SPEAKER ADAPTATION FOR MULTICHANNEL END-TO-END SPEECH RECOGNITION

Tsubasa Ochiai [1], *Shinji Watanabe* [2†], *Shigeru Katagiri* [1], *Takaaki Hori* [2], *John Hershey* [2]

[1] Graduate School of Science and Engineering, Doshisha University, Kyoto, Japan
[2] Mitsubishi Electric Research Laboratories, Massachusetts, USA

ABSTRACT

Recent work on multichannel end-to-end automatic speech recognition (ASR) has shown that multichannel speech enhancement and speech recognition functions can be integrated into a deep neural network (DNN)-based system, and promising experimental results have been shown using the CHiME-4 and AMI corpora. In other recent DNN-based hidden Markov model (DNN-HMM) hybrid architectures, the effectiveness of speaker adaptation has been well established. Motivated by these results, we propose a multi-path adaptation scheme for end-to-end multichannel ASR, which combines the unprocessed noisy speech features with a speech-enhanced pathway to improve upon previous end-to-end ASR approaches. Experimental results using CHiME-4 show that (1) our proposed multi-path adaptation scheme improves ASR performance and (2) adapting the encoder network is more effective than adapting the neural beamformer, attention mechanism, or decoder network.

Index Terms— multichannel end-to-end ASR, neural beamformer, attention-based encoder-decoder, speaker adaptation

1. INTRODUCTION

Over the last decade, with the advent of deep learning techniques, deep neural network (DNN)-hidden Markov model (HMM) hybrid architectures [1] have become a standard approach for automatic speech recognition (ASR). In parallel with this, there has been significant interest in developing fully end-to-end deep learning architectures, such as attention-based encoder-decoder networks [2, 3] and connectionist temporal classification (CTC) systems [4]. The benefits of such approaches include 1) structural simplicity (the entire procedure from input to output is learned from data in a monolithic neural network-based architecture) and 2) consistency in optimization (the entire system is optimized with a single ASR-level objective).

Previous studies on end-to-end architectures mainly focused on the development of ASR systems in a single-channel setup without speech enhancement. However, in more realistic scenarios, speech inputs to ASR systems are contaminated by background noise and reverberation. Therefore, it is clearly important to study the utility of the end-to-end architecture in a multichannel setup where multichannel speech enhancement is conducted. In light of this, we extended the attention-based encoder-decoder framework by incorporating multichannel speech enhancement components, and proposed a Multichannel End-to-End (ME2E) ASR architecture that directly converts multichannel speech signal to text [5, 6]. We also showed that the proposed ME2E architecture successfully learned speech enhancement (noise suppression) ability through the end-to-end optimization procedure and achieved higher recognition performance than the conventional single-channel end-to-end architecture [5, 6].

Recent studies using benchmark tasks (e.g. CHiME 3 and 4 challenges) [7, 8] provided several clues to performance improvement in difficult noisy ASR problems, e.g. 1) use of such multichannel signal processing techniques as the beamforming method, 2) use of a strong language model such as the LSTM-based RNN language model, and 3) use of speaker adaptation techniques. Taking into account these hints and another recent result (e.g. [9, 10]) in which speaker adaptation techniques were effectively applied to the DNN-HMM hybrid architecture, we naturally reach the expectation that speaker adaptation techniques can further improve the performance of the ME2E ASR architecture.

Motivated by the above, in this paper, we propose a *multi-pass adaptation* scheme for the ME2E ASR architecture where input speech data are transmitted through an unprocessed noisy speech path and an speech-enhanced path, and evaluate its effectiveness compared to a *single-path adaptation* scheme where input speech data are transmitted only through a speech-enhanced path. In addition, we analyze the effect of speaker adaptation in the ME2E ASR architecture, directing attention to the following questions:

1. Which is the most effective for speaker adaptation among ME2E ASR system components: neural beamformer, encoder network, attention mechanism, or decoder network?
2. How does the speaker adaptation procedure affect the inner behavior of the ME2E ASR system?

Our study in this paper is related to previous works on speaker adaptation for a conventional HMM-based ASR architecture (e.g. [9, 10, 11, 12]), except that we focus on speaker adaptation in the ME2E ASR framework. To the best of our knowledge, this paper shows the first results of the speaker adaptation for either the end-to-end ASR architecture or the neural beamformer.

2. OVERVIEW OF MULTICHANNEL END-TO-END ASR

In Figure 1, we illustrate an overview of the ME2E ASR architecture. The architecture adopts a mask-based neural beamformer [13, 14] as a speech enhancement component and the attention-based encoder-decoder [2, 3] as an ASR component, where the feature extraction function connects these components. Let $X^c = \{\mathbf{x}_t^c \in \mathbb{C}^F | t = 1, \cdots, T\}$ be a short-time Fourier transform (STFT) feature sequence recorded at c-th channel, where $\mathbf{x}_t^c$ is a F-dimensional STFT feature vector at input time step t, T is the input sequence length, and C is the number of channels. Given multichannel noisy speech inputs $\{X^c\}_{c=1}^C$, the ME2E ASR architecture directly estimates the *a posteriori* probabilities for output label sequence $Y = \{y_n \in \mathcal{V} | n = 1, \cdots, N\}$ using the fully neural network-based architecture, 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

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, Takaaki Hori, and John Hershey was supported by MERL.

† Currently, he is at Johns Hopkins University.

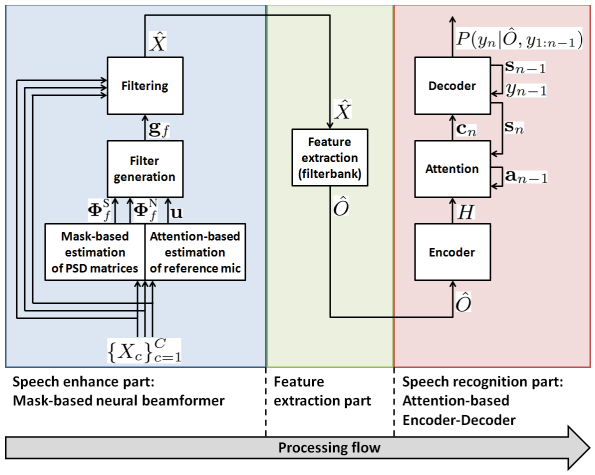

Fig. 1. Overview of multichannel end-to-end ASR architecture.

follows:

$$P(Y|\{X^c\}_{c=1}^{C};\Lambda_{\text{all}}) = \prod_n P(y_n|\{X^c\}_{c=1}^{C}, y_{1:n-1};\Lambda_{\text{all}}), \quad (1)$$

$$\hat{X} = \text{Beamformer}(\{X^c\}_{c=1}^{C};\Lambda_{\text{beam}}), \quad (2)$$

$$\hat{O} = \text{Feature}(\hat{X}), \quad (3)$$

$$H = \text{Encoder}(\hat{O};\Lambda_{\text{enc}}), \quad (4)$$

$$\mathbf{c}_n = \text{Attention}(\mathbf{a}_{n-1}, \mathbf{s}_n, H;\Lambda_{\text{att}}), \quad (5)$$

$$P(y_n|\{X^c\}_{c=1}^{C}, y_{1:n-1};\Lambda_{\text{all}}) = \text{Decoder}(\mathbf{c}_n, \mathbf{s}_{n-1}, y_{1:n-1};\Lambda_{\text{dec}}), \quad (6)$$

where $\Lambda_{\text{all}} = \{\Lambda_{\text{beam}}, \Lambda_{\text{enc}}, \Lambda_{\text{att}}, \Lambda_{\text{dec}}\}$ represents a total set of trainable model parameters. Λ_{beam}, Λ_{enc}, Λ_{att}, and Λ_{dec} correspond to the model parameters for each function.

First, $\text{Beamformer}(\cdot)$ estimates the beamforming filter $\mathbf{g}_f$ through the estimation of three statistics, (1) the cross-channel power spectral density matrix for speech $\mathbf{\Phi}_f^{\mathrm{S}}$, (2) the same type matrix for noise $\mathbf{\Phi}_f^{\mathrm{N}}$, and (3) the reference microphone vector $\mathbf{u}$, and it also integrates the multichannel noisy speech signals $\{X^c\}_{c=1}^{C}$ into a single-channel enhanced speech signal $\hat{X}$ by linear filtering. Next, $\text{Feature}(\cdot)$ converts the enhanced STFT feature sequence $\hat{X}$ to a log Mel filterbank (LMF) feature sequence $\hat{O} = \{\hat{\mathbf{o}}_t \in \mathbb{R}^{D_{\mathrm{O}}} | t = 1, \cdots, T\}$, where $\hat{\mathbf{o}}_t$ is a D_{O}-dimensional LMF feature vector at input time step t. Moreover, $\text{Encoder}(\cdot)$ transforms the enhanced LMF feature sequence $\hat{O}$ 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. $\text{Attention}(\cdot)$ 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. Finally, $\text{Decoder}(\cdot)$ 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. ADAPTATION OF MULTICHANNEL END-TO-END ASR

3.1. Basic formalization of speaker adaptation procedure

In this paper, we focus on an unsupervised speaker adaptation scenario, where hypothesized transcriptions are generated by the first-pass decoding with the speaker-independent (SI) end-to-end system and used as target labels in place of (correct) reference transcriptions. We also adopt a simple retraining-based adaptation scheme, where the network parameters of either all or some of the system components are re-estimated using a target speaker's speech data.

Let Λ_{adapt} be parameters to be adapted (adaptation parameters) in a speaker adaptation stage (e.g. $\Lambda_{\text{adapt}} = \Lambda_{\text{enc}}$). We also assume that training samples spoken by target speaker s, $\mathcal{X}_s = \{(\mathbf{X}_i, Y_i) | i = 1, \cdots, I\}$, are available for adaptation, where $\mathbf{X}_i$ is the i-th multichannel noisy speech sample, Y_i is its corresponding target label, and I is the number of such samples. Then, using adaptation objective $E(\mathbf{\Lambda}_{\text{all}}; \mathcal{X}_s)$, an optimization procedure for a speaker-adapted end-to-end system is formalized as follows:

$$\overline{\mathbf{\Lambda}}_{\text{adapt}} = \arg\min_{\mathbf{\Lambda}_{\text{adapt}}} E(\mathbf{\Lambda}_{\text{all}}; \mathcal{X}_s). \quad (7)$$

Considering the risk of overtraining and/or storage cost, the adaptation parameters should not be too large. In addition, it is also an interesting question which component in the ME2E system is more effective for speaker adaptation (in other words, more related to speaker characteristics). We experimentally investigate this point by changing the adaptation parameter selection, such as $\Lambda_{\text{adapt}} = \Lambda_{\text{beam}}$ or $\Lambda_{\text{adapt}} = \Lambda_{\text{enc}}$.

3.2. Multi-path adaptation for multichannel end-to-end ASR

Although the ME2E ASR architecture is fully based on neural network, it consists of such separately designed components as the speech enhancement component and the feature extraction component. Based on this, we can consider a multi-path adaptation scheme in the ME2E architecture by adopting the multi-condition training concept [8]. In Figure 2, we give an overview of the multi-path adaptation procedure. In addition to a speech-enhanced path that goes through the neural beamformer, i.e. the speech enhancement component, we set an unprocessed noisy speech path that goes directly from the input to the latter component, the attention-based encoder-decoder network. The resulting procedure makes it possible to optimize the entire network not only with the signal enhanced by the neural beamformer but with the unprocessed noisy signal. It is therefore expected that the attention-based encoder-decoder network learns the robustness against noisy speech and becomes a powerful ASR back-end component.

Let $\mathcal{L}^{\text{enhan}}(Y|\mathbf{X})$ be the joint CTC-attention loss [15] through the speech-enhanced path and $\mathcal{L}^{\text{noisy}}(Y|X^c)$ be the loss through the unprocessed noisy path. Then, the optimization of the multi-path adaptation procedure is formalized as follows:

$$\overline{\mathbf{\Lambda}}_{\text{adapt}} = \arg\min_{\mathbf{\Lambda}_{\text{adapt}}} \Big(E^{\text{enhan}}(\mathbf{\Lambda}_{\text{all}}; \mathcal{X}_s) + E^{\text{noisy}}(\mathbf{\Lambda}_{\text{all}}; \mathcal{X}_s) \Big), \quad (8)$$

where E^{enhan} and E^{noisy} are the accumulated loss defined as:

$$E^{\text{enhan}}(\mathbf{\Lambda}_{\text{all}}; \mathcal{X}_s) = \sum_{i=1}^{I} \mathcal{L}^{\text{enhan}}(Y_i|\mathbf{X}_i), \quad (9)$$

$$E^{\text{noisy}}(\mathbf{\Lambda}_{\text{all}}; \mathcal{X}_s) = \sum_{i=1}^{I} \sum_{c=1}^{C} \mathcal{L}^{\text{noise}}(Y_i|X_i^c), \quad (10)$$

where X_i^c is the i-th single-channel noisy speech sample recorded at c-th channel. In the successive experiment section, we compare the multi-path adaptation procedure and the single-path adaptation procedure that uses only the speech-enhanced path for optimization.

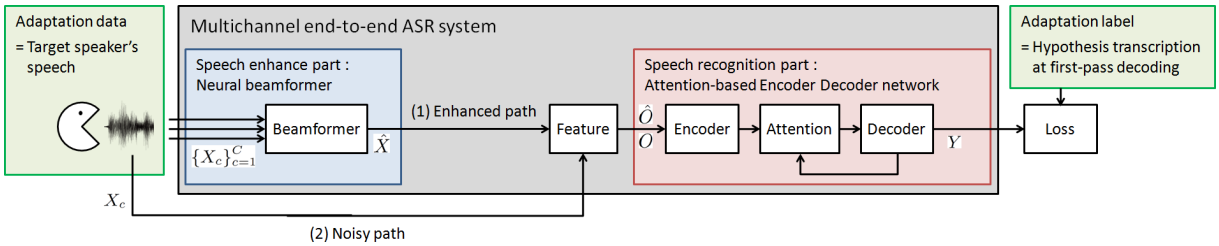

Fig. 2. Overview of multi-path adaptation scheme for multichannel end-to-end ASR architecture.

4. EXPERIMENT

4.1. Condition

We evaluated the speaker adaptation effect for the ME2E ASR architecture mainly using the CHiME-4 corpus, which is a well-known multichannel noisy ASR benchmark. The CHiME-4 corpus consists of real and simulated speech data recorded using a tablet device with 6-channel microphones in four environments: cafe (CAF), street junction (STR), public transportation (BUS), and pedestrian area (PED). The data are also grouped in three subsets: 1) training set, 2) development set, and 3) evaluation set. We used the training set to train the baseline SI ME2E system, which also worked as the initial seed model for the latter speaker adaptation procedure. We also used the development set to optimize such hyperparameters as the number of training epochs in the adaptation, the evaluation set to adapt the adaptation model parameters. We then evaluated the performances of the speaker-adapted ASR systems.

The experimental conditions were basically the same as our previous studies [5, 6]. Main differences of the conditions in the present paper were that we used an additional corpus for training the baseline SI ME2E system and an external language model. Our previous study [16] suggested that the amount of training data in the CHiME-4 corpus was not sufficient to learn the strong ASR back-end and especially to learn the appropriate language regularity. To compensate for such data insufficiency, we used in the paper the WSJ corpus [17] as additional training data and applied the LSTM-based RNN language model to the decoding procedure. We utilized the WSJ's single-channel clean data for training of the encoder-decoder network by bypassing the beamforming network. Because the external, LSTM-based language model was trained using a large amount of the WSJ text data, it provided more appropriate language regularity for the decoding procedure. Our adopted decoding procedure basically follows the one proposed in [18]. Other experimental conditions, such as 1) conditions related to feature extraction, 2) network configurations, and 3) conditions related to the training of the baseline SI ME2E system, were basically the same as those in [5, 6].

To evaluate the speaker adaptation effect in the ME2E ASR framework, we conducted the following three-step procedure that used the unsupervised adaptation setup: 1) generate hypothesized transcriptions by the first-pass decoding with the baseline SI ME2E system, 2) adapt (re-estimate) the network parameters of all or some of the system components, using a target speaker's speech data, and 3) again conduct decoding with the speaker adapted system. Note that, in the multi-path adaptation procedure, we shared the hypothesized transcriptions obtained by the decoding through the enhanced path as the target labels for the corresponding noisy-path optimization. The results were basically represented as word error rate (WER) for the real data of the evaluation set.

For the optimization in the adaptation stage, we used the stochastic gradient descent (SGD) algorithm, which enabled us to fine-tune the network parameters with a small learning rate, with the early stopping technique [19]. We set the learning rate as 0.005 based on preliminary experiments. We repeated 20 training epochs in the adaptation procedure, while calculating the WER scores every 5 epochs. Based on the WER scores for the real data of the development set, we set the number of epochs (i.e. 5, 10, 15, or 20) for the adaptation using the evaluation set. Regardless of the number of epochs, ASR performances were improved by the adaptation procedure. However, selecting appropriate number of adaptation epochs was important to suppress over-fitting and achieve better recognition performances.

Table 1. Word error rate [%] and character error rate [%] of baseline speaker-independent system for real data of evaluation set.

use external language model	WER	CER
No	41.7	20.9
Yes	28.8	17.2

4.2. Result

4.2.1. Baseline

First, we evaluated the effect of the external language model. In Table 1, we show the WER scores of the baseline SI ME2E system for the real data of the evaluation set. For reference, we also show the character error rate (CER) in the table. From the table, we find that the external language model was quite effective at improving the baseline performance, which suggests that the amount of training data in the CHiME-4 corpus was insufficient to learn the appropriate language regularity. Based on this, in successive experiments, we adopted the external language model in the decoding procedure of all of the evaluated systems.

4.2.2. Speaker adaptation

In Table 2, we show the WER scores of the speaker-adapted systems for the real data of the evaluation set. The upper row corresponds to the adaptation scheme. The left-end column shows the set of network parameters, which were adapted (re-estimated) as the adaptation model parameters. In this experiment, we investigated five different assignments of the adaptation model parameters: 1) whole network ($\Lambda_{\text{adapt}} = \Lambda_{\text{all}}$), 2) neural beamformer ($\Lambda_{\text{adapt}} = \Lambda_{\text{beam}}$), 3) encoder network ($\Lambda_{\text{adapt}} = \Lambda_{\text{enc}}$), 4) attention mechanism and decoder network ($\Lambda_{\text{adapt}} = \{\Lambda_{\text{att}}, \Lambda_{\text{dec}}\}$)[1], and 5) neural beamformer and encoder network ($\Lambda_{\text{adapt}} = \{\Lambda_{\text{beam}}, \Lambda_{\text{enc}}\}$).

From the table, we find that, compared to the non-adapted baseline system, the speaker-adapted systems basically achieved higher recognition performances, and that the speaker adaptation was effective even for the ME2E ASR architecture. Also, the comparison of the single-path adaptation and the multi-path adaptation validated

[1]To estimate the *a posteriori* probabilities for output label sequence, the decoder network works cooperatively with the attention mechanism. Hence, we treated them as a set of the adaptation model parameters.

Table 2. Word error rate [%] of speaker-adapted system for real data of evaluation set.

Λ_{adapt}	single-path	multi-path
$\{\Lambda_{\text{all}}\}$	27.1	26.4
$\{\Lambda_{\text{beam}}\}$	28.0	N/A
$\{\Lambda_{\text{enc}}\}$	27.1	26.2
$\{\Lambda_{\text{att}}, \Lambda_{\text{dec}}\}$	28.7	28.9
$\{\Lambda_{\text{beam}}, \Lambda_{\text{enc}}\}$	27.2	26.3

Table 3. Word error rate [%] of environment-adapted system for real data of evaluation set.

Λ_{adapt}	scheme	WER
$\{\Lambda_{\text{beam}}\}$	single-path	28.2
$\{\Lambda_{\text{beam}}, \Lambda_{\text{enc}}\}$	multi-path	27.4

the effectiveness of the multi-path adaptation scheme, which augmented the variability of the adaptation data using both the speech-enhanced and unprocessed noisy paths.

Furthermore, from the table, we can obtain the following findings with respect to the adaptation model parameters:

1. The adaptation of the decoder network did not contribute to improving the ASR performance. This is probably because the decoder network mainly handles language-level features, which are not closely related to speaker characteristics.
2. The adaptation of the encoder network achieved the lowest WER score, which is competitive or a little bit lower than that of the adaptation of the entire network. This result suggests that the adaptation of only the encoder network could be a sufficient replacement of the adaptation of the entire network. This is probably because the encoder network mainly handles the acoustic-signal-level features that contain speaker characteristics.
3. Although the adaptation of the neural beamformer contributed to improving the ASR performance, the improvement was smaller than that of the adaptation of the encoder network. The result suggests that the beamforming procedure was not directly related to the speaker characteristics, or that the beamforming filter estimated by the neural beamformer already possessed speaker variability to some extent.
4. Contrary to expectations, the adaptation effects of the neural beamformer and the encoder network were not complementary. The adaptation of both the neural beamformer and the encoder network did not contribute to further improving the ASR performance, compared to the adaptation of either of them alone.

4.2.3. Environment-level adaptation

There is the possibility that the neural beamformers are more effective for environment adaptation than for speaker adaptation because noise characteristics basically depend not on speakers but on speaking environment. Taking this into account, we additionally investigated the environment-adaptation effect.

In Table 3, we show the WER scores of the environment-adapted systems, which are related to the adaptation of the neural beamformers, for the real data of the evaluation set. From the table, we find that the performance of the environment-adapted systems was lower than that of the speaker-adapted systems. The results suggested that even the environment-level adaptation was not very effective for the adaptation of the neural beamformers and demonstrated that the speaker-level adaptation was more effective at improving the total ASR performance for the ME2E system.

Table 4. Signal-to-distortion ratio and perceptual evaluation of speech quality for simulation data of development set.

Λ_{adapt}	scheme	WER	SDR	PESQ
{ } (baseline)	N/A	14.7	9.10	2.33
$\{\Lambda_{\text{beam}}\}$	single-path	14.1	9.14	2.34
$\{\Lambda_{\text{beam}}, \Lambda_{\text{enc}}\}$	multi-path	12.5	9.13	2.34

4.2.4. Evaluation in signal-level measures

In addition to the above ASR-level evaluations, we conducted additional signal-level evaluation to analyze how the speaker adaptation procedure affected the behavior of the ME2E ASR architecture. To do this, we adopted two signal-level measures: 1) signal-to-distortion ratio (SDR) [20], and 2) perceptual evaluation of speech quality (PESQ) [21], which are commonly used for speech enhancement quality assessment. Then, using these measures, we evaluated the speech enhancement quality of the outputs of the beamforming component. The calculation of these measures requires a pair of estimated enhanced speech signals and its corresponding clean speech signals. Therefore, we used the simulation data of the development set for this evaluation.

In Table 4, we show WER, SDR, and PESQ scores for the baseline and the speaker-adapted systems, which are related to the adaptation of the neural beamformers. The first row indicates the result of the non-adapted baseline system. From the table, we clearly find that the speaker adaptation effect in terms of ASR-level measures (i.e. WER) appears for the simulation data of the development set. However, there is little difference in the signal-level measures (i.e. SDR and PESQ) between the speaker-adapted systems and the baseline system. This result suggests that the effect of the adaptation procedure was small for the behavior of the beamforming component, and it also validated the findings in Section 4.2.2 that the adaptation of the acoustic feature transformation was the most effective for the adaptation of the total ME2E ASR architecture.

5. CONCLUSION

In this paper, we conducted experimental evaluations of the speaker adaptation effect in the ME2E ASR architecture. The experimental results using the CHiME-4 corpus clearly demonstrated the effectiveness of the speaker adaptation in the ME2E ASR architecture. In addition, the experimental results led to the following findings: 1) our proposed multi-path adaptation procedure, which utilizes both the speech-enhanced and noisy paths, is effective at increasing adaptation performance, and 2) among the system components of the ME2E ASR architecture (the neural beamformer, the encoder network, the attention mechanism, and the decoder network), the adaptation of the encoder network was the most effective at adapting the target speaker's characteristics.

Furthermore, the experimental results showed that the adaptation procedure had little effect on the behavior of the beamforming component within the ME2E ASR system. This is probably due to the tight constraint that our adopted mask-based neural beamformer used beamforming filters based on the less flexible, minimum variance distortionless response (MVDR) beamformer [22]. The constraint would degrade the adaptability of the mask-based neural beamformer, and the adoption of other more flexible beamformers [23] will be an interesting future research topic.

6. REFERENCES

[1] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, 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 Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio, "Attention-based models for speech recognition," in *Advances in Neural Information Processing Systems (NIPS)*, 2015, pp. 577–585.

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

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

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

[6] Tsubasa Ochiai, Shinji Watanabe, Takaaki Hori, John R Hershey, and Xiong Xiao, "A unified architecture for multichannel end-to-end speech recognition with neural beamforming," *IEEE Journal of Selected Topics in Signal Processing*, 2017.

[7] Jon Barker, Ricard Marxer, Emmanuel Vincent, and Shinji Watanabe, "The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015, pp. 504–511.

[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] Hank Liao, "Speaker adaptation of context dependent deep neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7947–7951.

[10] G. Saon, H. Soltau, D. Nahamoo, and M. Picheny, "Speaker adaptation of neural network acoustic models using i-vectors," in *2013 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2013, pp. 55–59.

[11] Joao Neto, Luís Almeida, Mike Hochberg, Ciro Martins, Luis Nunes, Steve Renals, and Tony Robinson, "Speaker-adaptation for hybrid hmm-ann continuous speech recognition system," 1995.

[12] Roberto Gemello, Franco Mana, Stefano Scanzio, Pietro Laface, and Renato De Mori, "Linear hidden transformations for adaptation of hybrid ann/hmm models," *Speech Communication*, vol. 49, no. 10, pp. 827–835, 2007.

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

[14] Hakan Erdogan, Tomoki Hayashi, John R Hershey, Takaaki Hori, Chiori Hori, Wei-Ning Hsu, Suyoun Kim, Jonathan Le Roux, Zhong Meng, and Shinji Watanabe, "Multi-channel speech recognition: LSTMs all the way through," in *CHiME 2016 workshop*, 2016.

[15] 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.

[16] Tsubasa Ochiai, Shinji Watanabe, and Shigeru Katagiri, "Does speech enhancement work with end-to-end ASR objectives?: experimental analysis of multichannel end-to-end ASR," in *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2017.

[17] 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.

[18] Takaaki Hori, Shinji Watanabe, Yu Zhang, and William Chan, "Advances in joint CTC-attention based end-to-end speech recognition with a deep CNN encoder and RNN-LM," in *Interspeech*, 2017, pp. 949–953.

[19] Christopher M Bishop, *Pattern recognition and machine learning*, springer, 2006.

[20] 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.

[21] 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.

[22] 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.

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