UNSUPERVISED UTTERANCE-WISE BEAMFORMER ESTIMATION WITH SPEECH RECOGNITION-LEVEL CRITERION

Takuya Higuchi, Takuya Yoshioka, Keisuke Kinoshita and Tomohiro Nakatani

NTT Communication Science Laboratories, NTT Corporation
{higuchi.takuya, yoshioka.takuya, kinoshita.k, nakatani.tomohiro}@lab.ntt.co.jp

ABSTRACT

In this paper, we perform beamforming with a speech recognition-level criterion. A beamformer is usually designed by optimizing signal-level criteria, e.g., by minimizing the beamformer output covariance or by maximizing the signal-to-noise ratio (SNR). Such signal-level criteria do not always guarantee that the optimized beamformer is the best for noise robust automatic speech recognition. Recently, a few approaches have been proposed for performing beamforming with a speech recognition-level criterion. These approaches train beamformers along with an acoustic model by using multichannel training data and a parallel corpus of noisy and clean data. This paper proposes a novel approach for estimating the beamformer for every test utterance with a speech recognition-level criterion. We use an unsupervised acoustic model adaptation scheme to optimize our beamformer. Specifically, we first obtain decoding results with an initialized beamformer, and then we optimize our beamformer using back propagation to minimize the cross entropy between the first-pass decoding results and actual network outputs. With this approach, our beamformer can be trained to discriminate hidden Markov model states more clearly for every test utterance. Experimental results show that our beamformer outperforms a beamformer designed with a signal-level criterion.

Index Terms— Beamforming, automatic speech recognition, acoustic model adaptation

1. INTRODUCTION

This paper deals with beamformer estimation for noise robust automatic speech recognition (ASR). Beamforming is a well-known technique for background noise suppression and has been shown to be a promising approach for noise robust ASR [1, 2]. By applying a linear filter to multichannel signals recorded by a microphone array, a beamformer can enhance a target speech signal and helps us to achieve more accurate ASR.

To construct the beamforming filter, we need to design a criterion to be optimized. For example, a minimum variance distortion-less response (MVDR) beamformer can be obtained by minimizing the covariance of the beamformer outputs without introducing distortion into the speaker direction parameterized by a steering vector. A maximum SNR (max-SNR) beamformer can be obtained by maximizing the SNR of the beamformer outputs. A delay and sum beamformer can be obtained based on a direction-of-arrival (DOA) estimate and a plane wave assumption. These beamformers are estimated by optimizing signal-level criteria, therefore ASR systems that perform ASR after the beamforming are completely excluded from consideration. Although these beamformers are effective for ASR, their optimality for ASR cannot be guaranteed with such signal-level criteria.

A few approaches have recently been proposed for optimizing beamformers with an acoustic model [3, 4, 5]. Sainath et al. have proposed a convolutional neural network (CNN)-based approach, where a multichannel beamforming filter is implemented as a convolutional filter in a CNN and connected to subsequent neural networks. The overall network can be regarded as a large acoustic model, whose input is a multichannel time domain signal, and whose output is a hidden Markov model (HMM) state posterior. The acoustic model is trained on training data by minimizing the cross entropy between correct labels and acoustic model outputs. Similarly, Xiao et al. proposed deep beamforming networks [5] that estimate the beamforming filter from a generalized cross correlation (GCC) [6] between microphones. The deep beamforming networks are jointly trained with an acoustic model by minimizing the cross entropy. While these approaches allow us to perform beamforming with the cross entropy criterion, they do not adapt the beamformer to unseen environments. In addition, these approaches require a parallel corpus of clean and noisy data to perform multi-task learning [4] or to train the beamformer part of neural networks in advance [5].

This paper proposes a technique for beamformer estimation for test environments, where the estimation is performed with the cross entropy (CE) criterion. We recently proposed a front-end optimization method with the CE criterion [7], where denoising was achieved with time-frequency masking. In contrast to our previous work, this paper applies our front-end optimization method to beamforming, which we have shown to be more effective for noise robust ASR than time-frequency masking [1].

We employ an unsupervised acoustic model adaptation scheme (see, e.g., [8]), but the beamformer speech enhancement parameter for multichannel observation is optimized unlike the acoustic model adaptation. In particular, we first perform decoding with an initialized beamformer and obtain supervision binary labels. Then we perform back propagation to minimize the cross entropy between the supervision binary labels and actual acoustic model outputs. The beamformer is estimated for every test utterance to enable the subsequent acoustic model to better discriminate the HMM state posterior. The number of estimated parameters for the time-invariant beamformer is relatively small, which helps us to avoid overfitting to the errorful supervisions. Finally, we perform decoding with the estimated beamformer, and obtain the final decoding results. Our experimental results show the effectiveness of our proposed beamformer estimation in terms of word error rate (WER) compared with an MVDR beamformer estimated with a signal-level criterion.

2. BEAMFORMING

This section briefly describes a beamformer in the time-frequency domain. Let $y_{f,t,m}$ denote the m-th microphone signal at frequency f and time t. By using vector notation the signals from all M microphones can be represented as

$$\mathbf{y}_{f,t} = [y_{f,t,1}, \ldots, y_{f,t,M}]^{\mathrm{T}}, \tag{1}$$

where the superscript T denotes non-conjugate transposition.

The beamformer, which is represented as a linear filter in the frequency domain, can be described by using vector notation as

$$\mathbf{w}_f = [w_{f,1}, \ldots, w_{f,M}]^{\mathrm{T}}. \tag{2}$$

An enhanced speech signal $\hat{s}_{f,t}$ can be obtained by multiplying the filter by the observed signal as

$$\hat{s}_{f,t} = \mathbf{w}_f^{\mathrm{H}} \mathbf{y}_{f,t}, \tag{3}$$

where the superscript H denotes conjugate transposition.

The key to successful beamforming is the accurate estimation of the filter $\mathbf{w}_f$ for every test environment. Various criteria have been proposed for beamformer estimation, most of which are designed in the audio signal space. The aim of this study is to estimate the beamformer with the speech recognition-level criterion for every test utterance.

3. PROPOSED BEAMFORMER ESTIMATION

In this section, we first provide an overview of our proposed beamformer estimation, which comprises beamformer initialization, first-pass decoding and beamformer estimation by back propagation. Then we describe in detail these three steps of our proposed method. We also provide an interpretation of our proposed method in comparison with existing acoustic model adaptation approaches.

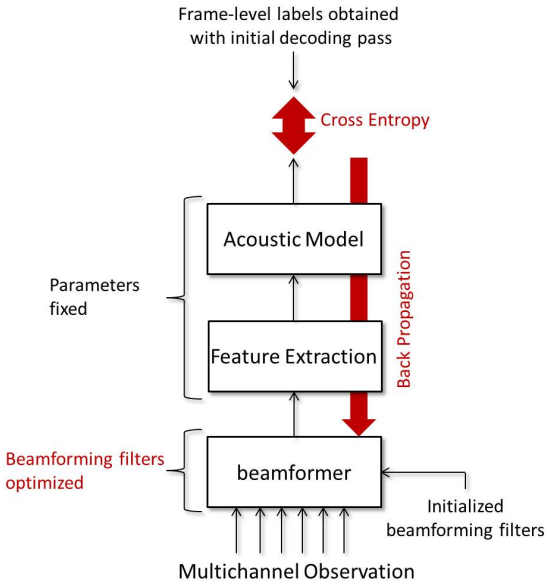

Fig. 1. Overview of our beamformer estimation.

3.1. Overview

Figure 1 shows an overview of our proposed beamformer estimation method, which is based on a standard unsupervised acoustic model adaptation approach. First, we initialize the beamformer filter by using an existing beamforming method to leverage recently developed powerful beamforming approaches. Second, we perform decoding with the initialized beamformer to obtain supervision labels for beamformer adaptation. The beamforming and feature extraction are reformulated as computational layers of neural networks, and connected to the acoustic model. The beamforming filters are finally optimized for every test utterance by using back propagation, which minimizes the cross entropy between the supervision labels and acoustic model outputs. This beamformer estimation allows the beamforming filters to take account of the subsequent recognition process, and the filters are optimized to make the recognizer better discriminate the HMM state posteriors. The optimized beamformer is used to yield the final deciding results.

3.2. Beamformer initialization

We can use any existing beamformer estimation approach to obtain initial values for the beamforming filters, which are subsequently optimized with our proposed method. Recently, several beamforming approaches have been shown to yield great performance gains [1, 2, 9], even though the back-end recognizer is not considered in the beamforming front-end. We can leverage these existing approaches by using the estimated filters as the initial values, and this actually yielded better performance in our experiments (see section 4 for more details).

3.3. First-pass decoding to obtain supervision labels

To perform beamformer adaptation with the CE criterion, we need to obtain supervision labels. We follow the unsupervised acoustic model adaptation scheme to obtain the supervision labels (see, e.g., [8] for details of deep neural network (DNN) adaptation). We first perform forwarding with the initialized beamformer and an acoustic model trained in advance to obtain estimated HMM state posteriors. The posteriors and a language model are used to obtain initial decoding results, i.e, a sequence of estimated words. Then, we perform forced alignment and obtain a supervision label corresponding to every time frame. Following this scheme, obtained supervision labels are different from the acoustic model outputs in the following two respects; the supervision labels are refined by the language model, and the supervisions are binary values while the acoustic model outputs are continuous values from zero to one. Minimizing the CE between the supervision labels and the acoustic model outputs makes the acoustic model outputs closer to binary values, which are refined by the language model. Even though adapting too many parameters would result in reproducing the same decoding results as those of first-pass decoding, we retrain just the beamformer filters, which can produce enhanced signals only by focusing on a specific direction. This constraint of the retrained parameters, i.e., beamforming filters, helps us to avoid overfitting to the errorful supervisions and to obtain performance gain by forwarding again with the adapted beamformer.

3.4. Back propagation for beamformer estimation

Our beamformer filters are retrained based on the gradient descent algorithm by performing back propagation with the CE criterion. Our objective function $\mathcal{L}(l^{sv}, \hat{l})$, the CE between the supervision labels and the acoustic model outputs, can be described as

$$\begin{aligned}\mathcal{L}(l^{sv}, \hat{l}) &= \sum_t \mathcal{L}(l_t^{sv}, \hat{l}_t) \\ &= \sum_t CrossEntropy(l_t^{sv}, \hat{l}_t),\end{aligned} \tag{4}$$

where l^{sv} is the 1-best binary labels generated by the initial decoding pass, and $\hat{l}$ is the acoustic model outputs, i.e., the estimated posteriors. The filter can be updated based on the gradient descent algorithm as

$$\mathbf{w}_f \leftarrow \mathbf{w}_f - \alpha \frac{1}{T} \sum_t \frac{\partial \mathcal{L}(l_t^{sv}, \hat{l}_t)}{\partial \mathbf{w}_f^*}, \tag{5}$$

where $\mathbf{w}_f^*$ and α denote the conjugate of the filter $\mathbf{w}_f$ and the learning rate respectively.

Now we reformulate beamforming and feature extraction as neural network layers to calculate the gradient with respect to the beamforming filter. The acoustic model outputs $\hat{l}$ can be calculated with Eq. (3) and

$$\hat{x} = FeatureExtract(\hat{s}), \tag{6}$$

$$\hat{l} = AcousticModel(\hat{x}), \tag{7}$$

where $\hat{x}$ denotes extracted features. $FeatureExtract(\hat{s})$ denotes a function for extracting features from the enhanced signals, and it is usually parameterized by fixed parameters (e.g. log-mel feature extraction and feature normalization by affine transform). $AcousticModel(\hat{x})$ denotes an acoustic model for computing HMM posteriors from the acoustic features. The acoustic model is trained in advance by using training data. The overall computation process can be regarded as a large model that outputs the HMM state posteriors from multichannel observations.

From Eqs. (3), (6) and (7), the gradient of the CE with respect to the conjugate of the beamforming filter can be computed by the chain rule as

$$\frac{\partial \mathcal{L}(l_t^{sv}, \hat{l}_t)}{\partial \mathbf{w}_f^*} = \frac{\partial \mathcal{L}(l_t^{sv}, \hat{l}_t)}{\partial \hat{x}_t} \cdot \frac{\partial \hat{x}_t}{\partial \hat{s}_{f,t}} \cdot \frac{\partial \hat{s}_{f,t}}{\partial \mathbf{w}_f^*}, \tag{8}$$

where $\frac{\partial \hat{s}_{f,t}}{\partial \mathbf{w}_f^*}$ can be described as

$$\frac{\partial \hat{s}_{f,t}}{\partial \mathbf{w}_f^*} = \mathbf{y}_{f,t}^{\mathrm{T}}, \tag{9}$$

from Eq. (3). $\frac{\partial \mathcal{L}(l_t^{sv}, \hat{l}_t)}{\partial \hat{x}_t}$ can be computed because the gradient is used in the acoustic model training. $\frac{\partial \hat{x}_t}{\partial \hat{s}_{f,t}}$ can also be computed because the feature extraction is described by basic computations, i.e., addition, multiplication, the log function and the power operation.

3.5. Interpretation of proposed method

The proposed adaptation scheme is related to other adaptation approaches such as the DNN retraining approach investigated in [10], the linear input network (LIN) [11], learning hidden unit contributions (LHUC) [12], and feature-space discriminative linear regression (fDLR) [13]. Indeed, when considering the combination of the beamformer and the acoustic model as a single neural network, the approach becomes similar to retraining the parameters of the first layer, LIN with a diagonal linear transformation matrix, or LHUC applied to the input layer. The major difference is that the proposed approach operates in the complex domain, whereas LIN and other adaptation approaches are usually employed after the feature extraction process. By performing adaptation in the complex domain and on multiple channel signals, we are able to exploit the spatial information. We will investigate the combination of our approach with other adaptation techniques in future work.

Table 1. WERs on the real data in the development and evaluation sets.

systems	dev					eval				
	avg	bus	caf	ped	str	avg	bus	caf	ped	str
w/o speech enhancement	10.33	15.93	8.61	7.08	9.72	16.80	24.30	18.01	12.91	11.97
Proposed beamformer w/ ref-mic init.	10.10	15.46	8.55	6.86	9.53	16.33	23.78	17.61	12.28	11.65
CGMM-MVDR beamformer [1, 2]	5.97	8.84	4.88	4.43	5.72	9.06	12.30	7.58	8.52	7.83
Proposed beamformer w/ CGMM-MVDR init.	**5.79**	8.28	4.85	4.40	5.62	**8.89**	11.95	7.45	8.31	7.84

4. EXPERIMENTAL EVALUATION

4.1. Settings

We conducted experiments using the CHiME-3 corpus [14] to evaluate the effectiveness of our proposed beamformer in terms of WER. The corpus consists of read utterances that were recorded with six microphones attached to a tablet device in four different environments: public tranportation (bus), café (caf), pedestrian area (ped), and street junction (str). The sentences were taken from the WSJ0 corpus. The training set comprises 1600 real and 7138 simulated utterances. The audio data from all six microphones were used for training, which amounts to about 108 hours. The development and evaluation sets consist of 1640 and 1320 real utterances, respectively.

In our experiments, we performed speaker independent decoding by using a deep convolutional neural network (CNN) acoustic model [1, 15, 16] and a recurrent neural network (RNN) language model [17, 18]. The CNN consisted of five convolution layers and two max-pooling layers, where all the layers contained 180 feature maps. The last convolution layer was followed by two fully connected layers with 2048 units and a softmax layer. The softmax layer contained 5976 units, i.e., context-dependent HMM states. The RNN language model used 10 classes and accommodated 250 units in the hidden recurrent layer. The inputs to the acoustic model comprised 80-dimensional log mel-filter bank channel outputs, where the filter bank outputs at 19 time frames were concatenated as an input for a center time frame. Utterance-wise mean normalization was performed after the log mel-filter bank feature extraction, which was followed by feature normalization using the first- and second-order statistics obtained from all the training data.

We considered two methods for our beamformer initialization. One is called reference microphone initialization, where all beamformer components are initialized to extract a single-channel observation recorded with a reference microphone. The other initialization method utilizes the unsupervised masking-based beamformer described in [1, 2]. In particular, we first performed time-frequency mask estimation based on a complex Gaussian mixture model (CGMM) by maximizing the log-likelihood criterion [19]. Then, the steering vector of the target speaker was extracted as the eigenvector associated with the maximum eigenvalue of the covariance matrix of the target speech, where the covariance was calculated with the estimated masks and multichannel observed signals. The initial beamformer was obtained as the MVDR beamformer that was parameterized by the estimated steering vector. The beamformer was updated 30 times with Eq. (5) by using an utterance-batch processing approach. The learning rate α was set at 4×10^3 for the reference microphone initialization, and set at 6×10^3 for the masking-based beamformer initialization. These learning rates were tuned using the development set. Other experimental conditions were set as in Table 2.

Table 2. Experimental conditions.

Sampling frequency	16 kHz
Frame length	25 ms
Frame shift	10 ms
Window function	Hanning

We compared the performance of the proposed method with the beamformer obtained with the signal-level criterion [1, 2] that was used for the beamformer initialization.

4.2. Results

Table 1 shows the WERs obtained with the proposed and competing methods. With the reference microphone initialization, our proposed beamformer estimation achieved a WER improvement compared with the WER obtained without speech enhancement, while its improvement was limited and smaller than that obtained with the MVDR beamformer. While the MVDR beamformer greatly reduced the WERs for all the environments, our proposed beamformer estimation yielded a further performance gain by using the MVDR beamformer for our beamformer initialization. These results shows the effectiveness of our proposed beamformer estimation.

5. CONCLUSION

This paper proposed a novel approach for unsupervised beamformer estimation with a CE criterion. We first performed initial decoding to obtain the supervision labels with the initialized beamformer. Then, our beamformer was estimated for every utterance by minimizing the CE between the supervision labels and the acoustic model outputs. The estimated beamformer enabled the acoustic model to better discriminate the HMM states. Experimental results showed that our beamformer outperformed the conventional beamformer obtained with a signal-level criterion in terms of WER for the CHiME-3 evaluation set.

6. REFERENCES

[1] T. Yoshioka, N. Ito, M. Delcroix, A. Ogawa, K. Kinoshita, M. Fujimoto, C. Yu, W. J. Fabian, M. Espi, T. Higuchi, S. Araki, and T. Nakatani, "The NTT CHiME-3 system: advances in speech enhancement and recognition for mobile multi-microphone devices," in *Proc. Worksh. Automat. Speech Recognition, Understanding*, 2015, pp. 436–443.

[2] T. Higuchi, T. Yoshioka, N. Ito, and T. Nakatani, "Robust MVDR beamforming using time-frequency masks for online/offline ASR in noise," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2016, pp. 5210–5214.

[3] Tara N. Sainath, Ron J. Weiss, Kevin W. Wilson, Arun Narayanan, Michiel Bacchiani, and A. Senior, "Speaker localization and microphone spacing invariant acoustic modeling from raw multichannel waveforms," in *Proc. Worksh. Automat. Speech Recognition, Understanding*, 2015, pp. 30–36.

[4] Tara N. Sainath, Ron J. Weiss, Kevin W. Wilson, Arun Narayanan, and Michiel Bacchiani, "Factored spatial and spectral multichannel raw waveform CLDNNs," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2016, pp. 5075–5079.

[5] Xiong Xiao, Shinji Watanabe, Hakan Erdogan, Liang Lu, John Hershey, Michael L Seltzer, Guoguo Chen, Yu Zhang, Michael Mandel, and Dong Yu, "Deep beamforming networks for multi-channel speech recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2016, pp. 5745–5749.

[6] Charles Knapp and Glifford Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976.

[7] T. Higuchi, T. Yoshioka, and T. Nakatani, "Optimization of speech enhancement front-end with speech recognition-level criterion," in *Proc. Interspeech*, 2016, pp. 3808–3812.

[8] Dong Yu and Li Deng, *Automatic speech recognition: a deep learning approach*, Springer, 2015.

[9] Jahn Heymann, Lukas Drude, and Reinhold Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2016, pp. 196–200.

[10] Hank Liao, "Speaker adaptation of context dependent deep neural networks," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2013, pp. 7947–7951.

[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] Pawel Swietojanski and Steve Renals, "Learning hidden unit contributions for unsupervised speaker adaptation of neural network acoustic models," in *Spoken Language Technology Workshop (SLT), 2014 IEEE*. IEEE, 2014, pp. 171–176.

[13] Frank Seide, Gang Li, Xie Chen, and Dong Yu, "Feature engineering in context-dependent deep neural networks for conversational speech transcription," in *Proc. Worksh. Automat. Speech Recognition, Understanding*, 2011, pp. 24–29.

[14] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME' speech separation and recognition challenge: dataset, task and baselines," in *Proc. Worksh. Automat. Speech Recognition, Understanding*, 2015, pp. 504–511.

[15] O. Abdel-Hamid, A.-r. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 10, pp. 1533–1545, 2014.

[16] T. N. Sainath, B. Kingsbury, G. Saon, H. Soltau, A.-r. Mohamed, G. Dahl, and B. Ramabhadrana, "Deep convolutional neural networks for large-scale speech tasks," *Neural Networks*, vol. 64, pp. 39–48, 2015.

[17] T. Mikolov, M. Karafiat, L. Burget, J. Cernocky, and S. Khudanpur, "Recurrent neural network based language model," in *Proc. Interspeech*, 2010, pp. 1045–1048.

[18] T. Mikolov, A. Deoras, S. Kombrink, L. Burget, and J. Cernocky, "Empirical evaluation and combination of advanced language modeling techniques," in *Proc. Interspeech*, 2011, pp. 605–608.

[19] N. Ito, S. Araki, T. Yoshioka, and T. Nakatani, "Relaxed disjointness based clustering for joint blind source separation and dereverberation," in *Proc. Int. Worksh. Acoust. Echo, Noise Contr.*, 2014, pp. 268–272.