estimation approach has the advantage of leveraging well-known techniques, but it requires parallel data composed of aligned clean and noisy speech, which are usually difficult to obtain without data simulation.

Recently, it has been reported that the mask estimation-based approaches (Yoshioka et al., 2015; Heymann et al., 2016; Erdogan et al., 2016) achieve great performance in noisy speech recognition benchmarks (e.g., CHiME 3 and 4 challenges). Although this paper proposes to incorporate both mask and filter estimation approaches in an end-to-end framework, motivated by those successes, we focus more on the mask estimation, implementing it along with the MVDR estimation as a differentiable network. Our MVDR formulation estimates the speech image at the *reference* microphone and includes selection of the reference microphone using an attention mechanism. By using channel-independent mask estimation along with this reference selection, the model can generalize to different microphone array geometries (number of channels, microphone locations, and ordering), unlike the filter estimation approach. Finally, because the masks are latent variables in the end-to-end training, we no longer need parallel clean and noisy speech.

The main advantages of our proposed multichannel end-to-end speech recognition system are:

1. Overall inference from speech enhancement to recognition is jointly optimized for the ASR objective.
2. The trained system can be used for input signals with arbitrary number and order of channels.
3. Parallel clean and noisy data are not required. We can optimize the speech enhancement component with noisy signals and their transcripts.

2. Overview of attention-based encoder-decoder networks

This section explains a conventional attention-based encoder-decoder framework, which is used to directly deal

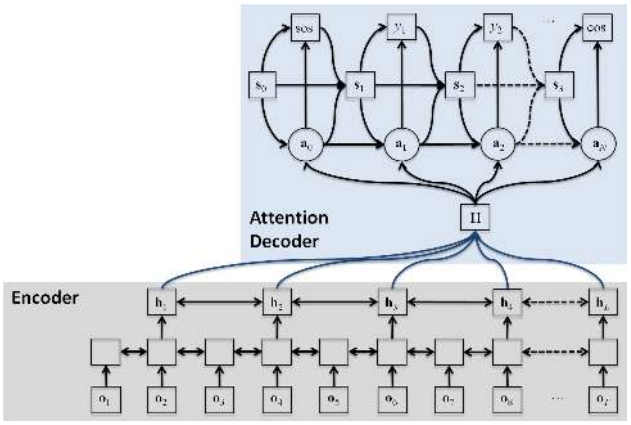


Figure 1. The structure of an attention-based encoder-decoder framework. The encoder transforms an input sequence O into a high-level feature sequence H, and then the decoder generates a character sequence Y through the attention mechanism.

with variable length input and output sequences. The framework consists of two RNNs, called encoder and decoder respectively, and an attention mechanism, which connects the encoder and decoder, as shown in Figure 1. Given a T-length sequence of input features $O = \{\mathbf{o}_t \in \mathbb{R}^{D_\mathrm{O}} | t = 1, \cdots, T\}$, the network generates an N-length sequence of output labels $Y = \{y_n \in \mathcal{V} | n = 1, \cdots, N\}$, where $\mathbf{o}_t$ is a D_O-dimensional feature vector (e.g., log Mel filterbank) at input time step t, and y_n is a label symbol (e.g., character) at output time step n in label set $\mathcal{V}$.

First, given an input sequence O, the encoder network transforms it to an L-length high-level feature sequence $H = \{\mathbf{h}_l \in \mathbb{R}^{D_\mathrm{H}} | l = 1, \cdots, L\}$, where $\mathbf{h}_l$ is a D_H-dimensional state vector at a time step l of encoder's top layer. In this work, the encoder network is composed of a bidirectional long short-term memory (BLSTM) recurrent network. To reduce the input sequence length, we apply a subsampling technique (Bahdanau et al., 2016) to some layers. Therefore, l represents the frame index subsampled from t and L is less than T.

Next, the attention mechanism integrates all encoder outputs H into a D_H-dimensional context vector $\mathbf{c}_n \in \mathbb{R}^{D_\mathrm{H}}$ based on an L-dimensional attention weight vector $\mathbf{a}_n \in [0, 1]^L$, which represents a soft alignment of encoder outputs at an output time step n. In this work, we adopt a location-based attention mechanism (Chorowski et al., 2015), and $\mathbf{a}_n$ and $\mathbf{c}_n$ are formalized as follows:

$$\mathbf{f}_n = \mathbf{F} * \mathbf{a}_{n-1}, \tag{1}$$

$$k_{n,l} = \mathbf{w}^\mathrm{T} \tanh(\mathbf{V}^\mathrm{S}\mathbf{s}_n + \mathbf{V}^\mathrm{H}\mathbf{h}_l + \mathbf{V}^\mathrm{F}\mathbf{f}_{n,l} + \mathbf{b}), \tag{2}$$

$$a_{n,l} = \frac{\exp(\alpha k_{n,l})}{\sum_{l=1}^{L} \exp(\alpha k_{n,l})}, \quad \mathbf{c}_n = \sum_{l=1}^{L} a_{n,l}\mathbf{h}_l, \tag{3}$$

where $\mathbf{w} \in \mathbb{R}^{1\times D_{\mathrm{W}}}$, $\mathbf{V}^{\mathrm{H}} \in \mathbb{R}^{D_{\mathrm{W}}\times D_{\mathrm{H}}}$, $\mathbf{V}^{\mathrm{S}} \in \mathbb{R}^{D_{\mathrm{W}}\times D_{\mathrm{S}}}$, $\mathbf{V}^{\mathrm{F}} \in \mathbb{R}^{D_{\mathrm{W}}\times D_{\mathrm{F}}}$ are trainable weight matrices, $\mathbf{b} \in \mathbb{R}^{D_{\mathrm{W}}}$ is a trainable bias vector, $\mathbf{F} \in \mathbb{R}^{D_{\mathrm{F}}\times 1\times D_{\mathrm{f}}}$ is a trainable convolution filter. $\mathbf{s}_n \in \mathbb{R}^{D_{\mathrm{S}}}$ is a D_{S}-dimensional hidden state vector obtained from an upper decoder network at n, and α is a sharpening factor (Chorowski et al., 2015). $*$ denotes the convolution operation.

Then, the decoder network incrementally updates a hidden state $\mathbf{s}_n$ and generates an output label y_n as follows:

$$\mathbf{s}_n = \mathrm{Update}(\mathbf{s}_{n-1}, \mathbf{c}_{n-1}, y_{n-1}), \tag{4}$$
$$y_n = \mathrm{Generate}(\mathbf{s}_n, \mathbf{c}_n), \tag{5}$$

where the $\mathrm{Generate}(\cdot)$ and $\mathrm{Update}(\cdot)$ functions are composed of a feed forward network and an LSTM-based recurrent network, respectively.

Now, we can summarize these procedures as follows:

$$P(Y|O) = \prod_n P(y_n|O, y_{1:n-1}), \tag{6}$$
$$H = \mathrm{Encoder}(O), \tag{7}$$
$$\mathbf{c}_n = \mathrm{Attention}(\mathbf{a}_{n-1}, \mathbf{s}_n, H), \tag{8}$$
$$y_n = \mathrm{Decoder}(\mathbf{c}_n, y_{1:n-1}), \tag{9}$$

where $\mathrm{Encoder}(\cdot) = \mathrm{BLSTM}(\cdot)$, $\mathrm{Attention}(\cdot)$ corresponds to Eqs. (1)-(3), and $\mathrm{Decoder}(\cdot)$ corresponds to Eqs. (4) and (5). Here, special tokens for start-of-sentence (sos) and end-of-sentence (eos) are added to the label set $\mathcal{V}$. The decoder starts the recurrent computation with the (sos) label and continues to generate output labels until the (eos) label is emitted. Figure 1 illustrates such procedures.

Based on the cross-entropy criterion, the loss function is defined using Eq. (6) as follows:

$$\mathcal{L} = -\ln P(Y^*|O) = -\sum_n \ln P(y_n^*|O, y_{1:n-1}^*), \tag{10}$$

where Y^* is the ground truth of a whole sequence of output labels and $y_{1:n-1}^*$ is the ground truth of its subsequence until an output time step $n-1$.

In this framework, the whole networks including the encoder, attention, and decoder can be optimized to generate the correct label sequence. This consistent optimization of all relevant procedures is the main motivation of the end-to-end framework.

3. Neural beamformers

This section explains neural beamformer techniques, which are integrated with the encoder-decoder network in the following section. This paper uses frequency-domain beamformers rather than time-domain ones, which achieve significant computational complexity reduction in multichannel neural processing (Li et al., 2016; Sainath et al., 2016).

In the frequency domain representation, a filter-and-sum beamformer obtains an enhanced signal by applying a linear filter as follows:

$$\hat{x}_{t,f} = \sum_{c=1}^{C} g_{t,f,c} x_{t,f,c}, \tag{11}$$

where $x_{t,f,c} \in \mathbb{C}$ is an STFT coefficient of c-th channel noisy signal at a time-frequency bin (t, f). $g_{t,f,c} \in \mathbb{C}$ is a corresponding beamforming filter coefficient. $\hat{x}_{t,f} \in \mathbb{C}$ is an enhanced STFT coefficient, and C is the numbers of channels.

In this paper, we adopt two types of neural beamformers, which basically follow Eq. (11); 1) filter estimation network and 2) mask estimation network. Figure 2 illustrates the schematic structure of each approach. The main difference between them is how to compute the filter coefficient $g_{t,f,c}$. The following subsections describe each approach.

3.1. Filter estimation network approach

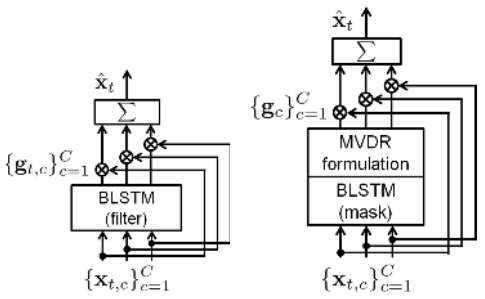
The filter estimation network directly estimates a *time-variant* filter coefficients $\{g_{t,f,c}\}_{t=1,f=1,c=1}^{T,F,C}$ as the outputs of the network, which was originally proposed in (Li et al., 2016). F is the dimension of STFT features.

This approach uses a single real-valued BLSTM network to predict the real and imaginary parts of the complex-valued filter coefficients at an every time step. Therefore, we introduce multiple $(2 \times C)$ output layers to separately compute the real and imaginary parts of the filter coefficients for each channel. Then, the network outputs time-variant filter coefficients $\mathbf{g}_{t,c} = \{g_{t,f,c}\}_{f=1}^{F} \in \mathbb{C}^F$ at a time step t for c-th channel as follows;

$$Z = \mathrm{BLSTM}(\{\bar{\mathbf{x}}_t\}_{t=1}^{T}), \tag{12}$$
$$\Re(\mathbf{g}_{t,c}) = \tanh(\mathbf{W}_c^{\Re}\mathbf{z}_t + \mathbf{b}_c^{\Re}), \tag{13}$$
$$\Im(\mathbf{g}_{t,c}) = \tanh(\mathbf{W}_c^{\Im}\mathbf{z}_t + \mathbf{b}_c^{\Im}), \tag{14}$$

where $Z = \{\mathbf{z}_t \in \mathbb{R}^{D_{\mathrm{Z}}} | t = 1, \cdots, T\}$is a sequence of D_{Z}-dimensional output vectors of the BLSTM network. $\bar{\mathbf{x}}_t = \{\Re(x_{t,f,c}), \Im(x_{t,f,c})\}_{f=1,c=1}^{F,C} \in \mathbb{R}^{2FC}$ is an input feature of a $2FC$-dimensional real-value vector for the BLSTM network. This is obtained by concatenating the real and imaginary parts of all STFT coefficients in all channels. $\Re(\mathbf{g}_{t,c})$ and $\Im(\mathbf{g}_{t,c})$ is the real and imaginary part of filter coefficients, $\mathbf{W}_c^{\Re} \in \mathbb{R}^{F\times D_{\mathrm{Z}}}$ and $\mathbf{W}_c^{\Im} \in \mathbb{R}^{F\times D_{\mathrm{Z}}}$ are the weight matrices of the output layer for c-th channel, and $\mathbf{b}_c^{\Re} \in \mathbb{R}^F$ and $\mathbf{b}_c^{\Im} \in \mathbb{R}^F$ are their corresponding bias vectors. Using the estimated filters $\mathbf{g}_{t,c}$, the enhanced STFT coefficients $\hat{x}_{t,f}$ are obtained based on Eq. (11).

This approach has several possible problems due to its formalization. The first issue is the high flexibility of the estimated filters $\{g_{t,f,c}\}_{t=1,f=1,c=1}^{T,F,C}$, which are composed of



(a) filter estimation (b) mask estimation

Figure 2. Structures of neural beamformers. (a) Filter estimation network, which directly estimates the filter coefficients. (b) Mask estimation network, which estimates time-frequency masks, and then get filter coefficients based on the MVDR formalization.

a large number of unconstrained variables ($2TFC$) estimated from few observations. This causes problems such as training difficulties and over-fitting. The second issue is that the network structure depends on the number and order of channels. Therefore, a new filter estimation network has to be trained when we change microphone configurations.

3.2. Mask estimation network approach

The key point of the mask estimation network approach is that it constrains the estimated filters based on well-founded array signal processing principles. Here, the network estimates the time-frequency masks, which are used to compute the *time-invariant* filter coefficients $\{g_{f,c}\}_{f=1,c=1}^{F,C}$ based on the MVDR formalizations. This is the main difference between this approach and the filter estimation network approach described in Section 3.1. Also, mask-based beamforming approaches have achieved great performance in noisy speech recognition benchmarks (Yoshioka et al., 2015; Heymann et al., 2016; Erdogan et al., 2016). Therefore, this paper proposes to use a mask-based MVDR beamformer, where overall procedures are formalized as a differentiable network for the subsequent end-to-end speech recognition system. Figure 3 summarizes the overall procedures to compute the filter coefficients, which is a detailed flow of Figure 2 (b).

3.2.1. Mask-based MVDR formalization

One of the MVDR formalizations computes the time-invariant filter coefficients $\mathbf{g}(f) = \{g_{f,c}\}_{c=1}^{C} \in \mathbb{C}^C$ in Eq. (11) as follows (Souden et al., 2010):

$$\mathbf{g}(f) = \frac{\mathbf{\Phi}^{\mathrm{N}}(f)^{-1}\mathbf{\Phi}^{\mathrm{S}}(f)}{\mathrm{Tr}(\mathbf{\Phi}^{\mathrm{N}}(f)^{-1}\mathbf{\Phi}^{\mathrm{S}}(f))}\mathbf{u}, \tag{15}$$

where $\mathbf{\Phi}^{\mathrm{S}}(f) \in \mathbb{C}^{C\times C}$ and $\mathbf{\Phi}^{\mathrm{N}}(f) \in \mathbb{C}^{C\times C}$ are the cross-channel power spectral density (PSD) matrices (also known as spatial covariance matrices) for speech and noise signals,

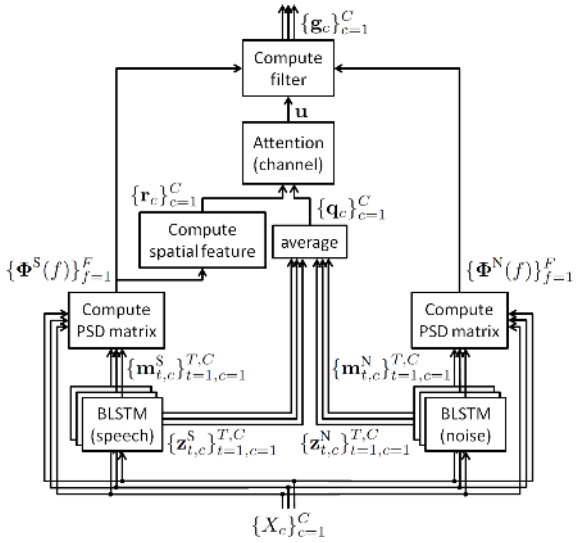


Figure 3. Overall procedures to compute filter coefficients in mask estimation network approach.

respectively. $\mathbf{u} \in \mathbb{R}^C$ is the one-hot vector representing a reference microphone, and $\mathrm{Tr}(\cdot)$ is the matrix trace operation. Note that although the formula contains a matrix inverse, the number of channels is relatively small, and so the forward pass and derivatives can be efficiently computed.

Based on (Yoshioka et al., 2015; Heymann et al., 2016), the PSD matrices are robustly estimated using the expectation with respect to time-frequency masks as follows:

$$\mathbf{\Phi}^{\mathrm{S}}(f) = \frac{1}{\sum_{t=1}^{T} m_{t,f}^{\mathrm{S}}} \sum_{t=1}^{T} m_{t,f}^{\mathrm{S}} \mathbf{x}_{t,f}\mathbf{x}_{t,f}^{\dagger}, \tag{16}$$

$$\mathbf{\Phi}^{\mathrm{N}}(f) = \frac{1}{\sum_{t=1}^{T} m_{t,f}^{\mathrm{N}}} \sum_{t=1}^{T} m_{t,f}^{\mathrm{N}} \mathbf{x}_{t,f}\mathbf{x}_{t,f}^{\dagger}, \tag{17}$$

where $\mathbf{x}_{t,f} = \{x_{t,f,c}\}_{c=1}^{C} \in \mathbb{C}^C$ is the spatial vector of an observed signal for each time-frequency bin, $m_{t,f}^{\mathrm{S}} \in [0,1]$ and $m_{t,f}^{\mathrm{N}} \in [0,1]$ are the time-frequency masks for speech and noise, respectively. $\dagger$ represents the conjugate transpose.

3.2.2. Mask estimation network

In the mask estimation network approach, we use two real-valued BLSTM networks; one for a speech mask and the other for a noise mask. Each network outputs the time-frequency mask as follows:

$$Z_c^{\mathrm{S}} = \mathrm{BLSTM}^{\mathrm{S}}(\{\bar{\mathbf{x}}_{t,c}\}_{t=1}^{T}), \tag{18}$$

$$\mathbf{m}_{t,c}^{\mathrm{S}} = \mathrm{sigmoid}(\mathbf{W}^{\mathrm{S}}\mathbf{z}_{t,c}^{\mathrm{S}} + \mathbf{b}^{\mathrm{S}}), \tag{19}$$

$$Z_c^{\mathrm{N}} = \mathrm{BLSTM}^{\mathrm{N}}(\{\bar{\mathbf{x}}_{t,c}\}_{t=1}^{T}), \tag{20}$$

$$\mathbf{m}_{t,c}^{\mathrm{N}} = \mathrm{sigmoid}(\mathbf{W}^{\mathrm{N}}\mathbf{z}_{t,c}^{\mathrm{N}} + \mathbf{b}^{\mathrm{N}}), \tag{21}$$

where $Z_c^{\mathrm{S}} = \{\mathbf{z}_{t,c}^{\mathrm{S}} \in \mathbb{R}^{D_{\mathrm{Z}}} | t = 1, \cdots, T\}$ is the output sequence of D_{Z}-dimensional vectors of the BLSTM network to obtain a speech mask over c-th channel's input STFTs. Z_c^{N} is the BLSTM output sequence for a noise mask. $\bar{\mathbf{x}}_{t,c} = \{\Re(x_{t,f,c}), \Im(x_{t,f,c})\}_{f=1}^{F} \in \mathbb{R}^{2F}$ is an input feature of a $2F$-dimensional real-value vector. This is obtained by concatenating the real and imaginary parts of all STFT features at c-th channel. $\mathbf{m}_{t,c}^{\mathrm{S}} = \{m_{t,f,c}^{\mathrm{S}}\}_{f=1}^{F} \in [0,1]^F$ and $\mathbf{m}_{t,c}^{\mathrm{N}}$ are the estimated speech and noise masks for every c-th channel at a time step t, respectively. $\mathbf{W}^{\mathrm{S}}, \mathbf{W}^{\mathrm{N}} \in \mathbb{R}^{F \times D_{\mathrm{Z}}}$ are the weight matrices of the output layers to finally output speech and noise masks, respectively, and $\mathbf{b}^{\mathrm{S}}, \mathbf{b}^{\mathrm{N}} \in \mathbb{R}^F$ are their corresponding bias vectors.

After computing the speech and noise masks for each channel, the averaged masks are obtained as follows:

$$\mathbf{m}_t^{\mathrm{S}} = \frac{1}{C}\sum_{c=1}^{C} \mathbf{m}_{t,c}^{\mathrm{S}}, \quad \mathbf{m}_t^{\mathrm{N}} = \frac{1}{C}\sum_{c=1}^{C} \mathbf{m}_{t,c}^{\mathrm{N}}. \tag{22}$$

We use these averaged masks to estimate the PSD matrices as described in Eqs. (16) and (17). The MVDR beamformer through this BLSTM mask estimation is originally proposed in (Heymann et al., 2016), but our neural beamformer further extends it with attention-based reference selection, which is described in the next subsection.

3.2.3. Attention-based reference selection

To incorporate the reference microphone selection in a neural beamformer framework, we use a soft-max for the vector $\mathbf{u}$ in Eq. (15) derived from an attention mechanism. In this approach, the reference microphone vector $\mathbf{u}$ is estimated from time-invariant feature vectors $\mathbf{q}_c$ and $\mathbf{r}_c$ as follows:

$$\tilde{k}_c = \mathbf{v}^{\mathrm{T}} \tanh(\mathbf{V}^{\mathrm{Q}}\mathbf{q}_c + \mathbf{V}^{\mathrm{R}}\mathbf{r}_c + \tilde{\mathbf{b}}), \tag{23}$$

$$u_c = \frac{\exp(\beta \tilde{k}_c)}{\sum_{c=1}^{C} \exp(\beta \tilde{k}_c)}, \tag{24}$$

where $\mathbf{v} \in \mathbb{R}^{1 \times D_{\mathrm{V}}}, \mathbf{V}^{\mathrm{Z}} \in \mathbb{R}^{D_{\mathrm{V}} \times 2D_{\mathrm{Z}}}, \mathbf{V}^{\mathrm{R}} \in \mathbb{R}^{D_{\mathrm{V}} \times 2F}$ are trainable weight parameters, $\tilde{\mathbf{b}} \in \mathbb{R}^{D_{\mathrm{V}}}$ is a trainable bias vector. β is the sharpening factor. We use two types of features; 1) the time-averaged state vector $\mathbf{q}_c \in \mathbf{R}^{2D_{\mathrm{Z}}}$ extracted from the BLSTM networks for speech and noise masks in Eqs. (18) and (20), i.e.,

$$\mathbf{q}_c = \frac{1}{T}\sum_{t=1}^{T} \{\mathbf{z}_{t,c}^{\mathrm{S}}, \mathbf{z}_{t,c}^{\mathrm{N}}\}, \tag{25}$$

and 2) the PSD feature $\mathbf{r}_c \in \mathbb{R}^{2F}$, which incorporates the spatial information into the attention mechanism. The following equation represents how to compute $\mathbf{r}_c$:

$$\mathbf{r}_c = \frac{1}{C-1}\sum_{c'=1, c' \neq c}^{C} \{\Re(\phi_{f,c,c'}^{\mathrm{S}}), \Im(\phi_{f,c,c'}^{\mathrm{S}})\}_{f=1}^{F}, \tag{26}$$

where $\phi_{f,c,c'}^{S} \in \mathbb{C}$ is the entry in c-th row and c'-th column of the speech PSD matrix $\mathbf{\Phi}^{\mathrm{S}}(f)$ in Eq. (16). The PSD matrix represents correlation information between channels. To select a reference microphone, the spatial correlation related to speech signals is more informative, and therefore, we only use the speech PSD matrix $\mathbf{\Phi}^{\mathrm{S}}(f)$ as a feature.

Note that, in this mask estimation based MVDR beamformer, masks for each channel are computed separately using the same BLSTM network unlike Eq. (12), and the mask estimation network is independent of channels. Similarly, the reference selection network is also independent of channels, and the beamformer deals with input signals with arbitrary number and order of channels without re-training or re-configuration of the network.

4. Multichannel end-to-end ASR

In this work, we propose a multichannel end-to-end speech recognition, which integrates all components with a single neural architecture. We adopt neural beamformers (Section 3) as a speech enhancement part, and the attention-based encoder-decoder (Section 2) as a speech recognition part.

The entire procedure to generate the sequence of output labels $\hat{Y}$ from the multichannel inputs $\{X_c\}_{c=1}^{C}$ is formalized as follows:

$$\hat{X} = \mathrm{Enhance}(\{X_c\}_{c=1}^{C}), \tag{27}$$

$$\hat{O} = \mathrm{Feature}(\hat{X}), \tag{28}$$

$$\hat{H} = \mathrm{Encoder}(\hat{O}), \tag{29}$$

$$\hat{\mathbf{c}}_n = \mathrm{Attention}(\hat{\mathbf{a}}_{n-1}, \hat{\mathbf{s}}_n, \hat{H}), \tag{30}$$

$$\hat{y}_n = \mathrm{Decoder}(\hat{\mathbf{c}}_n, \hat{y}_{1:n-1}). \tag{31}$$

$\mathrm{Enhance}(\cdot)$ is a speech enhancement function realized by the neural beamformer based on Eq. (11) with the filter or mask estimation network (Section 3.1 or 3.2).

$\mathrm{Feature}(\cdot)$ is a feature extraction function. In this work, we use a normalized log Mel filterbank transform to obtain $\hat{\mathbf{o}}_t \in \mathbb{R}^{D_{\mathrm{O}}}$ computed from the enhanced STFT coefficients $\hat{\mathbf{x}}_t \in \mathbb{C}^F$ as an input of attention-based encoder-decoder:

$$\mathbf{p}_t = \{\Re(\hat{x}_{t,f})^2 + \Im(\hat{x}_{t,f})^2\}_{f=1}^{F}, \tag{32}$$

$$\hat{\mathbf{o}}_t = \mathrm{Norm}(\log(\mathrm{Mel}(\mathbf{p}_t))), \tag{33}$$

where $\mathbf{p}_t \in \mathbb{R}^F$ is a real-valued vector of the power spectrum of the enhanced signal at a time step t, $\mathrm{Mel}(\cdot)$ is the operation of $D_{\mathrm{O}} \times F$ Mel matrix multiplication, and $\mathrm{Norm}(\cdot)$ is the operation of global mean and variance normalization so that its mean and variance become 0 and 1.

$\mathrm{Encoder}(\cdot)$, $\mathrm{Attention}(\cdot)$, and $\mathrm{Decoder}(\cdot)$ are defined in Eqs. (7), (8), and (9), respectively, with the sequence of the enhanced log Mel filterbank like features $\hat{O}$ as an input.

Thus, we can build a multichannel end-to-end speech recognition system, which converts multichannel speech signals to texts with a single network. Note that because all procedures, such as enhancement, feature extraction, encoder, attention, and decoder, are connected with differentiable graphs, we can optimize the overall inference to generate a correct label sequence.

Relation to prior works

There have been several related studies of neural beamformers based on the filter estimation (Li et al., 2016; Xiao et al., 2016a; Meng et al., 2017) and the mask estimation (Heymann et al., 2016; Erdogan et al., 2016; Xiao et al., 2016b). The main difference is that such preceding studies use a component-level training objective within the conventional hybrid frameworks, while our work focuses on the entire end-to-end objective. For example, Heymann et al., 2016; Erdogan et al., 2016 use a signal-level objective (binary mask classification or regression) to train a network given parallel clean and noisy speech data. Li et al., 2016; Xiao et al., 2016a; Meng et al., 2017; Xiao et al., 2016b use ASR objectives (HMM state classification or sequence discriminative training), but they are still based on the hybrid approach. Speech recognition with raw multichannel waveforms (Hoshen et al., 2015; Sainath et al., 2016) can also be seen as using a neural beamformer, where the filter coefficients are represented as network parameters, but again these methods are still based on the hybrid approach.

As regards end-to-end speech recognition, all existing studies are based on a single channel setup. For example, most studies focus on a standard clean speech recognition setup without speech enhancement. (Chorowski et al., 2014; Graves & Jaitly, 2014; Chorowski et al., 2015; Chan et al., 2016; Miao et al., 2015; Zhang et al., 2016; Kim et al., 2017; Lu et al., 2016). Amodei et al., 2016 discusses end-to-end speech recognition in a noisy environment, but this method deals with the noise robustness by preparing various types of simulated noisy speech for training data, and does not incorporate multichannel speech enhancement in their networks.

5. Experiments

We study the effectiveness of our multichannel end-to-end system compared to a baseline end-to-end system with noisy speech or beamformed inputs. We use the two multichannel speech recognition benchmarks, CHiME-4 (Vincent et al., 2016) and AMI (Hain et al., 2007).

CHiME-4 is a speech recognition task in public noisy environments, consisting of speech recorded using a tablet device with 6-channel microphones. It consists of real and simulated data. The training set consists of 3 hours of real speech data uttered by 4 speakers and 15 hours of simulation speech data uttered by 83 speakers. The development set consists of 2.9 hours of real and simulation speech data uttered by 4 speakers, respectively. The evaluation set consists of 2.2 hours of real and simulation speech data uttered by 4 speakers, respectively. We excluded the 2nd channel signals, which is captured at the microphone located on the backside of the tablet, and used 5 channels for the following multichannel experiments ($C = 5$).

AMI is a speech recognition task in meetings, consisting of speech recorded using 8-channel circular microphones ($C = 8$). It consists of only real data. The training set consists of about 78 hours of speech data uttered by 135 speakers. The development and evaluation sets consist of about 9 hours of speech data uttered by 18 and 16 speakers, respectively. The amount of training data (i.e., 78 hours) is larger than one for CHiME-4 (i.e., 18 hours), and we mainly used CHiME-4 data to demonstrate our experiments.

5.1. Configurations

5.1.1. Encoder-decoder networks

We used 40-dimensional log Mel filterbank coefficients as an input feature vector for both noisy and enhanced speech signals ($D_{\mathrm{O}} = 40$). In this experiment, we used 4-layer BLSTM with 320 cells in the encoder ($D_{\mathrm{H}} = 320$), and 1-layer LSTM with 320 cells in the decoder ($D_{\mathrm{S}} = 320$). In the encoder, we subsampled the hidden states of the first and second layers and used every second of hidden states for the subsequent layer's inputs. Therefore, the number of hidden states at the encoder's output layer is reduced to $L = T/4$. After every BLSTM layer, we used a linear projection layer with 320 units to combine the forward and backward LSTM outputs. For the attention mechanism, 10 centered convolution filters ($D_{\mathrm{F}} = 10$) of width 100 ($D_{\mathrm{f}} = 100$) were used to extract the convolutional features. We set the attention inner product dimension as 320 ($D_{\mathrm{W}} = 320$), and used the sharpening factor $\alpha = 2$. To boost the optimization in a noisy environment, we adopted a joint CTC-attention multi-task loss function (Kim et al., 2017), and set the CTC loss weight as 0.1.

For decoding, we used a beam search algorithm similar to (Sutskever et al., 2014) with the beam size 20 at each output step to reduce the computation cost. CTC scores were also used to re-score the hypotheses with 0.1 weight. We adopted a length penalty term (Chorowski et al., 2015) to the decoding objective and set the penalty weight as 0.3. In the CHiME-4 experiments, we only allowed the hypotheses whose length were within $0.3 \times L$ and $0.75 \times L$ during decoding, while the hypothesis lengths in the AMI experiments were automatically determined based on the above scores. Note that we pursued a pure end-to-end setup without using any external lexicon or language models, and

Table 1. Character error rate [%] for CHiME-4 corpus.

MODEL	DEV SIMU	DEV REAL	EVAL SIMU	EVAL REAL
NOISY	25.0	24.5	34.7	35.8
BEAMFORMIT	21.5	19.3	31.2	28.2
FILTER_NET	19.1	20.3	28.2	32.7
MASK_NET (REF)	15.5	18.6	**23.7**	28.8
MASK_NET (ATT)	**15.3**	**18.2**	**23.7**	**26.8**

Table 2. Character error rate [%] for AMI corpus.

MODEL	DEV	EVAL
NOISY	41.8	45.3
BEAMFORMIT	44.9	51.3
MASK_NET (ATT)	**35.7**	**39.0**

used CER as an evaluation metric.

5.1.2. NEURAL BEAMFORMERS

256 STFT coefficients and the offset were computed from 25ms-width hamming window with 10ms shift ($F = 257$). Both filter and mask estimation network approaches used similar a 3-layer BLSTM with 320 cells ($D_{\mathrm{Z}} = 320$) without the subsampling technique. For the reference selection attention mechanism, we used the same attention inner product dimension ($D_{\mathrm{V}} = 320$) and sharpening factor $\beta = 2$ as those of the encoder-decoder network.

5.1.3. SHARED CONFIGURATIONS

All the parameters are initialized with the range [-0.1, 0.1] of a uniform distribution. We used the AdaDelta algorithm (Zeiler, 2012) with gradient clipping (Pascanu et al., 2013) for optimization. We initialized the AdaDelta hyperparameters $\rho = 0.95$ and $\epsilon = 1^{-8}$. Once the loss over the validation set was degraded, we decreased the AdaDelta hyperparameter ϵ by multiplying it by 0.01 at each subsequent epoch. The training procedure was stopped after 15 epochs. During the training, we adopted multi-condition training strategy, i.e., in addition to the optimization with the enhanced features through the neural beamformers, we also used the noisy multichannel speech data as an input of encoder-decoder networks without through the neural beamformers. Preliminary experiments showed that the multi-condition training is essential to improve the robustness of the encoder-decoder networks. Note that we trained the entire network from scratch without any pre-training procedures. All the above networks are implemented by using Chainer (Tokui et al., 2015).

5.2. Results

Table 1 shows the recognition performances of CHiME-4 with the five systems: NOISY, BEAMFORMIT, FILTER_NET, MASK_NET (REF), and MASK_NET (ATT). NOISY and BEAMFORMIT were the baseline single-channel end-to-end systems, which did not include the speech enhancement part in their frameworks. Their end-to-end networks were trained only with noisy speech data by following a conventional multi-condition training strategy (Vincent et al., 2016). During decoding, NOISY used single-channel noisy speech data from 'isolated 1ch track' in CHiME-4 as an input, while BEAMFORMIT used the enhanced speech data obtained from 5-channel signals with BeamformIt (Anguera et al., 2007), which is well-known delay-and-sum beamformer, as an input.

FILTER_NET, MASK_NET (REF), and MASK_NET (ATT) were the multichannel end-to-end systems described in Section 4. To evaluate the validity of the reference selection, we prepared MASK_NET (ATT) based on the mask-based beamformer with attention-based reference selection described in Section 3.2.3, and MASK_NET (REF) with 5-th channel as a fixed reference microphone, which is located on the center front of the tablet device.

Table 1 shows that BEAMFORMIT, FILTER_NET, MASK_NET (REF), and MASK_NET (ATT) outperformed NOISY, which confirms the effectiveness of combining speech enhancement with the attention-based encoder-decoder framework. The comparison of MASK_NET (REF) and MASK_NET (ATT) validates the use of the attention-based mechanism for reference selection. FILTER_NET, which is based on the filter estimation network described in Section 3.1, also improved the performance compared to NOISY, but worse than MASK_NET (ATT). This is because it is difficult to optimize the filter estimation network due to a lack of restriction to estimate filter coefficients, and it needs some careful optimization, as suggested by (Xiao et al., 2016a). Finally, MASK_NET (ATT) achieved better recognition performance than BEAMFORMIT, which proves the effectiveness of our joint integration rather than a pipe-line combination of speech enhancement and (end-to-end) speech recognition.

To further investigate the effectiveness of our proposed multichannel end-to-end framework, we also conducted the experiment on the AMI corpus. Table 2 compares the recognition performance of the three systems: NOISY, BEAMFORMIT, and MASK_NET (ATT). In NOISY, we used noisy speech data from the 1st channel in AMI as an input to the system. Table 2 shows that, even in the AMI, our proposed MASK_NET (ATT) achieved better recognition performance than the attention-based baselines (NOISY and BEAMFORMIT), which also confirms the effectiveness of our proposed multichannel end-to-end

Table 3. CHiME-4 validation accuracies [%] for MASK_NET (ATT) with different numbers and orders of channels.

MODEL	CHANNEL	DEV
NOISY	ISOLATED_1CH_TRACK	87.9
MASK_NET (ATT)	5_6_4_3_1	91.2
MASK_NET (ATT)	3_4_1_5_6	91.2
MASK_NET (ATT)	5_6_4_1	91.1
MASK_NET (ATT)	5_6_4	90.9

framework. Note that BEAMFORMIT was worse than NOISY even with the enhanced signals. This phenomenon is sometimes observed in noisy speech recognition that the distortion caused by sole speech enhancement degrades the performance without re-training. Our end-to-end system jointly optimizes the speech enhancement part with the ASR objective, and can avoid such degradations.

5.3. Influence on the number and order of channels

As we discussed in Section 3.2, one unique characteristic of our proposed MASK_NET (ATT) is the robustness/invariance against the number and order of channels without re-training. Table 3 shows an influence of the CHiME-4 validation accuracies on the number and order of channels. The validation accuracy was computed conditioned on the ground truth labels $y^*_{1:n-1}$ in Eq. (10) during decoder's recursive character generation, which has a strong correlation with CER. The second column of the table represents the channel indices, which were used as an input of the same MASK_NET (ATT) network.

Comparison of $5\_6\_4\_3\_1$ and $3\_4\_1\_5\_6$ shows that the order of channels did not affect the recognition performance of MASK_NET (ATT) at all, as we expected. In addition, even when we used fewer three or four channels as an input, MASK_NET (ATT) still outperformed NOISY (single channel). These results confirm that our proposed multichannel end-to-end system can deal with input signals with arbitrary number and order of channels, without any reconfiguration and re-training.

5.4. Visualization of beamformed features

To analyze the behavior of our developed speech enhancement component with a neural beamformer, Figure 4 visualizes the spectrograms of the same CHiME-4 utterance with the 5-th channel noisy signal, enhanced signal with BeamformIt, and enhanced signal with our proposed MASK_NET (ATT). We could confirm that the BeamformIt and MASK_NET (ATT) successfully suppressed the noises comparing to the 5-th channel signal by eliminating blurred red areas overall. In addition, by focusing on

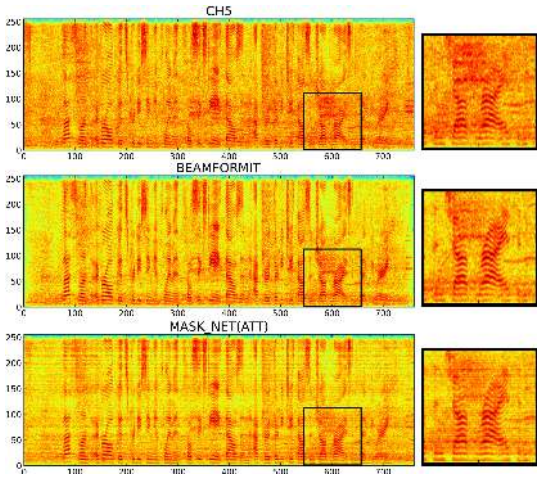


Figure 4. Comparison of the log-magnitude spectrograms of the same CHiME-4 utterance .

the insides of black boxes, the harmonic structure, which was corrupted in the 5-th channel signal, was recovered in BeamformIt and MASK_NET (ATT).

This result suggests that our proposed MASK_NET (ATT) successfully learned a noise suppression function similar to the conventional beamformer, although it is optimized based on the end-to-end ASR objective, without explicitly using clean data as a target.

6. Conclusions

In this paper, we extended an existing attention-based encoder-decoder framework by integrating a neural beamformer and proposed a multichannel end-to-end speech recognition framework. It can jointly optimize the overall inference in multichannel speech recognition (i.e., from speech enhancement to speech recognition) based on the end-to-end ASR objective, and it can generalize to different numbers and configurations of microphones. Our proposed architecture will potentially expand the scope of application of existing single-channel sequence-to-sequence problems to multichannel sequence-to-sequence problems. The experimental results on challenging noisy speech recognition benchmarks, CHiME-4 and AMI, show that the proposed framework outperformed the end-to-end baseline with noisy and delay-and-sum beamformed inputs.

The current system still has data sparseness issues due to the lack of lexicon and language models, unlike the conventional hybrid approach. Therefore, the results reported in the paper did not reach the state-of-the-art performance in these benchmarks, but they are still convincing to show the effectiveness of the proposed framework. Our most important future work is to overcome these data sparseness issues by developing adaptation techniques of an end-to-end framework with the incorporation of linguistic resources.

Acknowledgements

Tsubasa Ochiai was supported by JSPS Grants-in-Aid for Scientific Research No. 26280063. Shinji Watanabe, Takaaki Hori, and John Hershey were supported by MERL.

References

Amodei, Dario, Anubhai, Rishita, Battenberg, Eric, Case, Carl, Casper, Jared, Catanzaro, Bryan, Chen, Jingdong, Chrzanowski, Mike, Coates, Adam, Diamos, Greg, et al. Deep speech 2: End-to-end speech recognition in English and Mandarin. *International Conference on Machine Learning (ICML)*, pp. 173–182, 2016.

Anguera, Xavier, Wooters, Chuck, and Hernando, Javier. Acoustic beamforming for speaker diarization of meetings. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(7):2011–2022, 2007.

Bahdanau, Dzmitry, Chorowski, Jan, Serdyuk, Dmitriy, Brakel, Philemon, and Bengio, Yoshua. End-to-end attention-based large vocabulary speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4945–4949, 2016.

Barker, Jon, Marxer, Ricard, Vincent, Emmanuel, and Watanabe, Shinji. The third 'CHiME' speech separation and recognition challenge: Analysis and outcomes. *Computer Speech & Language*, 2016.

Benesty, Jacob, Chen, Jingdong, and Huang, Yiteng. *Microphone array signal processing*, volume 1. Springer Science & Business Media, 2008.

Bourlard, Hervé and Morgan, Nelson. *Connectionist speech recognition: A hybrid approach*. Kluwer Academic Publishers, 1994.

Capon, Jack. High-resolution frequency-wavenumber spectrum analysis. *Proceedings of the IEEE*, 57(8): 1408–1418, 1969.

Chan, William, Jaitly, Navdeep, Le, Quoc, and Vinyals, Oriol. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4960–4964, 2016.

Chorowski, Jan, Bahdanau, Dzmitry, Cho, Kyunghyun, and Bengio, Yoshua. End-to-end continuous speech recognition using attention-based recurrent NN: First results. *arXiv preprint arXiv:1412.1602*, 2014.

Chorowski, Jan K, Bahdanau, Dzmitry, Serdyuk, Dmitriy, Cho, Kyunghyun, and Bengio, Yoshua. Attention-based models for speech recognition. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 577–585, 2015.

Erdogan, Hakan, Hershey, John R, Watanabe, Shinji, Mandel, Michael, and Le Roux, Jonathan. Improved MVDR beamforming using single-channel mask prediction networks. In *Interspeech*, pp. 1981–1985, 2016.

Graves, Alex and Jaitly, Navdeep. Towards end-to-end speech recognition with recurrent neural networks. In *International Conference on Machine Learning (ICML)*, pp. 1764–1772, 2014.

Hain, Thomas, Burget, Lukas, Dines, John, Garau, Giulia, Wan, Vincent, Karafi, Martin, Vepa, Jithendra, and Lincoln, Mike. The AMI system for the transcription of speech in meetings. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 357–360, 2007.

Heymann, Jahn, Drude, Lukas, and Haeb-Umbach, Reinhold. Neural network based spectral mask estimation for acoustic beamforming. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 196–200, 2016.

Hinton, Geoffrey, Deng, Li, Yu, Dong, Dahl, George E, Mohamed, Abdel-rahman, Jaitly, Navdeep, Senior, Andrew, Vanhoucke, Vincent, Nguyen, Patrick, Sainath, Tara N, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29 (6):82–97, 2012.

Hoshen, Yedid, Weiss, Ron J, and Wilson, Kevin W. Speech acoustic modeling from raw multichannel waveforms. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4624–4628, 2015.

Jelinek, Frederick. Continuous speech recognition by statistical methods. *Proceedings of the IEEE*, 64(4):532–556, 1976.

Kim, Suyoun, Hori, Takaaki, and Watanabe, Shinji. Joint CTC-attention based end-to-end speech recognition using multi-task learning. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4835–4839, 2017.

Kinoshita, Keisuke, Delcroix, Marc, Gannot, Sharon, Habets, Emanuël AP, Haeb-Umbach, Reinhold, Kellermann, Walter, Leutnant, Volker, Maas, Roland, Nakatani, Tomohiro, Raj, Bhiksha, et al. A summary of the REVERB challenge: state-of-the-art and remaining challenges in reverberant speech processing research. *EURASIP Journal on Advances in Signal Processing*, 2016(1):1–19, 2016.

Li, Bo, Sainath, Tara N, Weiss, Ron J, Wilson, Kevin W, and Bacchiani, Michiel. Neural network adaptive beamforming for robust multichannel speech recognition. In *Interspeech*, pp. 1976–1980, 2016.

Lu, Liang, Zhang, Xingxing, and Renals, Steve. On training the recurrent neural network encoder-decoder for large vocabulary end-to-end speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5060–5064, 2016.

Meng, Zhong, Watanabe, Shinji, Hershey, John R, and Erdogan, Hakan. Deep long short-term memory adaptive beamforming networks for multichannel robust speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 271–275, 2017.

Miao, Yajie, Gowayyed, Mohammad, and Metze, Florian. EESEN: End-to-end speech recognition using deep RNN models and WFST-based decoding. In *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 167–174, 2015.

Pascanu, Razvan, Mikolov, Tomas, and Bengio, Yoshua. On the difficulty of training recurrent neural networks. *International Conference on Machine Learning (ICML)*, pp. 1310–1318, 2013.

Sainath, Tara N, Narayanan, Arun, Weiss, Ron J, Variani, Ehsan, Wilson, Kevin W, Bacchiani, Michiel, and Shafran, Izhak. Reducing the computational complexity of multimicrophone acoustic models with integrated feature extraction. In *Interspeech*, pp. 1971–1975, 2016.

Souden, Mehrez, Benesty, Jacob, and Affes, Sofiène. On optimal frequency-domain multichannel linear filtering for noise reduction. *IEEE Transactions on audio, speech, and language processing*, 18(2):260–276, 2010.

Sutskever, Ilya, Vinyals, Oriol, and Le, Quoc V. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems (NIPS)*, pp. 3104–3112, 2014.

Tokui, Seiya, Oono, Kenta, Hido, Shohei, and Clayton, Justin. Chainer: a next-generation open source framework for deep learning. In *Proceedings of Workshop on Machine Learning Systems (LearningSys) in NIPS*, 2015.

Van Veen, Barry D and Buckley, Kevin M. Beamforming: A versatile approach to spatial filtering. *IEEE ASSP Magazine*, 5(2):4–24, 1988.

Vincent, Emmanuel, Watanabe, Shinji, Nugraha, Aditya Arie, Barker, Jon, and Marxer, Ricard. An analysis of environment, microphone and data simulation mismatches in robust speech recognition. *Computer Speech & Language*, 2016.

Wölfel, Matthias and McDonough, John. *Distant speech recognition*. John Wiley & Sons, 2009.

Xiao, Xiong, Watanabe, Shinji, Erdogan, Hakan, Lu, Liang, Hershey, John, Seltzer, Michael L, Chen, Guoguo, Zhang, Yu, Mandel, Michael, and Yu, Dong. Deep beamforming networks for multi-channel speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5745–5749, 2016a.

Xiao, Xiong, Xu, Chenglin, Zhang, Zhaofeng, Zhao, Shengkui, Sun, Sining, and Watanabe, Shinji. A study of learning based beamforming methods for speech recognition. In *CHiME 2016 workshop*, pp. 26–31, 2016b.

Yoshioka, Takuya, Ito, Nobutaka, Delcroix, Marc, Ogawa, Atsunori, Kinoshita, Keisuke, Fujimoto, Masakiyo, Yu, Chengzhu, Fabian, Wojciech J, Espi, Miquel, Higuchi, Takuya, et al. The NTT CHiME-3 system: Advances in speech enhancement and recognition for mobile multimicrophone devices. In *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 436–443, 2015.

Zeiler, Matthew D. ADADELTA: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.

Zhang, Yu, Chan, William, and Jaitly, Navdeep. Very deep convolutional networks for end-to-end speech recognition. *arXiv preprint arXiv:1610.03022*, 2016.