2020 RISP International Workshop on Nonlinear Circuits, Communications and Signal Processing (NCSP 2020)
Honolulu, Hawaii, USA, February 28 - March 2, 2020

Underdetermined Multichannel Speech Enhancement Using Time-Frequency-Bin-Wise Switching Beamformer and Gated CNN-Based Time-Frequency Mask for Reverberant Environments

Riki Takahashi[1], Kouei Yamaoka[2], Li Li[1], Shoji Makino[1], Takeshi Yamada[1], Mitsuo Matsumoto[1]

[1]University of Tsukuba, Japan
E-mail: {r.takahashi@mmlab.cs, lili@mmlab.cs, maki@tara, takeshi@cs, makilab-research@tara}.tsukuba.ac.jp
[2]Tokyo Metropolitan University, Japan
E-mail: yamaoka-kouei@ed.tmu.ac.jp

Abstract

To achieve high performance of speech enhancement with low signal distortion under underdetermined conditions, the time-frequency-bin-wise switching (TFS) beamformer has recently been proposed. Its extension that applies a time-frequency (TF) mask constructed on the basis of direction of arrival (DOA) estimation for postprocessing, has been shown to further improve the performance. However, the precision of the TF mask estimated using spatial information strongly depends on acoustic factors such as reverberation. To address this issue and improve the accuracy of the mask estimation, in this paper, we propose a novel method that combines the TFS beamformer with a deep clustering (DC)-based TF mask, which is estimated using monaural signals. We investigate the performance of the TFS beamformer under various reverberation conditions. The experiments revealed that the proposed method further improved the performance of the TFS beamformer in speech enhancement even under heavy reverberation, whereas the conventional method using DOA estimation failed.

1. Introduction

In recent years, with the development of automatic speech recognition (ASR) and robot hearing, the importance of speech enhancement has considerably increased. Owing to the wide usage of stereo microphone built-in small devices such as smartphones and voice recorders, speech enhancement that serves dual-channel signals is a particular requisite. Beamforming and blind source separation (BSS) [1] are the two main methods to deal with this problem. Beamforming and BSS methods using spatial filtering [2–4] are noteworthy in the low distortion of the enhanced target speech. However, their performance degrades when there are fewer microphones than sources, i.e., underdetermined conditions.

On the other hand, methods applying time-frequency (TF) masking or multichannel Wiener filters [5–8] can sufficiently suppress noise and interferers under underdetermined conditions, while they are tend to distort the target speech. To accomplish high noise suppression performance with low distortion under underdetermined conditions, some attempts have recently been made to combine beamforming with masking techniques on the basis of the sparseness assumption of speeches [9–11]. In [9], a limited amount of pretrained dictionaries of mechanical noise covariance to construct the most suitable filter in each TF bin so that the output includes fewer noise residuals. Similar to [9], in [10,11], the time-frequency-bin-wise switching (TFS) beamformer, which applies the optimal filter selected from a set of predefined beamformers according to the minimum absolute value of outputs to each TF bin, has been proposed for underdetermined speech enhancement. All of these methods can be interpreted as adopting beamformers to multiple masked spectrograms summing up to the observed one, which exploits the benefits of both the beamforming technique and TF masking.

Taking account of its low computational cost and high noise suppression performance, in this paper, we consider the TFS beamformer. With the assumption that only the target speech and one interferer exist at each TF bin, all $N-1$ interferers can be suppressed by constructing $N-1$ beamformers and switching them among the TF bins. Note that N can be larger than the number of microphones M. However, there are still interferer components remaining when the assumption does not hold, namely, multiple interferers exist simultaneously at one TF bin. To further suppress these residuals, a TF mask constructed on the basis of the directions of arrival (DOA) of the sources is employed for postprocessing [10]. It has been shown to be effective in improving the performance of speech enhancement. However, the precision of the constructed TF mask tends to decrease in highly reverberant environments, where accurately estimating the DOA of sources becomes an extremely difficult task [12–14]. To address this drawback and increase the accuracy of mask estimation, in this paper, we propose the use of a mask estimation method based on deep neural networks for the postprocessing of the TFS beamformer, where a TF mask is estimated using monaural signals instead of multichannel signals. DNN-based TF masking [15,16] has recently gained much attention because of its impressive performance in speech separation. Specifically, we adopt deep clustering (DC) [17] with gated convolutional neural networks (CNNs) to take the advantage that high mask estimation accuracy can be achieved with a remarkably small amount of training data [18].

2. Time-frequency-bin-wise switching beamformer

Let $x_i(\omega,t)$ be the ith microphone signal at the angular frequency ω in the tth time frame in the short-time Fourier transform (STFT) domain. With two microphones, a beamformer is generally given by the following equations:

$$y(\omega,t) = \mathbf{w}^{\mathrm{H}}(\omega)\mathbf{x}(\omega,t), \quad (1)$$

$$\mathbf{x}(\omega,t) = [x_1(\omega,t), x_2(\omega,t)]^{\mathrm{T}}, \quad (2)$$

$$\mathbf{w}(\omega) = [w_1(\omega), w_2(\omega)]^{\mathrm{T}}, \quad (3)$$

Table 1: Dominant sound patterns at each TF bin of different signals. tgt, i_1, and i_2 stand for target speech, interferer 1, and interferer 2, respectively.

row	$\mathbf{x}$	y_1	y_2	y_{TFS}	y_{DC}
1	tgt	tgt	tgt	tgt	tgt
2	i_1	0	i_1	0	0
3	i_2	i_2	0	0	0
4	tgt, i_1	tgt	tgt, i_1	tgt	tgt
5	tgt, i_2	tgt, i_2	tgt	tgt	tgt
6	i_1, i_2	i_2	i_1	i_1 or i_2	0
7	tgt, i_1 , i_2	tgt, i_2	tgt, i_1	tgt, i_1 or tgt, i_2	tgt, i_1 or tgt, i_2

where $y(\omega, t)$ is the output of the beamformer, $\mathbf{w}(\omega)$ denotes the spatial filter vector, $(\cdot)^{\mathrm{T}}$ denotes the transpose, and $(\cdot)^{\mathrm{H}}$ denotes the Hermitian transpose. For the design of the spatial filter $\mathbf{w}(\omega)$, adaptive beamformers, e.g., the minimum variance distortionless response (MVDR) beamformer [19, 20], are widely used. Theoretically, a beamformer with M microphones can suppress up to $M-1$ interferers, which means that speech enhancement with a beamformer cannot achieve satisfactory performance under underdetermined conditions ($M < N$).

To achieve sufficient speech enhancement performance under such conditions, the TFS beamformer [10] has been proposed, where the optimal filter selected from multiple beamformers is used at each TF bin. For simplicity, we consider a case where a three-source mixture $\mathbf{x}(\omega, t)$ consisting of a target speech and interferers 1 and 2 is observed by two microphones. In this case, we cannot construct a null beamformer that suppresses both interferers simultaneously. Supposing that only the target speech and interferer k are observed (k = 1, 2), we can construct the beamformer k with a spatial filter $\mathbf{w}_k(\omega)$ that suppresses the interferer k by a conventional linear beamforming method. By applying the K constructed beamformers, we obtain K outputs from the observed signal $\mathbf{x}(\omega, t)$:

$$y_k(\omega, t) = \mathbf{w}_k^{\mathrm{H}}(\omega)\mathbf{x}(\omega, t), k = 1, \ldots, K. \quad (4)$$

Here, $\mathbf{w}_k(\omega)$ is the spatial filter of the beamformer k. By considering statistically independent target speech and interferers, we can assume that the magnitude of the target speech is smaller than that of the sum of the target speech and any interferer. The TFS beamformer is then performed according to the following criterion:

$$y_{\mathrm{TFS}}(\omega, t) = \begin{cases} y_1(\omega, t) & \text{if } |y_1(\omega, t)| \leq |y_2(\omega, t)|, \\ y_2(\omega, t) & \text{otherwise.} \end{cases} \quad (5)$$

Equation (5) implies that the TFS beamformer selects the output of the beamformer with the smaller magnitude, which is equivalent to perform optimal filter selection in terms of minimizing the output. Table 1 summarizes the seven patterns of the dominant sound source at each TF bin in the three-source mixture case. In the case in the 5th or 6th row of Table 1, one beamformer outputs the target speech, and the other one outputs the target speech and a slightly altered version of an interferer. According to the previously mentioned assumption, the TFS beamformer selects the output only including the target speech. In the case in the 7th or 8th row of Table 1, the TFS beamformer cannot suppress both interferers because there are M or more interferers in a TF bin, whereas the TFS beamformer suppresses only the most dominant interferer.

3. Postprocessing with deep clustering

3.1 Proposed method

Speech enhancement by the TFS beamformer resulted in a performance of high noise reduction with less distortion. However, as shown in the 7th and 8th rows of Table 1, either interferer 1 or 2 still remains. To reduce these residual components without distorting the target speech, we consider using the following TF mask

$$\hat{y}_(\omega, t) = m(\omega, t) y_{\mathrm{TFS}}(\omega, t), \quad (6)$$

where $m(\omega, t)$ is a soft or binary mask, the element of which equals to 1 at those TF bins including the target speech components regardless of the existence of interferers. $\hat{y}_(\omega, t)$ represents $y_{\mathrm{TFS}}(\omega, t)$ for the conventional method and $y_{\mathrm{DC}}(\omega, t)$ for the proposed method. In [10], the mask is constructed on the basis of DOA estimation, where the accuracy of the mask directly depends on the DOA estimates. In long reverberant environments, however, it is difficult to estimate DOA accurately [12–14] since the correct spatial information, namely, the phase difference, is difficult to obtain because of reflections. Moreover, since DOA estimation is performed for the observed multichannel signals, which is independent of the TFS beamformer, the benefits gained from the higher signal-to-noise ratio (SNR) thus cannot be exploited for the DOA estimation. To deal with this issue, motivated by the remarkable results recently achieved using a DNN in monaural source separation, we propose using a DNN to estimate the TF mask. Using training data and appropriately designed labels, namely, the speech enhanced by the TFS beamformer and DNN-based mask defined as

$$m(\omega, t) = \begin{cases} 1 & |s(\omega, t)| > \text{threshold} \\ 0 & \text{otherwise,} \end{cases} \quad (7)$$

we can train a DNN to estimate the TF mask for the output of the TFS beamformer. Here, $s(\omega, t)$ denotes the target speech. Note that the W-DO assumption [21] is not required. The enhanced speech $y(\omega, t)$ is finally obtained using for (6). The process flow of the proposed method is shown in Fig. 1. Specifically, we employ DC [17] with a gated CNN [18] for DNN-based mask estimation. The main difference between gated CNN and typical CNN is that a gated linear unit (GLU), the second term of (9), is used as nonlinear activation function instead of, e.g., regular rectified linear units. Since a GLU controls information passed on in the hierarchy, the gated CNN can capture long-term dependences without the vanishing gradient problem even with many layers.

3.2 Deep clustering with gated CNN

On the basis of the assumption that each TF bin of a mixture signal is dominated by a single source, i.e., W-DO, DC [17] separates sources by clustering feature vectors embedded from each TF bin. Let $\mathbf{X} = \{X_c\} \in \mathbb{R}^{C \times 1}$, where c

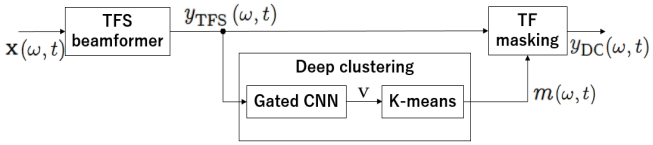

Figure 1: Process flow of TFS+DC.

denotes a pair of frequency and time indices (ω, t) and C is equal to the number of TF bins $\Omega \times T$. A set of feature vectors $\mathbf{V}_c = (V_{c,1}, \ldots, V_{c,D})^T$, which are called embedding vectors, are obtained by a DNN $\mathbf{V} = g_\Theta(\mathbf{X})$, where $g(\cdot)$ denotes a nonlinear function, Θ denotes the parameters of the network, and $\mathbf{V} = \{V_c\} \in \mathbb{R}^{C \times D}$ is a matrix consisting of embedding vectors in a row. The network is trained by minimizing the following objective function:

$$\begin{aligned}\mathcal{J}(\mathbf{V}) &= ||\mathbf{V}\mathbf{V}^T - \mathbf{Y}\mathbf{Y}^T||_F^2 \\ &= ||\mathbf{V}^T\mathbf{V}||_F^2 - 2||\mathbf{V}^T\mathbf{Y}||_F^2 + ||\mathbf{Y}^T\mathbf{Y}||_F^2, \end{aligned} \quad (8)$$

where $||\cdot||$ is the squared Frobenius norm and $\mathbf{Y} = \{Y_{c,n}\} \in \mathbb{R}^{C \times N}$ is a matrix composed of N-dimensional one-hot vectors in a row indicating to which source each TF bin belongs. During the test, a clustering algorithm (e.g., K-means) is applied to the assigned embedding vectors to construct binary masks for each source. Since the aim of using DC is to further enhance the speech outputted by the TFS beamformer, in the proposed method, the input is $y_{\mathrm{TFS}}(\omega, t)$ and only the mask for the target speech is estimated. For the network architecture, we use a gated CNN [18].

By using $\mathbf{H}_{l-1}$ to denote the output of the $(l-1)$th layer, we obtain the output $\mathbf{H}_l$ of the lth layer as

$$\mathbf{H}_l = (\mathbf{H}_{l-1} * \mathbf{W}_l^{\mathrm{f}} + b_l^{\mathrm{f}}) \otimes \sigma(\mathbf{H}_{l-1} * \mathbf{W}_l^{\mathrm{g}} + b_l^{\mathrm{g}}), \quad (9)$$

where $\mathbf{W}_l^{\mathrm{f}}$ and $\mathbf{W}_l^{\mathrm{g}}$ are the weight parameters, b_l^{f} and b_l^{g} are the bias parameters of the lth layer, $\otimes$ denotes element-wise multiplication, and σ is the sigmoid function.

3.3 Advantages of combining TFS beamformer with DNN-based TF masking

Comparing with the method combining the TFS beamformer with DOA-based TF masking, the proposed method has the following merits. First, since the TFS beamformer suppresses one of the interferers in each TF bin, the input signal of the DNN, i.e., the output of the TFS beamformer, is more sparse than the observed signals, which facilitates the mask estimation. Second, the performance of noise reduction can be controlled by adjusting the threshold in (7). For example, by setting the threshold at a sufficiently small value (e.g., 10^{-5}), we can mainly suppress interferers described at the 7th row in Table 1 while holding the distortion of the target speech minimum. Since the TFS beamformer can guarantee a distortionless response to the target speech, forcing the TF mask to keep this property is important to realize distortionless enhancement. If we set the threshold at a relatively large value, the DNN is trained to suppress all the components having less energy than the threshold so that interferers described at the 8th row in Table 1 having small energy can also be suppressed.

4. Experiments

4.1 Experimental conditions

To evaluate the effect of the proposed method, we conducted experiments using the ATR503 database [22]. The database consists of four male and six female speakers, and the number of audio files uttered by each speaker is 503. We divided the entire database into training and test sets. The training set consisted of six speakers and 450 utterances of each speaker. The remaining 53 utterances spoken by the other four speakers were used as the test set so that the experiments was conducted in a speaker- and utterance-independent manner.

We generated a total of 720 mixture signals (about 1 hour) for training the network by summing up three convolved speech signals randomly selected from the training set. Room impulse responses (RIRs) with reverberation times of {300, 400, 500, 600, 700, 800}ms were generated using the image method with the codes available in [23]. The DOA of the target speech was fixed at 90° and those of interferers were set at 70° and 130°. All the audio files were downsampled to 8 kHz. We computed complex spectrograms using a Hann window with a length of 512 ms and shift of 256 ms. To obtain the speeches enhanced by the TFS beamformers y_{TFS} and the corresponding labels, we applied a TFS beamformer with an MVDR beamformer [10] to the generated mixture signals and computed the label using (7) with threshold set at 10^{-5}. We trained two networks using data under each reverberation condition.

Mixture signals for the test with reverberation times at {300, 400, 500, 600, 700, 800}ms were generated in the same way, whereas DOAs of interferers were randomly selected from $\{10^\circ, 30^\circ, 50^\circ, 70^\circ, 110^\circ, 130^\circ, 150^\circ, 170^\circ\}$. For each reverberation condition, we generated 10 mixture signals, which were about 5 seconds long. Other experimental conditions are shown in Table 2.

We calculated signal-to-distortion ratio (SDR), signal-to-interference ratio (SIR), and signal-to-artifacts ratio (SAR) [24] and compared the performance of the proposed method TFS+DC with the degenerate unmixing estimation technique (DUET) [6], which is BSS via TF masking with a stereo microphone, the TFS beamformer (TFS) [10], and the TFS beamformer with the DOA-based mask (TFS+DOA) [10].

4.2 Results and discussion

Figure 2 shows results achieved by each method under closed reverberation time conditions. It is noteworthy that TFS+DC led to few artifacts, which are detrimental to many speech processing applications such as ASR. The methods using the TFS beamformer significantly outperformed DUET. The SDRs obtained by TFS+DOA and TFS+DC were higher than or equivalent to that achieved by the TFS beamformer, which confirmed the effectiveness of applying TF masking for postprocessing in improving the speech enhancement performance. Moreover, the proposed method achieved the best scores, especially in the relatively highly reverberant case where TFS+DOA failed to achieve further improvement from

Table 2: Experimental conditions.

Number of microphones	2
Distance between microphones	4 cm
Reverberation time (training/test)	300, 500 ms / 300, 400, 500, 600, 700, 800 ms
DOA of the target (training/test)	90° / 90°
DOA of interferers (training/test)	70°, 130°/ 10°, 30°, 50°, 70°, 110°, 130°, 150°, 170°
Distance between sources and microphones	2 m
Embedding dimension	20
Number of hidden layers	5
Number of channels	128
Learning rate	0.001
Optimizer	Adam
Minibatch size	4
Number of epochs	50

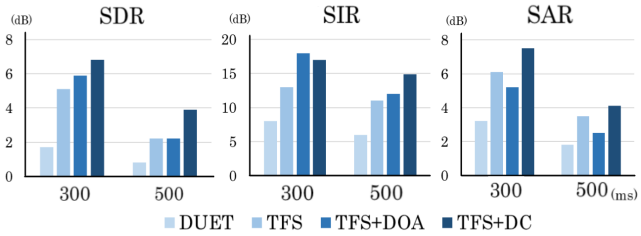

Figure 2: Average SDR, SIR, and SAR over 10 test signals under each reverberation condition.

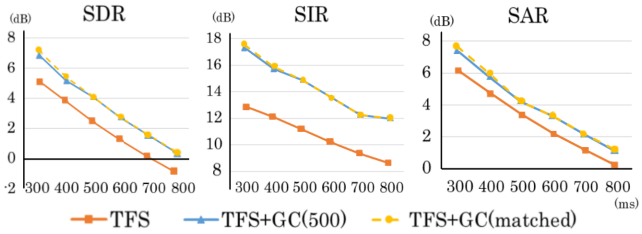

Figure 3: Results of speech enhancement by TFS and TFS+DC for reverberation times of 300–800 ms.

the TFS beamformer. These results confirmed the observation that the performance of TFS+DOA tends to decrease under highly reverberant conditions. They also confirmed the effectiveness of the proposed method.

To confirm the generalization ability of the trained network, we conducted an experiment employing the model trained under the 500 ms reverberant condition (TFS+DC(500)) to enhance signals under various reverberation conditions. The results are shown in Fig. 3. Although, as the reverberation time increased, the performances of the TFS beamformer and TFS+DC decreased in terms of all the criteria, TFS+DC always exceeded the TFS beamformer at a definite degree. The performance of using the model of TFS+DC(500) was very close to the results obtained using the model of TFS+DC(matched) which is trained on the data with the same reverberation time as the test data. The results indicates that the proposed method can accomplish further improvement regardless of the reverberation conditions. Furthermore, these results confirmed that the trained network generalized well to cope with a wide range of reverberation times.

5. Conclusion

In this paper, we proposed a combination of the TFS beamformer and TF masking based on DC with a gated CNN as an underdetermined speech enhancement method having high noise reduction performance with low signal distortion. The TFS beamformer efficiently suppresses interferers even under underdetermined conditions. Furthermore, by applying DNN-based TF masking for postprocessing, it is possible to suppress interferers that cannot be suppressed by the TFS beamformer. To confirm the effectiveness of the proposed method, we conducted simulation experiments in reverberant environments. As the results, we confirmed that the proposed method achieved superior speech enhancement performance to the conventional method with the DOA-based TF masking, especially in long reverberant environments. Additionally, we also confirmed that the proposed method can enhance the target speech while keeping the distortion of the target speech to the minimum with appropriately designed labels.

Acknowledgments

This work was supported by Japan Society for the Promotion of Science (JSPS) KAKENHI Grant Numbers JPH04131 and JP19J20420, and the SECOM Science and Technology Foundation.

References

[1] S. Makino *et al.*, *Blind speech separation*, Springer, 2007.
[2] P. Smaragdis, *Neurocomputing*, vol. 22, no. 1-3, pp. 21–34, 1998.
[3] D. Kitamura *et al.*, *IEEE/ACM TASLP*, vol. 24, no. 9, pp. 1622–1637, 2016.
[4] T. Higuchi *et al.*, *IEEE/ACM TASLP.*, vol. 25, no. 4, pp. 780–793, 2017.
[5] O. Yilmaz *et al.*, *IEEE TSP*, vol. 52, no. 7, pp. 1830–1847, 2004.
[6] S. Rickard, in *Blind Speech Separation*, pp. 217–241, Springer, 2007.
[7] H. Sawada *et al.*, *IEEE TASLP*, vol. 19, no. 3, pp. 516–527, 2011.
[8] H. Sawada *et al.*, *IEEE TASLP.*, vol. 21, no. 5, pp. 971–982, 2013.
[9] M. Togami *et al.*, *Proc. EUSIPCO*, pp. 741–745, 2010.
[10] K. Yamaoka *et al.*, *Proc. EUSIPCO*, pp. 1582–1586, 2018.
[11] K. Yamaoka *et al.*, *Proc. ICASSP*, pp. 7908–7912, 2019.
[12] O. Schwartz *et al.*, *IEEE/ACM TASLP*, vol. 22, no. 2, pp. 392–402, 2013.
[13] O. Schwartz *et al.*, *Proc. IWAENC*, pp. 1–5, 2016.
[14] J. R. Jensen *et al.*, *Proc. ICASSP*, pp. 176–180, 2016.
[15] P.-S. Huang *et al.*, *Proc. ICASSP*, pp. 1562–1566, 2014.
[16] D. Wang *et al.*, *IEEE/ACM TASLP*, vol. 26, no. 10, pp. 1702–1726, 2018.
[17] J. R. Hershey *et al.*, *Proc. ICASSP*, pp. 31–35, 2016.
[18] L. Li *et al.*, *Proc. ICASSP*, pp. 16–20, 2018.
[19] H. L. Van Trees, *Optimum array processing*, 2002.
[20] O. L. Frost, *Proceedings of the IEEE*, vol. 60, no. 8, pp. 926–935, 1972.
[21] O. Yılmaz *et al.*, *IEEE TSP*, vol. 52, no. 7, pp. 1830–1847, 2004.
[22] A. Kurematsu *et al.*, *Speech communication*, vol. 9, no. 4, pp. 357–363, 1990.
[23] E. A. P. Habets, Available at: https://www.audiolabs-erlangen.de/fau/professor/habets/software/rir-generator, 2008.
[24] E. Vincent *et al.*, *IEEE TASLP*, vol. 14, no. 4, pp. 1462–1469, 2006.