Multiple fixed beamformers with a spacial Wiener-form postfilter for far-field speech recognition

Sining Sun*, Shuran Zhou†, Mei-Yuh Hwang†, Lei Xie* , Qin Li†, Xin Lei†
* School of Computer Science, Northwestern Polytechnical University, Xi'an, China
† Mobvoi AI Lab, Seattle, USA

***Abstract*—Far-field speech recognition is becoming a hot topic in research and industrial applications. In this paper, in order to improve far-field speech recognition performance, we propose to use multiple fixed beamformers with a spacial Wiener-form postfilter (MFB-SWP) to suppress noise and interference. Our proposed method consists of two parts, beamforming and post-filter estimation. First, multiple fixed beamformers are designed and each of them aims at one specific direction. Next the target speech is estimated using the fixed beamformer aiming to the target direction, and the noise and interference signals are estimated using the remaining beamformers. After that, we calculate a spacial Wiener-form time-frequency and frame-level gains, as postfilter to further reduce the residual noise and interference. Compared with a single fixed beamformer, the proposed MFB-SWP method can suppress noise and interference significantly. It is also computationally more efficient comparing with other adaptive beamforming methods. Our experiments showed that proposed method achieved 16-50% relative character error rate (CER) reduction compared with using the single fixed beamformer only.**

***Index Terms*—far-field speech recognition, fixed beamformer, spacial Wiener postfilter, MFB-SWP**

I. Introduction

In recent years, thanks to the development of deep learning techniques, big data and powerful computation resources, the performance of automatic speech recognition (ASR) has been improving significantly. With the increasing popularity of far-field applications, far-field speech recognition has become a hot topic. However, it remains a tough problem due to the complexity of acoustic environments in far-filed scenarios. First, the energy of speech signals decays with the propagated distance from the sound source to the microphone, and thus the captured signals are weak the further they go. Secondly, room reverberation can severely degrade ASR performance, especially the late reverberation [1]. Thirdly, interference by surrounding speech, music and noise is very common in far-field scenarios. Often seen in real applications, all the three problems can occur simultaneously. Room reverberation and noise interference, compounded with the decay by distance, often result in low signal-to-noise ratio and ASR becomes highly sophisticated.

In order to tackle the foregoing problems, many algorithms have been proposed. Based on the number of recording microphones, these methods can be divided into two categories, single-channel methods and multi-channel methods. In this paper, we focus on the multi-channel case. Compared with the single channel case, multiple microphones can give much more spatial information. For example, in order to alleviate the effect of reverberation, a generalized multi-channel weighted prediction error (WPE) method [2] was proposed. When multiple microphones are available, beamforming techniques, such as minimum variance distortionless response (MVDR) and generalized sidelobe canceller (GSC) beamforming [3], can be used to perform signal enhancement and interference suppression. Recently, due to successful application to single channel speech enhancement [4], [5], deep learning is also applied to estimate adaptive beamforming parameters [6], [7].

In complex acoustic environments, current beamforming techniques have very limited capability in de-noising and suppressing interference. Usually beamforming is followed by a post-filter to reduce the residual noise [8]. The speech distortion weighted multi-channel Wiener filter (SDW-MWF) [9] is a popular postfilter, which can be decomposed into an MVDR beamfomer and a single-channel post-filter. In [10], Warsitz et al. proposed a generalized eigenvalue (GEV) decomposition based beamforming approach. Because the GEV beamformer is designed to maximize the output SNR, it may introduce speech distortion. In order to control the distortions, they proposed the blind analytical normalization (BAN) postfilter for compensation. They further proved that GEV with the BAN postfilter equals to MVDR. Some single channel speech enhancement methods can also be combined with beamforming techniques. For example, Cohen et al. [11], [12] combined the modified version of log-spectral amplitude estimator (OMLSA) [13] with a generalized sidelode canceller (GSC) structure.

Our previous work [14] combined multiple fixed beamform-

Lei Xie is the corresponding author.

ers with ROVER [15] to improve far-field speech recognition system performance. In this paper, we propose to use **m**ultiple **f**ixed **b**eamformers with a **s**pacial **W**iener-form **p**ostfilter (MFB-SWP) to suppress the noises and interference from non-target directions. First, multiple fixed beamformers are designed and each of them only steers to a specified direction. Once the target speech direction is known, the output of the beamformer which steers to this direction will be selected as the estimated target speech, and the remaining beamformers' output will be regarded as the estimated noise and interference. After we obtain these estimations, a Wiener-form time-frequency gain can be calculated as a postfilter. Compared with single fixed beamformer (FB), the proposed MFB-SWP method can suppress noises significantly. Compared with other adaptive beamforming methods such as MVDR and GSC, our MFB-SWP is more computationally efficient. We evaluate the proposed front-end enhancement method in a Mandarin speech recognition task. Instead of simulating far-field multi-channel speech, we recorded several test corpora with multiple microphones in real conditions. Our experimental results showed that the proposed MFB-SWP approach can reduce character error rates (CER) compared with the FB and GSC beamformers, up to 50% relative improvement.

II. The MFB-SWP method

As mentioned earlier, we use multiple fixed beamformers and each of them is steered to one specific direction. We follow the fixed beamformer proposed by [16]. In this section, we first introduce the basic idea of fixed beamformers, and then we elaborate our proposed method.

A. Fixed beamformers with the white noise gain constraint

One of the disadvantages of fixed beamformers is the poor white noise gain (WNG), especially in the low frequency range. In paper [16], the authors introduced the WNG constraint during the fixed beamformer design. For a microphone array with N sensors, the beamformer's coefficient at frequency ω is $\mathbf{w}(\omega) = [w_0(\omega), ..., w_{N-1}(\omega)]^T$. The response of the filter is given by Reference [3]:

$$B(\omega, \theta) = \sum_{n=0}^{N-1} w_n(\omega) e^{-j\omega\tau_n(\theta)} \tag{1}$$

where θ is the angle of arrival, τ_n is the relative delay time. The WNG is defined by

$$A(\omega) = \frac{|\mathbf{w}^T(\omega)\mathbf{d}(\omega)|^2}{|\mathbf{w}^H(\omega)\mathbf{w}(\omega)|} \tag{2}$$

where $\mathbf{d}^T(\omega) = [exp(-j\omega\tau_0\theta_d), ..., exp(-j\omega\tau_{N-1}\theta_d)]$ denotes the steering vector and θ_d is the true direction where the speech is from. Superscript H is the conjugate transpose operation.

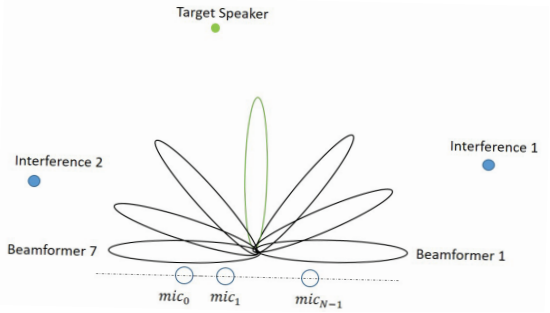

Fig. 1. The design of multiple fixed beamformers.

The idea behind the design is to optimally approximate a desired response, $B^*(\omega, \theta)$, by $B(\omega, \theta)$ in the least square sense with some constraints. Typically, a numerical solution is obtained by dividing the frequency range into F frequency bins $\omega_f, f = 0, ...F-1$, and the possible source direction into M angles $\theta_m, m = 0, ...M-1$. The WNG constraint puts a minimum value on $A(\omega) \geq \gamma > 0$, which enables the problem to be formulated as a convex problem. The MATLAB CVX Toolbox [17] is used in this paper. Design details can be found in [16].

B. MFB-SWP

The noisy signal captured by a microphone array is $\mathbf{y}(\omega, t) = [y_0(\omega, t), ..., y_{N-1}(\omega, t)]^T$ in the frequency domain. Once the beamformer's coefficient $\mathbf{w}(\omega)$ is designed, the enhanced signal can be calculated by

$$y(\omega, t) = \mathbf{w}^H(\omega)\mathbf{y}(\omega, t) \tag{3}$$

The beamformer described in Section II-A can suppress interference from the non-target direction to some extent, but the performance is still limited. There are still a lot of residual noises in the enhanced signal. Further noise reduction is necessary. We propose to use multiple fixed beamformers with each directed to one specific direction, as shown in Figure 1. In the figure, we show there is one speech signal from a target speaker at one position, and two interference signals from two different positions. Here we have a linear microphone array with N microphones and seven fixed beamformers are applied. The green microphone is steered to the target speaker and the others steered to other directions, including the directions of the interference. Note that, we can use several fixed beamformers to cover the whole space because we can control the beam width during beamformer's design.

To formulate the design, assume there are P fixed beamformers, $\mathbf{w}_1(\omega), \mathbf{w}_2(\omega), ..., \mathbf{w}_P(\omega)$. Further assume beamformer Q directs to the target direction, then the enhanced target signal is

$$\hat{y}(\omega, t) = \mathbf{w}_Q^H(\omega)\mathbf{y}(\omega, t) \tag{4}$$

Even the signal is enhanced by the target direction beamformer, there are still a lot of residual interference signals. In order to estimate the interference, the signals from other non-target directions will also be estimated by the remaining beamformers

$$\hat{n}(\omega,t) = \sum_{p=1,p\neq Q}^{P} \mathbf{w}_p^H(\omega)\mathbf{y}(\omega,t) \tag{5}$$

Then, a Wiener form gain in time-frequency domain can be estimated by

$$G_{tf}(\omega) = \frac{\hat{y}(\omega,t)}{\hat{y}(\omega,t)+\hat{n}(\omega,t)} = \frac{\hat{\xi}(\omega,t)}{\hat{\xi}(\omega,t)+1} \tag{6}$$

where

$$\hat{\xi}_{\omega,t} = \frac{\hat{y}(\omega,t)}{\hat{n}(\omega,t)} \tag{7}$$

is an estimated SNR. Because of the spacial leakage of the fixed beamformers, the estimated interference signals' energy is usually larger than the original one, so a discount factor is introduced here. Equation 7 shares a similar form as the parametric Wiener filter [18], except that the spacial information is used to estimate the interference signals. Therefore we call this postfilter a spacial Wiener postfilter (SWP). Note that when there is no interference, $\hat{n}(\omega,t)$ should be very small, and $G_{tf}(\omega)$ should be 1.0. Otherwise when there is a strong interference, $G_{tf}(\omega)$ should be small.

Furthermore, we calculate a time level gain by summing $G_{tf}(\omega)$ over all frequencies

$$G_t = \sum_{\omega} G_{tf}(\omega) \tag{8}$$

The final enhanced signal can be obtained by applying Equations 7 and 8 to Equation 4

$$s(\omega,t) = \hat{y}(\omega,t)G_{tf}(\omega)G_t \tag{9}$$

III. Experiments

In this paper, we evaluate the effectiveness of speech enhancement algorithms by speech recognition error rates. In particular, we carried our experiments on Mandarin speech recognition and hence character error rates (CER) are measured. We performed experiments on various conditions, including different noise conditions, reverberation and the geometries of microphone array arrangement.

A. Data collection

Instead of simulating multi-channel data, we recorded the test corpora using smart TVs and smart home speakers, with various noise and reverberation conditions in order to cover the real application scenarios. Note that in this paper, as we only focus on beamforming techniques, the direction of the target speaker is known and fixed. The test corpora consist of natural sentence queries, keywords, and some command and control, in Mandarin. Some examples are provided in Figure 3.

1) Recording devices: As shown in Figure 2, the recording device is Mobvoi's AI module with a microphone array development kit. The microphone array has two different geometries. The first one is a 4-element uniform linear array (ULA) with 4cm mic interval. The second one is a 6-element uniform circular array (UCA) with a customized diameter from 2cm to 12cm. In this paper we used an UCA with 4cm diameter, and we do not use the data collected by the two microphones at 0° and 180°.

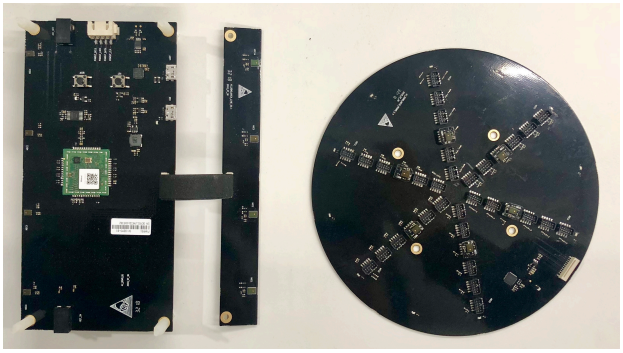

Fig. 2. Mobvoi's microphone array modules. The left one is the uniform linear array (ULA) and the right one is the uniform circle array (UCA).

2) Recording rooms, noise and interference: Two rooms with different reverberation were selected during our data collection. First, we want to verify the performance when there is no to little reverberation. This room is regarded as a virtually anechoic room. In this anechoic room, several volunteers read the transcripts from a fixed angle and fixed distance away from the microphone array, in total 330 utterances. Two background noises (speech and/or music) were replayed simultaneously by two high-fidelity (Hi-Fi) speakers at two different but fixed angles and distances away from the microphone array. Thus this test set is a noisy test set.

The second room is a normal sized (5m by 8m) living room with reverberation, which is the most common scenario that we care about. In the living room, 2700 clean utterances were recorded first, with a single close-distance microphone. Then a clean test set of 2700 utterances was created by replaying these utterances via a Hi-Fi speaker from a fixed angle and fixed distance away from the microphone array. Separately, the same 2700 utterances were replayed the second time with one background noise (music and speech) replayed simultaneously at a different but fixed angle and distance away from the microphone array. So for the living room scenario, there are two test sets: one clean and one noisy test set. Table I shows the details of our recording setups.

Chinese queries	English equivalent
播放权力的游戏 第 1 季第 1 集	Play Game Throne Season 1 Episode 1
我想看阿甘正传	I want to watch Forrest Gump
嗨，小问	Hi, Xiao Wen
你好，问问	Hello, Wen Wen

Fig. 3. Example queries of our test corpora.

TABLE I
DETAILS OF OUR RECORDING SETUPS.

Room	RT60 (ms)	Device	Noise	Num. of utterances	Speaker
Anechoic	∼50	4mic ULA 4mic UCA	Yes	330	Human
Living	∼250	4mic ULA	Yes &No	2700	Hi-Fi Speaker

B. Acoustic and language model

The acoustic model used in our work is an 11-layer TDNN, trained with the lattice-free MMI criterion [19] using Kaldi. The model was trained using 16,000 hours of near-field speech data and thus is not optimized for far-field ASR. The language model is a 4-gram word-based model with a vocabulary size of 220k.

C. Speech recognition results

Since our focus is on far-field speech recognition, we only evaluate our proposed method by reporting speech recognition error rates.

Our front-end speech enhancement pipeline includes multi-channel beamforming, followed by single channel noise reduction (NR) [13] to further reduce noise. We compared our proposed MFB-SWP approach with GSC beamformer [12], and also with fixed beamformer (FB) as mentioned in Section II-A. Another comparison is with blind source separation (BSS), where a multi-channel BSS method proposed by [20] was used.

We first conducted our experiments in a virtually anechoic room, in order to exclude the effect of reverberation. The target speaker was in a known and fixed direction, while interference speakers were in other directions. As shown in Table II, the proposed MFB-SWP method outperforms all other methods in all microphone array geometries. Note that, for the 2-mic ULA array, 2 microphones with 8cm interval were selected from the 4-mic ULA array. Because our data is very noisy and SNR is pretty low, the single-channel speech recognition performance is very bad. Comparing with FB+NR, our proposed method obtained 16.7%, 26.0% and 50.8% relative CER reduction in the 2-mic ULA, 4-mic ULA and 4-mic UCA respectively.

The next set of experiments was conducted in a normal-sized living room, with a certain level of reverberation. As shown in Table III, on the noisy test set, our proposed MFB-SWP method outperformed all other methods. Particularly it achieved a relative 5% CER reduction compared with FB+NR, or 16.9% compared to NR only. On the clean test set, although it was worse than GSC+NR, it still reduced CER relatively by 13.7% compared to NR only.

From Table II and Table III, we see our algorithm is more effective in eliminating noises than de-reverberation. Therefore our future work would be to combine de-reverberation with beamforming to further improve ASR.

TABLE II
CER IN A VIRTUALLY ANECHOIC ROOM, WHERE THERE IS NO REVERBERATION, BUT NOISE AND INTERFERENCE EXIST. NR=SINGLE CHANNEL NOISE REDUCTION.

Setup	Anechoic Room		
	2 Mic ULA	4 Mic ULA	4Mic UCA
NR	80.71%	77.90%	73.97%
BSS + NR	71.36%	66.42%	62.87%
GSC + NR	70.23%	50.97%	54.72%
FB + NR	71.15%	49.27%	69.49%
MFB-SWP + NR	**59.29%**	**36.44%**	**34.22%**

TABLE III
CER OF CLEAN AND NOISY TEST DATA IN A NORMAL SIZED LIVING ROOM, WHERE REVERBERATION EXISTS.

Setup	Living Room	
	4 Mic ULA	
	Clean	Noisy
NR	27.35%	76.07%
BSS + NR	29.16%	80.54%
GSC + NR	**20.25%**	65.52%
FB + NR	37.93%	66.44%
MFB-SWP + NR	23.61%	**63.19%**

IV. CONCLUSIONS

In this paper, we propose a novel and highly efficient beamformer with a postfilter, named MFB-SWP, that can be used with different microphone geometries, and on resource-limited embedded devices. We demonstrate its effectiveness in real user scenarios. One significant advantage of our algorithm is that it requires fewer computing resources, comparing to other adaptive beamforming methods. In this case, filters can be generated with offline tools and deployed directly to embedded devices as weight parameters. On the other hand, the direction of the target speech is assumed to be known in all experiments reported here. Our future work will involve combining the proposed method with direction of arrival (DOA) estimation and de-reverberation algorithms to perfect our microphone array modules.

V. ACKNOWLEDGEMENT

This research work is supported by the National Natural Science Foundation of China (No.61571363).

REFERENCES

[1] Keisuke Kinoshita, Marc Delcroix, Takuya Yoshioka, Tomohiro Nakatani, Armin Sehr, Walter Kellermann, and Roland Maas, "The reverb challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," in *Applications of Signal Processing To Audio and Acoustics*, 2013, pp. 1–4.

[2] Takuya Yoshioka and Tomohiro Nakatani, "Generalization of multi-channel linear prediction methods for blind mimo impulse response shortening," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 10, pp. 2707–2720, 2012.

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

[4] Yong Xu, Jun Du, Li-Rong Dai, and Chin-Hui Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal processing letters*, vol. 21, no. 1, pp. 65–68, 2014.

[5] DeLiang Wang and Jitong Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2018.

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

[7] Xiong Xiao, Chenglin Xu, Zhaofeng Zhang, Shengkui Zhao, Sining Sun, Shinji Watanabe, Longbiao Wang, Lei Xie, Douglas L Jones, Eng Siong Chng, et al., "A study of learning based beamforming methods for speech recognition," .

[8] Sharon Gannot, Emmanuel Vincent, Shmulik Markovich-Golan, Alexey Ozerov, Sharon Gannot, Emmanuel Vincent, Shmulik Markovich-Golan, and Alexey Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 25, no. 4, pp. 692–730, 2017.

[9] Simon Doclo, Ann Spriet, Jan Wouters, and Marc Moonen, *Speech Distortion Weighted Multichannel Wiener Filtering Techniques for Noise Reduction*, 2005.

[10] Ernst Warsitz and Reinhold Haeb-Umbach, "Blind acoustic beamforming based on generalized eigenvalue decomposition," *IEEE Transactions on Audio Speech & Language Processing*, vol. 15, no. 5, pp. 1529–1539, 2007.

[11] Israel Cohen, Sharon Gannot, and Baruch Berdugo, "An integrated real-time beamforming and postfiltering system for nonstationary noise environments," *Eurasip Journal on Advances in Signal Processing*, vol. 2003, no. 11, pp. 1–10, 2003.

[12] S. Gannot and I. Cohen, "Speech enhancement based on the general transfer function gsc and postfiltering," *IEEE Transactions on Speech & Audio Processing*, vol. 12, no. 6, pp. 561–571, 2004.

[13] Israel Cohen and Baruch Berdugo, "Speech enhancement for non-stationary noise environments," *Signal Processing*, vol. 81, no. 11, pp. 2403–2418, 2001.

[14] Sining Sun, Yangyang Shi, Ching-feng Yeh, Suliang Bu, and Computer Science, "Multiple Beamformers with ROVER for the CHiME-5 Challenge," .

[15] J. G Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (rover)," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 1997, pp. 347–354.

[16] Edwin Mabande, Adrian Schad, and Walter Kellermann, "Design of robust superdirective beamformers as a convex optimization problem," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009, pp. 77–80.

[17] Michael Grant and Stephen Boyd, "CVX: Matlab software for disciplined convex programming, version 2.1," Mar. 2014.

[18] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586–1604, Dec 1979.

[19] Daniel Povey, Vijayaditya Peddinti, Daniel Galvez, Pegah Ghahremani, Vimal Manohar, Xingyu Na, Yiming Wang, and Sanjeev Khudanpur, "Purely sequence-trained neural networks for asr based on lattice-free mmi.," in *Interspeech*, 2016, pp. 2751–2755.

[20] Daichi Kitamura, Nobutaka Ono, Hiroshi Sawada, Hirokazu Kameoka, and Hiroshi Saruwatari, "Determined blind source separation with independent low-rank matrix analysis," in *Audio Source Separation*, pp. 125–155. Springer, 2018.