Does speech enhancement work with end-to-end ASR objectives?: Experimental analysis of multichannel end-to-end ASR

Ochiai, T.; Watanabe, S.; Katagiri, S.

TR2017-139 October 2017

Abstract

Recently we proposed a novel multichannel end-to-end speech recognition architecture that integrates the components of multichannel speech enhancement and speech recognition into a single neural-network-based architecture and demonstrated its fundamental utility for automatic speech recognition (ASR). However, the behavior of the proposed integrated system remains insufficiently clarified. An open question is whether the speech enhancement component really gains speech enhancement (noise suppression) ability, because it is optimized based on end-to-end ASR objectives instead of speech enhancement objectives. In this paper, we solve this question by conducting systematic evaluation experiments using the CHiME-4 corpus. We first show that the integrated end-to-end architecture successfully obtains adequate speech enhancement ability that is superior to that of a conventional alternative (a delay-and-sum beamformer) by observing two signal-level measures: the signal-todistortion ratio and the perceptual evaluation of speech quality. Our findings suggest that to further increase the performances of an integrated system, we must boost the power of the latter-stage speech recognition component. However, an insufficient amount of multichannel noisy speech data is available. Based on these situations, we next investigate the effect of using a large amount of single-channel clean speech data, e.g., the WSJ corpus, for additional training of the speech recognition component. We also show that our approach with clean speech significantly improves the total performance of multichannel end-to-end architecture in the multichannel noisy ASR tasks.

IEEE International Workshop on Machine Learning for Signal Processing (MLSP)

2017 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, SEPT. 25–28, 2017, TOKYO, JAPAN

DOES SPEECH ENHANCEMENT WORK WITH END-TO-END ASR OBJECTIVES?: EXPERIMENTAL ANALYSIS OF MULTICHANNEL END-TO-END ASR

Tsubasa Ochiai[1], *Shinji Watanabe*[2], *Shigeru Katagiri*[1]

[1] Doshisha University
[2] Mitsubishi Electric Research Laboratories

ABSTRACT

Recently we proposed a novel multichannel end-to-end speech recognition architecture that integrates the components of multichannel speech enhancement and speech recognition into a single neural-network-based architecture and demonstrated its fundamental utility for automatic speech recognition (ASR). However, the behavior of the proposed integrated system remains insufficiently clarified. An open question is whether the speech enhancement component really gains speech enhancement (noise suppression) ability, because it is optimized based on end-to-end ASR objectives instead of speech enhancement objectives. In this paper, we solve this question by conducting systematic evaluation experiments using the CHiME-4 corpus. We first show that the integrated end-to-end architecture successfully obtains adequate speech enhancement ability that is superior to that of a conventional alternative (a delay-and-sum beamformer) by observing two signal-level measures: the signal-to-distortion ratio and the perceptual evaluation of speech quality. Our findings suggest that to further increase the performances of an integrated system, we must boost the power of the latter-stage speech recognition component. However, an insufficient amount of multichannel noisy speech data is available. Based on these situations, we next investigate the effect of using a large amount of single-channel clean speech data, e.g., the WSJ corpus, for additional training of the speech recognition component. We also show that our approach with clean speech significantly improves the total performance of multichannel end-to-end architecture in the multichannel noisy ASR tasks.

Index Terms— Multichannel end-to-end automatic speech recognition, neural beamformer, encoder-decoder network

1. INTRODUCTION

Motivated by the recent rise of Deep Neural Network (DNN) technologies, the hybrid Automatic Speech Recognition (ASR) system architecture that combines DNN and the Hidden Markov Model (HMM) has attracted great research interest [1]. Usually in this approach, DNN and HMM are separately developed and simply combined. However, such a simple combination does not necessarily lead to the best performances, and it raises a question about if there can be other types of approaches, e.g., a tightly-coupled combination and the integrated design of the overall system. To challenge this question, an alternative to using DNN-HMM hybrid architecture, i.e., an end-to-end development scheme for ASR systems, was proposed whose feasibility was studied in tasks of recognizing clean speech inputs [2, 3].

In reality, speech inputs for ASR systems are generally contaminated by background noise and reverberation. Therefore, emerging end-to-end design schemes are obviously expected to run well for noisy speech inputs. In this light, we proposed a novel end-to-end development scheme that encompasses the entire ASR system, i.e., its components of multichannel speech enhancement and recognition, which directly transforms continuous speech inputs to text character sequence outputs. This is realized by integrating a neural-network-based multichannel speech enhancement technique, which we refer to as neural beamformer [4, 5], (for the enhancement component) and an attention-based encoder-decoder framework [2] (for the recognition component). For clarity, we refer to our scheme as the Multichannel End-to-End (M-E2E) ASR framework [6].

To evaluate our M-E2E framework, we experimentally compared it with a combination of the preceding end-to-end framework, which assumes single-channel speech inputs and does not include speech enhancement functions within the framework, and the delay-and-sum beamformer speech enhancement method, i.e., BeamformIt [7], in the widely adopted multichannel noisy speech benchmark tasks of CHiME-4 [8] and AMI [9]. For simplicity, we refer to the competitor as the End-to-End framework attached by BeamformIt (E2E-BIt). Experimental results showed that our M-E2E approach outperforms E2E-BIt in terms of such ASR-oriented measures as the Character Error Rate (CER). However, the behavior or the characteristics of M-E2E-based ASR systems have not yet been sufficiently clarified. An important open question remains whether the M-E2E framework realizes speech enhancement functions inside fully neural-network-based ASR systems. In addition, the achieved recognition accuracies of M-E2E-based systems are still lower than the conventional DNN-HMM hybrid systems. Therefore, in this paper we elaborate the behavior of the ASR systems developed by the M-E2E framework and also aim at improving their recognition performances. The concrete research issues in the paper are summarized as follows:

1. So far, integrated training based on the M-E2E framework for fully neural-network-based ASR systems has only been conducted using ASR-oriented training criterion. Because the criterion has no direct link with speech enhancement, it remains unknown whether the trained ASR system really gained speech enhancement ability in its front-end multichannel speech enhancement component (neural beamformer). This question is also re-expressed as whether the M-E2E-based training actually extracts speech components (or their corresponding representation) from noisy speech inputs or simply converts the inputs into something useful for ASR. To answer these questions, we analyze the inner speech representation produced by the M-E2E-based systems using

Tsubasa Ochiai and Shigeru Katagiri was supported in part by JSPS Grants-in-Aid for Scientific Research No. 26280063, MEXT-Supported Program Driver-in-the-Loop, and Grant-in-Aid for JSPS Fellows. Shinji Watanabe was supported by MERL.

the signal-level criterion, i.e., the Signal-to-Distortion Ratio (SDR) and Perceptual Evaluation of Speech Quality (PESQ), which are commonly used for speech enhancement quality assessment.

2. If the M-E2E framework implements speech enhancement (noise suppression) inside fully neural-network-based systems, are the resulting enhanced speech signals clean enough to achieve the same level of ASR-oriented performances (in terms of Word Error Rate (WER)) as DNN-HMM hybrid systems?
3. Compared to the training of the DNN-based acoustic models in the DNN-HMM hybrid framework, M-E2E-based training essentially requires more training data, because M-E2E training must also learn regularities in language. However, the amount of data in the existing multichannel noisy ASR corpora, e.g., CHiME-4, is often much smaller than such single-channel clean speech datasets as the WSJ corpus [10]. In addition, it is essentially difficult in the end-to-end development framework to use a large amount of text data, which is available nowadays and critically important for gaining effective language models in the DNN-HMM hybrid framework: The end-to-end framework basically requires speech inputs. Considering this practical restriction, we study how to solve the insufficiency of the training speech data to improve the decoding quality of fully neural-network-based systems in the multichannel noisy ASR tasks.

2. MULTICHANNEL END-TO-END ASR

2.1. Overview

Figure 1 illustrates an overview of the ASR system architecture based on our M-E2E ASR framework. The system consists of a mask-based neural beamformer and an attention-based encoder-decoder network. Given multichannel noisy speech inputs $\{X_c\}_{c=1}^{C}$, where C is the number of channels and X_c is the c-th channel input that consists of a short-time Fourier transform (STFT) feature sequence, the system first filters and integrates the multichannel inputs into a single-channel (hopefully noise-suppressed) input in the mask-based neural beamformer stage, converts the filtered input to a sequence of logarithmic Mel-scale filterbank (LMF) outputs $\hat{O}$ in the feature extraction stage, and transforms (decodes) the sequence of LFM features to such class labels Y as the sequence of characters by the estimation of the *a posteriori* probabilities in the final, attention-based encoder-decoder stage.

We represent the entire procedure of the M-E2E-based system in the following functional forms:

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

$$\hat{O} = \text{Feature}(\hat{X}), \tag{2}$$

$$P(Y|\hat{O}) = \text{E2E\_ASR}(\hat{O}). \tag{3}$$

Here, $\text{Enhance}(\cdot)$, which is a speech enhancement function realized by the mask-based neural beamformer, converts multichannel STFT feature inputs $\{X_c\}_{c=1}^{C}$ to a sequence of single-channel enhanced STFT features $\hat{X}$. $\text{Feature}(\cdot)$ is a feature extraction function, which bridges the speech enhancement and ASR components. In this paper, we adopt the normalized LMF function for $\text{Feature}(\cdot)$, which converts $\hat{X}$ to $\hat{O}$. Subsequently, $\text{E2E\_ASR}(\cdot)$, which is an end-to-end ASR function realized by the attention-based encoder-decoder, estimates the *a posteriori* probabilities for output labels Y.

All of the above steps are represented as differentiable graphs. We train/optimize them only using the training sample pairs, each of which consists of multichannel noisy speech samples and its corresponding class labels, to satisfy such ASR-oriented criteria as CER as much as possible.

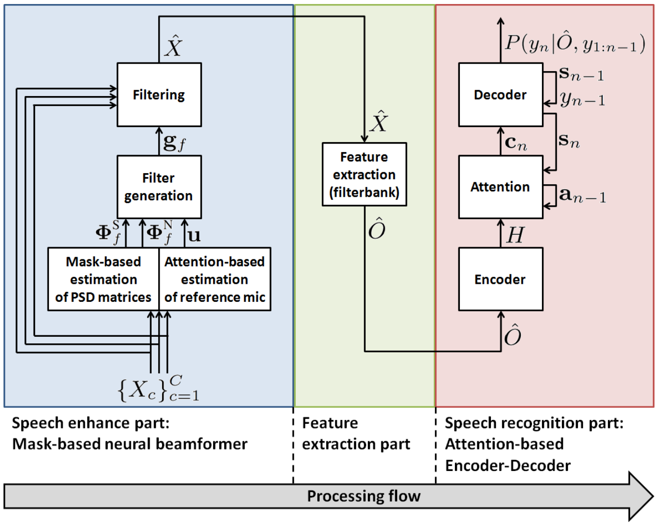
$\hat{X}$
Filtering
$\mathbf{g}_f$
Filter generation
$\mathbf{\Phi}_f^{\mathrm{S}}$
$\mathbf{\Phi}_f^{\mathrm{N}}$
$\mathbf{u}$
Mask-based estimation of PSD matrices
Attention-based estimation of reference mic
$\{X_c\}_{c=1}^{C}$
$\hat{X}$
Feature extraction (filterbank)
$\hat{O}$
$P(y_n|\hat{O}, y_{1:n-1})$
Decoder
$\mathbf{s}_{n-1}$
y_{n-1}
$\mathbf{c}_n$
$\mathbf{s}_n$
Attention
$\mathbf{a}_{n-1}$
H
Encoder
$\hat{O}$
Speech enhance part: Mask-based neural beamformer
Feature extraction part
Speech recognition part: Attention-based Encoder-Decoder
Processing flow

Fig. 1. Overview of multichannel end-to-end ASR system architecture: Mask-based neural beamformer works as a speech enhancement module and the attention-based encoder-decoder network works as an ASR module. Feature extraction function bridges these two modules.

2.2. Mask-based neural beamformer

The left side of Fig. 1 outlines our adopted mask-based neural beamformer.

The mask-based neural beamformer technique is based on linear filtering in the time-frequency domain. Let $x_{t,f,c}(\in \mathbb{C})$ and $g_{f,c}(\in \mathbb{C})$ be the STFT coefficients of the c-th channel noisy signal at time-frequency bin (t, f) and its corresponding beamforming filter coefficients, respectively. Then enhanced STFT coefficients $\hat{x}_{t,f}(\in \mathbb{C})$ are obtained as follows:

$$\hat{x}_{t,f} = \mathbf{g}_f^{\dagger}\mathbf{x}_{t,f}, \tag{4}$$

where $\mathbf{x}_{t,f} = \{x_{t,f,c}\}_{c=1}^{C}(\in \mathbb{C}^{C})$ is the spatial vector that represents the signals obtained from all the microphones for each time-frequency bin (t, f), $\mathbf{g}_f = \{g_{f,c}\}_{c=1}^{C}(\in \mathbb{C}^{C})$ is the time-invariant beamforming filter coefficients for all of the C channels, and † represents the conjugate transpose.

Filter coefficients $\mathbf{g}_f$ in Eq. (4) are computed based on the following Minimum Variance Distortionless Response (MVDR) formalization [11]:

$$\mathbf{g}_f = \frac{(\mathbf{\Phi}_f^{\mathrm{N}})^{-1}\mathbf{\Phi}_f^{\mathrm{S}}}{\mathrm{Tr}((\mathbf{\Phi}_f^{\mathrm{N}})^{-1}\mathbf{\Phi}_f^{\mathrm{S}})}\mathbf{u}, \tag{5}$$

where $\mathbf{\Phi}_f^{\mathrm{S}}(\in \mathbb{C}^{C\times C})$ and $\mathbf{\Phi}_f^{\mathrm{N}}(\in \mathbb{C}^{C\times C})$ are the cross-channel power spectral density (PSD) matrices (also known as spatial covariance matrices) for speech and noise signals, respectively, $\mathbf{u}(\in \mathbb{R}^{C})$ is a vector representing a selected reference microphone, and $\mathrm{Tr}(\cdot)$ represents the matrix trace operation.

Let $m_{t,f}^{\mathrm{S}}(\in [0, 1])$ and $m_{t,f}^{\mathrm{N}}(\in [0, 1])$ be the time-frequency masks for the speech and noise signals, respectively. Based on previous work [12], the PSD matrices are robustly estimated as the expected value with respect to the time-frequency masks as follows:

$$\mathbf{\Phi}_f^{\mathrm{S}} = \frac{1}{\sum_{t=1}^{T} m_{t,f}^{\mathrm{S}}} \sum_{t=1}^{T} m_{t,f}^{\mathrm{S}} \mathbf{x}_{t,f} \mathbf{x}_{t,f}^{\dagger}, \tag{6}$$

$$\mathbf{\Phi}_f^{\mathrm{N}} = \frac{1}{\sum_{t=1}^{T} m_{t,f}^{\mathrm{N}}} \sum_{t=1}^{T} m_{t,f}^{\mathrm{N}} \mathbf{x}_{t,f} \mathbf{x}_{t,f}^{\dagger}. \tag{7}$$

The time-frequency masks, $m_{t,f}^{\mathrm{S}}, m_{t,f}^{\mathrm{N}}$, are estimated with Bidirectional Long Short-Term Memory (BLSTM)-based recurrent networks. Reference microphone vector $\mathbf{u}$ is estimated with our proposed attention-based reference estimation mechanism. The details of the estimation procedures are described in our previous study [6]. The entire beamforming procedures described in this subsection corresponds to $\mathrm{Enhance}(\cdot)$ in Eq. (1).

2.3. Attention-based encoder-decoder network

The right-side of Fig. 1 illustrates an overview of the attention-based encoder-decoder network, which consists of two Recurrent Neural Networks (RNNs), one for the encoder and another for the decoder, both of which are connected by an attention mechanism.

Given feature sequence $O = \{\mathbf{o}_t \in \mathbb{R}^{D_{\mathrm{O}}} | t = 1, \cdots, T\}$, where $\mathbf{o}_t$ is a D_{O}-dimensional (LMF) feature vector at input time step t and T is the input sequence length, the network estimates the *a posteriori* probabilities for output label sequence $Y = \{y_n \in \mathcal{V} | n = 1, \cdots, N\}$, where y_n is a label symbol (e.g., character) at output time step n, N is the output sequence length, and $\mathcal{V}$ is a set of labels as follows:

$$P(Y|O) = \prod_n P(y_n|O, y_{1:n-1}), \tag{8}$$

$$H = \mathrm{Encoder}(O), \tag{9}$$

$$\mathbf{c}_n = \mathrm{Attention}(\mathbf{a}_{n-1}, \mathbf{s}_n, H), \tag{10}$$

$$P(y_n|O, y_{1:n-1}) = \mathrm{Decoder}(\mathbf{c}_n, \mathbf{s}_{n-1}, y_{1:n-1}), \tag{11}$$

where $y_{1:n-1}$ is a label sequence that consists of y_1 through y_{n-1}. Eqs. (8)-(11) correspond to $\mathrm{E2E\_ASR}(\cdot)$ in Eq. (3).

For input sequence O, the encoder RNN in Eq. (9) first transforms it to the L-length 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 of the encoder's top layer at subsampled time step l. Next the attention mechanism in Eq. (10) integrates all encoder outputs H into a D_{H}-dimensional context vector $\mathbf{c}_n \in \mathbb{R}^{D_{\mathrm{H}}}$ using L-dimensional attention weight vector $\mathbf{a}_n \in [0, 1]^L$ that represents a soft alignment of the encoder outputs at output time step n. Then the decoder RNN in Eq. (11) updates hidden state $\mathbf{s}_n$, estimates the *a posteriori* probability for output label y_n at output time step n, and further estimates the *a posteriori* probabilities for output sequence Y, based on the RNN recursiveness.

3. EXPERIMENTS

3.1. Outline of analyses

3.1.1. Evaluation in signal-to-distortion ratio and perceptual evaluation of speech quality

To evaluate the speech enhancement quality of the outputs of the beamformer modules, we adopted two criteria: 1) a signal-to-distortion ratio (SDR) [13] and 2) a perceptual evaluation of speech quality (PESQ) [14].

The SDR criterion is a quantitative measure representing the ratio between the target signal components and such distortion components as interference, noise, and artifact errors, which is commonly used in the literature of speech separation and enhancement. On the other hand, the PESQ criterion is a quantitative measure that considers human perception characteristics, which is commonly used as an industry standard for speech quality assessments in telecommunications. In both criteria, a higher score indicates that its corresponding estimated signal obtained higher quality.

The score calculation for these two criteria needs a pair of estimated enhanced speech signals and its corresponding clean speech signals. Therefore, we used the development set of the simulation data in the CHiME-4 corpus for this evaluation. We utilized the BSS EVAL toolbox[1] and Loizou's toolbox[2] for calculating the SDR and PESQ scores, respectively.

3.1.2. Evaluation of word error rates decoded with DNN-HMM hybrid system

In addition to the above evaluation using signal-level criteria, we evaluated the speech enhancement quality by inputting the enhanced speech signals to the conventional DNN-HMM hybrid system. Through the evaluation here, we investigate whether the enhanced speech signal obtained by our M-E2E framework is effective not only for the end-to-end framework but also for the conventional DNN-HMM hybrid framework.

We summarize the evaluation procedures as follows:

1. Extract the beamformer module from our M-E2E-based system, which was already trained with end-to-end ASR objectives.

2. Using the extracted module, obtain an enhanced speech signal for each utterance in the development and evaluation datasets of the CHiME-4 corpus.

3. Using the enhanced signals, compute the WER scores using CHiME-4's official baseline DNN-HMM hybrid system, provided in the CHiME-4 corpus [8].

Although we evaluated WER using the DNN-HMM hybrid system, the above front-end, mask-based neural beamformer was developed within our M-E2E framework.

3.1.3. Effects of increasing data for training neural recognition module

In addition to the above two evaluation scopes, we evaluated the effects of increasing the training data to improve the discriminative power of the attention-based encoder-decoder recognition module in the M-E2E framework. The decoder network in the recognition module plays a language model role in the DNN-HMM hybrid framework. However, in the CHiME-4 setup, the training data for the decoder network might be insufficient to gain language regularity, which is generally obtained using a large amount of text data. The size of CHiME-4 in terms of the amount of text data is obviously smaller than that in standard cases.

In this experiment, we take a practical approach that uses such large-scale clean speech datasets as the WSJ corpus to increase the amount of training data for the decoder network: although the size of the noisy speech dataset is often limited, many large-scale clean

[1] http://bass-db.gforge.inria.fr/bss_eval/bss_eval.zip

[2] http://ecs.utdallas.edu/loizou/speech/composite.zip

Table 1. Network configurations of end-to-end ASR system.

Model	Layer	Units	Type	Activation
Encoder	L1 - L4	320	BLSTM + Projection	tanh
Decoder	L1	320	LSTM	tanh
	L2	48	Linear	softmax
Beamformer	L1 - L3	320	BLSTM + Projection	tanh
	L4	514	Linear	sigmoid

Table 2. Training and decoding conditions for end-to-end ASR system.

Parameter initialization	Uniform distribution ([-0.1, 0.1])
Optimization technique	AdaDelta [15] + gradient clipping [16]
Training objective	Joint CTC-attention loss [17]
Training epoch	15
Beam size [18]	20
Length penalty [2]	0.3

speech datasets are available. Our experiment will clarify the effectiveness of an expedient but realistic way for improving the end-to-end ASR system development in the multichannel noisy ASR tasks.

3.2. Data corpora and representation

We conducted experiments with two corpora: 1) CHiME-4, a multichannel noisy ASR corpus whose training data length is 18 hours, and 2) WSJ, a single-channel clean ASR corpus whose training data length is 81 hours.

CHiME-4 consists of speech data recorded using a tablet device with 6-channel microphones in the following four environments: 1) in a cafe, 2) at a street junction, 3) on public transportation, and 4) in a pedestrian area. The data were grouped in two subsets: real data, which were actually recorded in one of the above environments, and simulated data, which were synthesized by adding recorded environment sounds (noise signals) to clean speech data that enabled enhancement evaluation using the clean speech as a reference. The whole data were also grouped in three subsets: 1) a training set (3 hours for real data and 15 hours for simulation data), 2) a development set (2.9 hours for real data and 2.9 hours for simulation data), and 3) an evaluation set (2.2 hours for real data and 2.2 hours for simulation data). The training set was used to train systems to be evaluated, the development set was used to set the hyper-parameters in training and determined the final status of the trained systems, and the evaluation set was used to test the trained systems independently from the training/development sets.

Among the 6-channel microphone outputs, we used 5-channel outputs ($C = 5$) recorded by the microphones on the front of tablet.

Using the Fourier transform, we converted the signal of each channel to a sequence of 257-dimensional STFT features ($F = 257$) and input the converted sequences for all of the channels to the mask-based neural beamformer.

In the feature extraction stage between the front-end neural beamformer and the back-end attention-based encoder-decoder, we adopted 40-dimensional LFM outputs ($D_{\mathrm{O}} = 40$).

3.3. Enhanced signals and ASR systems for comparison

To evaluate the characteristics of the outputs of the mask-based neural beamformer in our M-E2E-based system, we prepared (for comparison) the five following types of signals: 1) NOISY, 2) BEAMFORMIT, 3) MULTI_END2END, 4) ERDOGAN's_MVDR, and 5) HEYMANN's_GEV. NOISY is the single-channel noisy signal from 'isolated 1ch track' in CHiME-4. BEAMFORMIT is an enhanced signal with BeamformIt [7], which is a well-known weighted delay-and-sum beamformer. MULTI_END2END is a signal enhanced by our own neural beamformer in the M-E2E ASR system. ERDOGAN's_MVDR and HEYMANN's_GEV are signals enhanced by two types of state-of-the-art neural beamformers [4, 5]. ERDOGAN's_MVDR and HEYMANN's_GEV are known for their high ASR performances achieved with the conventional DNN-HMM hybrid system in the recent CHiME-4 challenge. To obtain enhanced signals for HEYMANN's_GEV, we used the software tools provided at their GitHub repository[3].

There were several differences in the beamforming formalization and network configurations among the three neural beamformers, especially beamformers that generate the ERDOGAN's_MVDR and HEYMANN's_GEV signals used the parallel data of clean and noisy speech for training; our beamformer in MULTI_END2END only used noisy data and their transcripts. More specifically, the mask estimation networks in ERDOGAN's_MVDR and HEYMANN's_GEV were optimized to estimate the ideal binary masks that are defined with parallel data of clean and noisy speech, while the mask estimation network in MULTI_END2END was optimized under the end-to-end ASR criterion using pairs of noisy speech and transcribed labels. Because we pursued a pure end-to-end setup, we did not use the parallel data of clean and noisy speech to train our M-E2E ASR system. The details of the neural beamformers for ERDOGANs MVDR and HEYMANNs GEV were described in the literature [4, 5], respectively.

For comparison purposes, we also prepared two kinds of ASR systems: a purely neural-network-based end-to-end ASR system and a conventional DNN-HMM hybrid system. We summarize the fundamental specifications and conditions of the M-E2E-based system in Tables 1 and 2. The M-E2E-based system was trained with the conventional multi-condition training strategy [8]. Note that the network configurations described in Table 1 were used for generating the signals in MULTI_END2END. On the other hand, we adopted the official baseline DNN-HMM hybrid ASR system that was included in the ChiME-4 corpus. The system was optimized using sequence-discriminative training with 5-th channel noisy speech data and it applied the language model re-scoring technique. The detail descriptions of the system are shown in a Kaldi recipe[4].

3.4. Results

3.4.1. Evaluation in signal-to-distortion ratio and perceptual evaluation of speech quality

To evaluate the speech enhancement quality obtained with the M-E2E-based system, we summarize the SDR and PESQ scores for the simulation data in the CHiME-4 development set in Figs. 2 and 3, respectively. Because the SDR and PESQ scores are computed for each utterance, we overlaid the standard deviation value (thin line) on the mean value (blue bar) in the figures. The results in both of the scores show that our M-E2E development framework achieved reasonable speech enhancement and its enhancement quality is competitive to or better than the cases of BEAMFORMIT and HEYMANN's_GEV, suggesting that the neural beamformer developed in our M-E2E framework successfully learns speech enhancement (noise suppression) ability, although it was optimized under the

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

[4] `https://github.com/kaldi-asr/kaldi/tree/master/egs/chime4/s5_6ch` Kaldi is a popular open-source toolkit for conducting ASR experiments [19].

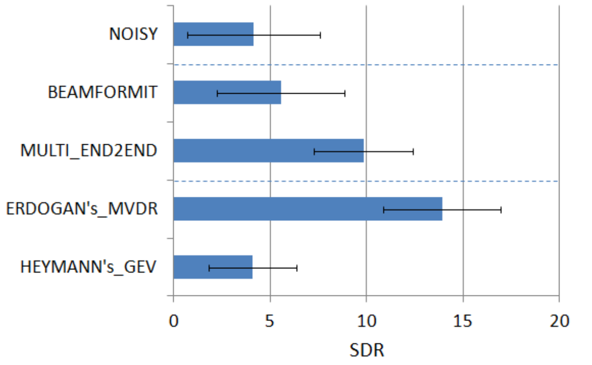

Fig. 2. Signal-to-distortion ratio (SDR) for simulation data in development set. Each bar indicates the mean in terms of utterances; a thin line indicates the standard deviation.

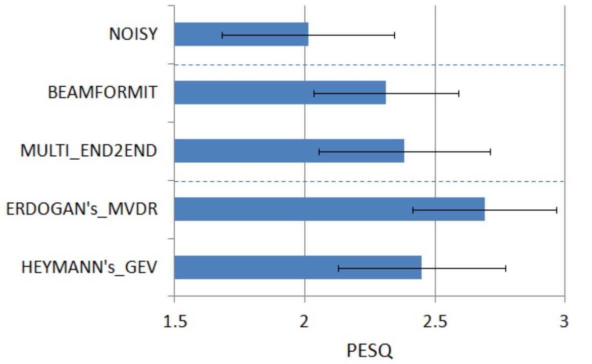

Fig. 3. Perceptual Evaluation of Speech Quality (PESQ) for simulation data in development set.

end-to-end ASR-oriented criterion.

3.4.2. Evaluation with DNN-HMM hybrid framework

To compare the discriminative power of the M-E2E-based enhanced speech signal itself, we performed ASR experiments by applying the baseline DNN-HMM hybrid system to the five prepared signals, NOISY through HEYMANN's_GEV, in the CHiME-4 corpus. Table 3 shows the WERs for the five cases. "Input signal" denotes the type of signal input to the DNN-HMM hybrid system, "Parallel use" indicates whether clean speech was used to train the beamformer, and "Dev-simu," "Dev-real," "Eval-simu," and "Eval-real" denote the WER scores of the development sets of the simulated and real speech data, the evaluation sets of simulated and real speech data, respectively.

The results show that the input signal enhanced by our M-E2E framework (in MULTI_END2END) achieved lower WER values than the signal enhanced by the BeamformIt enhancement (in BEAMFORMIT) as well as the original noisy speech (in NOISY). These results suggest that the neural beamformer in our M-E2E framework produced more suitable enhanced speech inputs at least for the DNN-HMM hybrid system than the standard beamformer, BeamformIt. On the other hand, the results of the M-E2E framework remain lower than those of the state-of-the-art neural beamformers in ERDOGAN's_MVDR and HEYMANN's_GEV. The main reason is probably that both competitors utilize clean and noisy speech data in parallel for speech enhancement. Another possible reason is that the competitors adopted such sophisticated tuning techniques as dropout [20] and batch normalization [21] in their beamformer training, while the M-E2E-based training for its beamformer did not use any of them.

Table 3. Word error rates [%] obtained with DNN-HMM hybrid system.

Input signal	Parallel use	Dev-simu	Dev-real	Eval-simu	Eval-real
NOISY	N/A	13.0	11.6	20.8	23.7
BEAMFORMIT	No	6.8	5.8	10.9	11.5
MULTI_END2END	No	5.2	5.5	7.1	10.1
ERDOGAN's_MVDR	Yes	4.0	4.5	4.1	8.0
HEYMANN's_GEV	Yes	5.3	5.0	6.7	7.3

3.4.3. Effects of increasing training data for end-to-end ASR framework

To analyze the effects of increasing the amount of training data, we compared the five cases, i.e., NOISY through HEYMANN's_GEV, in terms of CER, and summarized the CER scores obtained just using the CHiME-4 data and those obtained with the CHiME-4 data plus the WSJ clean speech data in Tables 4 and 5, respectively. In the testing stage for this purpose, we input one of the signals in the five cases (e.g., the original noisy speech or the signals enhanced by the M-E2E framework) to the attention-based encoder-decoder ASR system, which is not the DNN-HMM hybrid but it has a fully neural architecture. In addition, we only re-trained the back-end, attention-based encoder-decoder recognition module but did not re-train the front-end, mask-based neural beamformer in our M-E2E development framework when adding the WSJ speech data: The added speech data are single-channel clean data and therefore we cannot train the mask-based beamformer in principle.

A comparison of the two tables clearly shows that using additional data improved the recognition performances in all of the cases, even though the additional data were single-channel clean speech signals. From the results, we also found that our M-E2E-based ASR systems outperformed the E2E-BIt-based ASR systems, i.e., the combination of the neural recognition module and the BeamformIt, in all of the conditions in the tables. This clearly demonstrates the rationality and effectiveness of our M-E2E development framework.

Surprisingly, the accuracies in HEYMANN's_GEV were very low. A possible reason of this was that the signals produced by the generalized eigenvalue (GEV)-based beamformer produced a mismatch with the back-end neural recognition module because of the additional speech distortions caused by the GEV procedure.

4. CONCLUSION

We conducted the following two experimental analyses to investigate whether the speech enhancement component in our M-E2E development framework really learned a speech enhancement (noise suppression) capability: 1) evaluation of the speech enhancement quality in terms of two signal-level measures, i.e., SDR and PESQ, 2) evaluation of the speech enhancement performance in terms of the ASR-level measure, i.e., WER, in recognition experiments using the DNN-HMM hybrid system. The experimental results showed that our proposed M-E2E framework successfully achieved a speech enhancement (noise suppression) capability, although it was optimized based on the end-to-end ASR-oriented objective for generating correct label sequences.

In addition to the above experiment, we evaluated the effects of increasing the training data for back-end, encoder-decoder net-

Table 4. Character error rates [%] obtained with end-to-end ASR system over CHiME-4 corpus.

Input signal	Parallel use	Dev simu	Dev real	Eval simu	Eval real
NOISY	×	25.0	24.5	34.7	35.8
BEAMFORMIT	No	21.5	19.3	31.2	28.2
MULTI_END2END	No	15.3	18.2	23.7	26.8
ERDOGAN's_MVDR	Yes	16.2	18.2	24.3	26.7
HEYMANN's_GEV	Yes	61.5	65.3	56.8	65.5

Table 5. Character error rates [%] obtained with end-to-end ASR system over CHiME-4 and WSJ corpora.

Input signal	Parallel use	Dev simu	Dev real	Eval simu	Eval real
NOISY	×	19.4	19.0	28.7	29.6
BEAMFORMIT	No	17.4	15.3	28.1	24.2
MULTI_END2END	No	11.8	13.7	20.5	21.5
ERDOGAN's_MVDR	Yes	10.5	11.9	18.0	19.2
HEYMANN's_GEV	Yes	35.3	35.7	37.4	42.8

works without increasing the training data for the front-end, neural beamformer. The experimental result showed that even increasing the single-channel clean training data is effective to improve the recognition performance of the M-E2E development framework in the multichannel noisy ASR tasks.

The above findings also suggest that the M-E2E development framework probably already achieved reasonable beamformers, and such back-end recognition modules as attention-based encoder-decoder networks must be further improved to boost the discriminative power of the total end-to-end ASR system. To meet this requirement, developing training algorithms to optimize decoder networks only with text data (without speech signal data) is an important future subject.

In this work, we selected the multichannel speech enhancement front-end for our end-to-end ASR framework, because the performance of multichannel speech enhancement was shown to be basically better than that of the single-channel technique [22]. However, when considering such resource-constrained devices as mobile phones, we also find the importance of studying the single-channel enhancement front-end in the end-to-end ASR context.

5. REFERENCES

[1] Geoffrey Hinton, Li Deng, Dong Yu, et al., "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[2] Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio, "Attention-based models for speech recognition," in *Proc. NIPS*, 2015, pp. 577–585.

[3] Alex Graves and Navdeep Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *Proc. ICML*, 2014, pp. 1764–1772.

[4] Hakan Erdogan, Tomoki Hayashi, John R Hershey, et al., "Multi-channel speech recognition: LSTMs all the way through," in *CHiME 2016 workshop*, 2016.

[5] Jahn Heymann, Lukas Drude, and Reinhold Haeb-Umbach, "Wide residual blstm network with discriminative speaker adaptation for robust speech recognition," in *CHiME 2016 workshop*, 2016.

[6] Tsubasa Ochiai, Shinji Watanabe, Takaaki Hori, and John R Hershey, "Multichannel end-to-end speech recognition," in *International Conference on Machine Learing (ICML)*, 2017.

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

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

[9] Thomas Hain, Lukas Burget, John Dines, et al., "The AMI system for the transcription of speech in meetings," in *Proc. ICASSP*, 2007, pp. 357–360.

[10] Douglas B Paul and Janet M Baker, "The design for the wall street journal-based csr corpus," in *Proc. workshop on Speech and Natural Language*, 1992, pp. 357–362.

[11] Mehrez Souden, Jacob Benesty, and Sofiène Affes, "On optimal frequency-domain multichannel linear filtering for noise reduction," *IEEE Transactions on audio, speech, and language processing*, vol. 18, no. 2, pp. 260–276, 2010.

[12] Takuya Yoshioka, Nobutaka Ito, Marc Delcroix, et al., "The NTT CHiME-3 system: Advances in speech enhancement and recognition for mobile multi-microphone devices," in *Proc. ASRU*, 2015, pp. 436–443.

[13] Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[14] Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. ICASSP*, 2001, vol. 2, pp. 749–752.

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

[16] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio, "On the difficulty of training recurrent neural networks," *Proc. ICML*, pp. 1310–1318, 2013.

[17] Suyoun Kim, Takaaki Hori, and Shinji Watanabe, "Joint CTC-attention based end-to-end speech recognition using multi-task learning," in *Proc. ICASSP*, 2017, pp. 4835–4839.

[18] Ilya Sutskever, Oriol Vinyals, and Quoc V Le, "Sequence to sequence learning with neural networks," in *Proc. NIPS*, 2014, pp. 3104–3112.

[19] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, et al., "The kaldi speech recognition toolkit," in *IEEE workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2011.

[20] Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting.," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[21] Sergey Ioffe and Christian Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.