International Symposium on Chinese Spoken Language Processing (ISCSLP 2008)
Kunming, China
December 16-19, 2008

MICROPHONE ARRAY POST-FILTER BASED ON AUDITORY FILTERING

Peng Li [1], Fengchai Liao [2], Ning Cheng [3], Bo Xu [1,3], Wenju Liu [3]
[1] Digital Content Technique Research Center, Institute of Automation, Chinese Academy of Sciences
[2] Department of Math and Computer Science, Sanming University
[3] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences

ABSTRACT

In this paper, an auditory filtering based microphone array post-filter is proposed to enhance the quality of the output signal. By using a gammatone filterbank to band pass each input of the array, the input signals are decomposed into a two-dimensional T-F representation. Then, for each auditory filter channel, the post-filter's coefficients are estimated in each frame using the decomposed multi-channel input signals. Followed by the post-filtering and synthesis processing, the enhanced speech with better quality is acquired. Systematical evaluations on the CMU microphone array database prove that the proposed method could improve not only the noise reduction measure but also the speech quality measures.

Index Terms—Microphone array, speech enhancement, auditory filter

1. INTRODUCTION

Microphone array permits distant, hands-free signal acquisition. It has the ability to suppress interfering signals coming from undesired directions by providing spatial filtering to the sound field. The researches on microphone arrays would greatly facilitate many applications, such as speech enhancement, source localization and so on.

In recent years, many microphone array based speech enhancement techniques have been proposed [1]-[4]. Since the desired speech signal has an extremely wide bandwidth relative to its center frequency, conventional narrowband techniques can not be applied directly to microphone array signal processing. Thus, short time Fourier transform (STFT) is employed in most of the microphone array speech enhancement methods and the objective of the processing is always concentrated on improving the signal to noise ratio (SNR) at the output. However, researches on the human perception have proved that the responses of human auditory system to different frequencies do not have the same characteristics with STFT. Moreover, a higher SNR does not mean a higher speech quality. Considering the facts described above, in this paper, we attempt to introduce auditory filter, one of the main achievements of auditory perception researches, into the microphone array speech enhancement to further improve the speech quality of the output signal. For convenience, we only focus on the post-filter technique, which is one of the most typical methods in microphone array speech enhancement.

The organization of this paper is as follows: Section 2 gives a brief review of the theoretical basis of microphone array post-filtering. Section 3 explains the proposed speech enhancement method based on auditory filtering in detail. In Section 4, the performance of the proposed method is systematically evaluated with CMU multi-channel noisy office recordings. Finally, a conclusion is given in Section 5.

2. THEORETICAL FRAMEWORK OF MICROPHONE ARRAY POST-FILTER

In microphone array processing, the received multi-channel inputs are modeled as the desired signal filtered by the acoustic path to each microphone, plus an additive noise component on each channel. That is (omitting the frequency dependence for clarity)

$$\mathbf{x}' = s\mathbf{d} + \mathbf{n}' \tag{1}$$

where s is the desired signal, $\mathbf{d}$ is the propagation vector of the signal source

$$\mathbf{d} = [d_1\ d_2 \ldots \mathrm{d_N}]^T \tag{2}$$

and $\mathbf{n}'$ is similarly the vector of additive noise signals

$$\mathbf{n}' = [n'_1\ n'_2 \ldots \mathrm{n'_N}]^T \tag{3}$$

where N is the number of sensors in the array.

It has been demonstrated that, with this model, the optimal broadband Minimum Mean Square Error (MMSE) filter solution (that is, the multi-channel Wiener filter) can be factorized into a classical Minimum Variance Distortionless Response (MVDR) beamformer followed by a single-channel Wiener post-filter [2], that is

$$\mathbf{W}_{opt} = \left[\frac{\phi_{ss}}{\phi_{ss} + \phi_{nn}}\right] \frac{\Phi_{nn}^{-1}\mathbf{d}}{\mathbf{d}^H \Phi_{nn}^{-1}\mathbf{d}} \tag{4}$$

where $\mathbf{W}_{opt}$ is the vector of optimal filter coefficients, ϕ_{ss} and ϕ_{nn} are the (single-channel) signal and noise auto-spectral density vectors respectively, and Φ_{nn}^{-1} is the (multi-channel) noise cross-spectral matrix. The bracketed factor in eq.(4) corresponds to a single-channel Wiener filter, while the remaining corresponds to a MVDR beamformer [2].

Based on the above equation, an optimal array processing structure could be constructed. The transfer function of the single-channel Wiener post-filter is typically estimated from the aligned multi-channel inputs. The MVDR beamformer first maximizes the directivity of the array response, and then the post-filter further enhances the output broadband SNR. Therefore, the key problem is how to estimate the bracketed factor in eq.(4).

3. MICROPHONE ARRAY POST-FILTER BASED ON AUDITORY FILTERING

3.1. Overview

Figure 1 shows the structure of the proposed auditory filtering based microphone array post-filter. It is constructed according to the theoretical basis of the post-filter. From Figure 1 we can easily find that the proposed structure is similar with the Zelinsiki post-filter [1].

The proposed structure consists of five stages. First, the multi-channel input signals pass through the time alignment module to account for the effect of the propagation. Second, the aligned signals are band pass filtered by the auditory filterbanks and decomposed into two-dimensional T-F representations. Third, the MVDR bemforming is executed in each auditory filter channel. Then, the post-filter is applied to the MVDR output by using the coefficients estimate from the decomposed signals. At last, the output is synthesized with the method proposed by Weintraub [5]. Detailed explanations will be provided below.

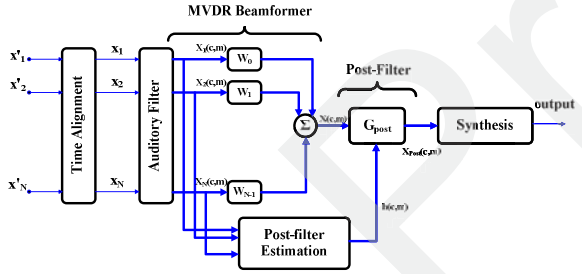

Figure 1: Structure of the proposed post-filter

3.2. Time alignment

According to Figure 1, the input signals are first scaled and aligned to account for the effect of the propagation vector **d** at the time alignment module. Thus, the signals at the output of the time alignment can be formulated as

$$\mathbf{x} = s + \mathbf{n} \tag{5}$$

where **n** is the noise signal vector after time alignment for the desired signal.

3.3. Auditory filter

Different from the general microphone array based speech enhancement method, in this paper, an auditory filterbank is used to pre-processing every input signal instead of STFT. The auditory filterbank here used is the gammatone filterbank [6], which is a standard model of cochlear filtering simulating the function of basilar membrane. The impulse response of a gammatone filter is:

$$g(t) = \begin{cases} t^{l-1}\exp(-2\pi bt)\cos(2\pi ft), & t \geq 0 \\ 0, & else \end{cases} \tag{6}$$

where $l = 4$ is the order, b is the equivalent rectangular bandwidth, and f is the center frequency of the filter.

In implementation, a 128-channel gammatone filterbank whose center frequencies are quasi-logarithmically spaced from 80 to 7000 Hz is adopted to filter the input signal (whose sampling frequency is 16 kHz) of each sensor. Then, the output of each filter channel is divided into frames of 400 samples (25ms) with overlap of 300 samples (≈19ms) between consecutive frames.

Note that, by above processing, the input signals are decomposed into two-dimensional T-F representations. It is different from the STFT processing. First, in each filter channel, the output is a time domain signal having the same length with the original input, which offers the opportunity to execute the MVDR beamforming and estimate the post-filter's coefficients in the time-domain instead of the frequency domain. Second, the output in each filter channel could be approximately treated as a "narrow-band" signal (especially in the low-frequency channel, higher frequency resolution can be obtained), which makes it possible to directly introduce the narrow-band signal processing methods to speech signal processing. In this paper, we only concentrate on the former one. The value of auditory filtering on introducing the narrow-band signal processing methods to wideband signal will be studied in the future.

3.4. MVDR beamformer

In section 3.3 we have described that, the output of each filter channel is a time-domain signal having the same length with the original input. Therefore, it is easy to fulfill the MVDR beamformer in each filter channel in the time-domain directly, which could be formulated as (omitting the time dependence for clarity):

$$\mathbf{x}(c,m) = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i(c,m)\ ,\quad c \in \{0,1,\cdots 127\} \tag{7}$$

where c is the channel number of the auditory filter, m is the frame number, $\mathrm{x}(c,m)$ is the output of the MVDR beamformer in the mth frame, cth gammatone filter; and $\mathrm{x}_i(c,m)$ is the aligned signal of the ith sensor filtered by the cth gammatone filter.

3.5. Post-filter

3.5.1. Estimation of the Post- filter coefficients

For a certain frame m, the auto- and cross-correlation of the aligned signals on sensors i and j in the cth filter channel can be calculated as

$$R_{x_i x_j}(0;c,m) = R_{ss}(0;c,m) + R_{n_i n_j}(0;c,m) + 2R_{sn_i}(0;c,m) \tag{8}$$

and $R_{x_i x_j}(0;c,m) = R_{ss}(0;c,m) + R_{n_i n_j}(0;c,m) + R_{sn_j}(0;c,m) + R_{n_i s}(0;c,m)$ (9)

where $R_{yz}(0;c,m) = E\{y_{c,m}(t) z_{c,m}(t)\}$.

Based on the assumptions that

1) both the signal and the noise have zero averages ($E\{s\}=0$, $E\{n_i\}=0, \forall i$);
2) the signal and noise are uncorrelated ($R_{n_i s}(0;c)=0, \forall i$);
3) the noise auto-correlation is the same on all sensors ($R_{n_i n_i}(0;c,m) = R_{nn}(0;c,m), \forall i$);
4) the noise between sensors is uncorrelated ($R_{n_i n_j}(0;c,m)=0, \ \forall i \neq j$).

eq. (8) and (9) can be reduced to

$$R_{x_i x_i}(0;c,m) = R_{ss}(0;c,m) + R_{nn}(0;c,m) \quad (10)$$

$$R_{x_i x_j}(0;c,m) = R_{ss}(0;c,m) \quad (11)$$

Here, $R_{x_i x_i}(0;c,m)$ and $R_{x_i x_j}(0;c,m)$ are the auto- and cross-correlations of the time-aligned inputs X. They can be computed using the short-time estimation method [7]:

$$\hat{R}_{x_i x_j}(0;c,m) = \alpha \hat{R}'_{x_i x_j}(0;c,m) + (1-\alpha) R_{x_i x_j}(0;c,m) \quad (12)$$

where $\hat{R}'_{x_i x_j}(0;c,m)$ and $\hat{R}_{x_i x_j}(0;c,m)$ are the estimates for the previous and current frames respectively. The term α is a number close to unity. With this recursive form, smoother and more accurate estimates can be obtained.

It is obvious that the numerator and denominator of the Wiener filter transfer function in eq.(4) can be estimated from the cross- and auto-spectral densities of the input channels, respectively. Since for a signal whose average is zero, the auto- and cross-correlation without lag is equal to the integral of the auto- and cross-spectral densities, respectively. Therefore the coefficients of the post-filter can be estimated as:

$$h(c,m) = \frac{R_{ss}(0;c,m)}{R_{ss}(0;c,m) + R_{nn}(0;c,m)} \quad (13)$$

By averaging the correlations over all possible sensor combinations, this estimate can be made more robust. According to this, the coefficients of the post-filter can be estimate as

$$\hat{h}(c,m) = \frac{\frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \hat{R}_{x_i x_j}(0;c,m)}{\frac{1}{N} \sum_{i=1}^{N-1} \hat{R}_{x_i x_i}(0;c,m)} \quad (14)$$

In practice, the noise field does not always satisfy the assumption that the noise between sensors is uncorrelated. In addition, the estimate can contain negative values over a wide frequency range. Applying this estimated filter to the output signal of the conventional beamformer may result in a severe distorted output signal. To solve this problem, we take the modulus of the spatial cross power densities:

$$\hat{h}(c,m) = \frac{\left| \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \hat{R}_{x_i x_j}(0;c,m) \right|}{\frac{1}{N} \sum_{i=1}^{N-1} \hat{R}_{x_i x_i}(0;c,m)} \quad (15)$$

3.5.2. Wiener post- filtering

Having estimated the filter coefficient for a certain frame *m* in channel *c* as eq.(15), the Wiener post-filtering could be implemented by multiplying every sample in that frame by the estimated coefficient, which can be formulated as:

$$\mathbf{x}_{\text{post}}(c,m) = \hat{h}(c,m) \cdot \mathbf{x}(c,m) \quad c \in \{0,1,\cdots 127\} \quad (16)$$

3.6. Synthesis

After the post-filter is finished, a speech waveform could be synthesized according to the method proposed by Weintraub. More details about the synthesis can be referred to [5].

4. EVALUATIONS

In this section, the effectiveness of the proposed post-filter method is evaluated on a standard microphone array database. Comparisons with some other multi-channel noise reduction techniques, including the MVDR beamformer, the Zelinski post-filter and the MVDR beamformer in the proposed method are also provided.

4.1. Evaluation corpus

The standard corpus used for the evaluations is the CMU microphone array database [8]. The corpus consists of 130 utterances, 10 speakers of 13 utterances each. All the recordings were collected by a linear microphone array with 8 sensors spaced 7 cm apart, at a sampling rate of 16 kHz. The array was placed on a desk and the subject sat directly in front of the array at a distance of 1 meter from the center. For each array recording, a close-talking control signal corresponding to clean speech is provided.

4.2. Evaluation measures

In order to compare the proposed approach with the other multi-channel reduction methods, four different objective speech quality measures are utilized. The segmental SNR enhancement (SSNRE), which is the dB difference between the segmental SNRs of the enhanced output and the noisy inputs average, is utilized to evaluate the noise reduction, while the perceptual evaluation of speech quality (PESQ), the log-area-ratio distance (LAR) and the log-spectral distance (LSD), which are found to have a high correlation with the human perception, are adopted to assess the speech quality of the enhanced output signal [4]. Note that high values of the SSNRE and PESQ denote high speech quality, while low values of the LAR and LSD denote high quality.

4.3. Evaluation results

Table 1 lists the SSNRE, PESQ, LAR and LSD results averaged across the entire database for all the studied enhancement algorithms and the noisy input at sensor 1 of the microphone array. From table 1, it can be seen that the

proposed speech enhancement methods outperform the traditional multi-channel methods since they consistently produce better results for both the noise reduction measures and the perceptual quality measures in the given database.

Table1. Speech quality results of CMU database
*(**Noisy:** Noisy signal at sensor 1;*
***MVDR:** MVDR output; **PF:** Zelinski post-filter output.*
***P_MVDR:**Proposed MVDR output. **P_PF:**Proposed post-filter output.)*

	SSNRE	PESQ	LAR	LSD
Nosiy	-	2.39	9.73	3.64
MVDR	1.02	2.39	10.05	2.46
P_MVDR	3.34	2.65	8.10	1.71
PF	1.33	2.62	14.01	2.32
P_PF	3.76	2.65	11.85	1.60

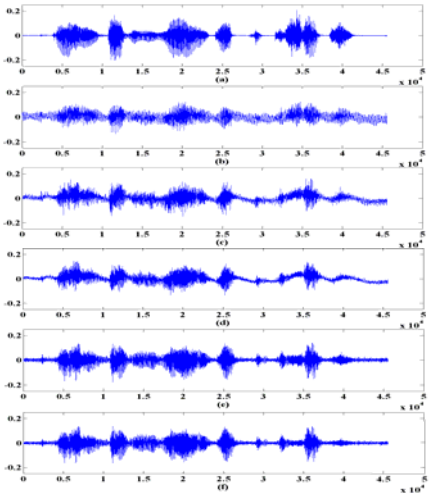

Figure 2: *Spectrograms for an utterance ''r-e-w-y-8-56''.*
(a) Original clean speech. (b) Noisy signal at sensor 1.
(c) MVDR output. (d) Zelinski post-filter output.
(e) Proposed MVDR output. (f)Proposed post-filter output.

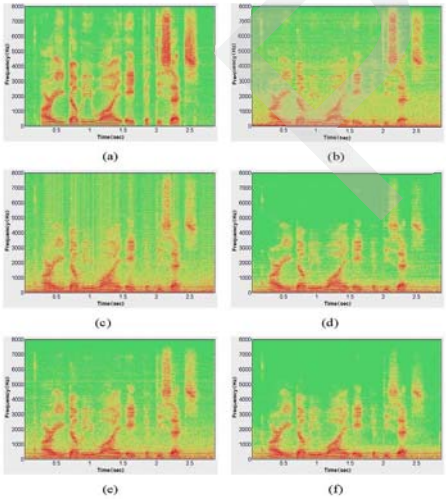

Figure 3: *Spectrograms for an utterance ''r-e-w-y-8-56''.*
(a) Original clean speech. (b) Noisy signal at sensor 1.
(c) MVDR output. (d) Zelinski post-filter output.
(e) Proposed MVDR output. (f)Proposed post-filter output.

In addition, we also present typical waveforms and spectrograms in Figure 2 and 3 respectively, for comparison between the clean signal, the noisy input at sensor 1 and the output signals of the studied multi-channel methods. From the figures, it can also be seen that the closest to the clean speech is that derived by the proposed approach.

5. CONCLUSIONS

In this paper, an auditory filtering based microphone array post-filter is proposed. Systematical evaluation results proved that the proposed microphone array post-filter method can achieve better speech quality then conventional methods. Moreover, the introducing of the processing of auditory filtering also makes it possible to employ the narrow-band array signal processing methods to solve the wideband speech problem, which shows a promising application prospect.

6. ACKNOWLEDGEMENTS

This work is supported by the National Grand Fundamental Research 973 Program of China under Grant No. 2004CB318105 and the National High-Tech Research and Development Plan of China under Grant No. 2006AA010103 and 2006AA01Z19.

7. REFERENCES

[1] R. Zelinski, "A microphone array with adaptive post-filtering for noise reduction in reverberant rooms", in Proc. of ICASSP'88, 1988, Vol. 5, pp. 2578-2581.

[2] K. Uwe Simmer, et al, "Post-filtering techniques", In: Brandstein, M., Ward, D. (Eds.), Microphone Arrays: Signal Processing Techniques and Applications. New York: Springer, Verlag, Chapter 3, pp. 36–60, 2001.

[3] Iain A. McCowan, Hervé Bourlard, "Microphone array post-filter based on noise field coherence", IEEE Trans. Speech Audio Process., Vol.11, pp.709-715, Nov. 2003.

[4] S. Lefkimmiatis and P. Maragos, "A generalized estimation approach for linear and nonlinear microphone array post-filters", Speech Comm., 49, 657-666, 2007.

[5] M. Weintraub, "A theory and computational model of auditory monaural sound separation," Ph.D. dissertation, Dept. Elect. Eng., Stanford Univ., Stanford, CA, 1985.

[6] Patterson, R. D., Nimmo-Smith, I., Holdsworth, J., Rice, P., 1988. An efficient auditory filterbank based on the gammatone function, Appl. Psychol. Unit, Cambridge Univ., Cambridge, U.K., APU Rep. 2341.

[7]J. B. Allen, D. A. Berkley, and J. Blauert, "Multimicrophone signal-processing technique to remove room reberberation from speech signals," J. Acoust. Soc. Amer., vol. 62, no. 4, pp. 912–915, Oct. 1977.

[8]Sullivan, T., 1996. CMU microphone array database. http://www.speech.cs.cmu.edu/databases/micarray.