A Dual-Channel Beamformer Based on Time-delay Compensation Estimator and Shifted PCA for Speech Enhancement

Jie Zhang
Open Lab on Human-Robot Interaction
School of Electrical and Computer Engineering
Shenzhen Graduate School
Peking University, Shenzhen, 518055
Email: zhangjie827@sz.pku.edu.cn

Hong Liu
Open Lab on Human-Robot Interaction
School of Electrical and Computer Engineering
Shenzhen Graduate School
Peking University, Beijing, China, 100871
Email: hongliu@pku.edu.cn

***Abstract*—Speech enhancement is an essential technique to process degraded audio in various applications. Beamforming to eliminate interferences based on sensor arrays is the most well-known method for this issue. However, traditional beamformers often face magnitude incoherence towards received signals due to directional weighting. Therefore, a novel dual-channel beamformer based on time-delay compensation (TDC) and shifted principal components analysis (PCA) is presented in this work. Firstly, our enhancement algorithm utilizes TDC estimator to preserve binaural cues, including interaural time-delay and intensity difference. Then the estimated cues are comprised to improve the shifted PCA, which can reduce noise by extracting primary components. Finally, the aforehand processed audio are input to a beamformer with post-filter to obtain enhanced speech. Experiments have demonstrated that the proposed algorithm could achieve some superiorities in speech intelligibility compared with the state-of-the-arts against real scenarios.**

I. Introduction

Speech enhancement to improve speech intelligibility and quality is a popular research area recently by overcoming the presence of interferences, which has been widely comprised in various applications, e.g., speech communications, recognition, hearing aids and teleconferencing. It can be classified into single-channel and multi-channel scenarios. The potential of single-channel based enhancement algorithms is limited for they only use spectral information [1], yet multi-channel based algorithms typically incorporate both spatial and spectral information. Multi-channel beamforming therefore has attracted large interests in last decades, which can directionally eliminate noise to enhance the source signals [2].

Fixed beamformers, including delay-and-sum and superdirective beamformer [3], are designed to concentrate on the source speech by combining the delayed and weighted versions of the received speech on each sensors. Adaptive beamformers, e.g., generalized sidelobe canceller (GSC) and minimum variance distortionless response (MVDR), have been investigated. They also employ the spectral properties of captured speech by the array to further reject undesired signals from other orientations compared with the fixed one. Frost has proposed the classical constrained minimum power adaptive beamforming for array enhancement [4]. Griffiths and Jim suggested the GSC algorithm by taking a blocking matrix to produce noise reference signals to improve the previous beamformer [5]. Zelinski added an additional Wiener post-filter to further enhance speech [6]. The noise coherence of diffuse field could be used to improve the generalization of Zelinski's post-filter [7]. Yousefian and Louizou discussed the coherence function between the target and noise signals as a criterion for noise reduction [8]. Azarpour *et al*. presented a binaural noise reduction system based on adaptive matched filter and post-filtering [9].

However, the inputs of traditional beamformers would face magnitude incoherence, because they account for the spatial diversity of desired speech and noise sources by combining multiple noisy input signals only after tap delay, i.e., time alignment. In dual-channel speech enhancement, the signal on the left (right) microphone would achieve more information than the other one when source is located in the left (right) plane. Although there are some works trying to manage this incoherence, e.g., applying independent component analysis (ICA) [10], they cannot be applied into reverberant scenarios. For these, we exploit shifted principal components analysis (PCA) to overcome the incoherence based on interaural-time delay (ITD) and interaural intensity difference (IID), which are yielded by time-delay compensation (TDC) estimator concurrently. Besides, a frequency-domain filter is introduced to extend the shifted PCA to adapt the reverberant environments. Therefore, our enhancement system consists of three modules, which is depicted in Fig.1.

The rest of this paper is organized as follows: TDC is briefly introduced in Sect. II. The modified shifted PCA version is presented in Sect. III. Sect. IV gives the beamforming with post-filter. Then experiments are available in Sect. V. At last, conclusions are drawn in Sect. VI.

II. Time-Delay Compensation Estimator

This section will address TDC estimator for binaural cues preservation in the time-frequency domain. The concept of TDC was foremost proposed by us in [11] [12] [13]. Let $s(n)$ denote the source speech, and the dual-channel received signals as $x_i(n), i \in \{l, r\}$, respectively. Assuming that binaural signals are counterparts of sound source with time-delay and

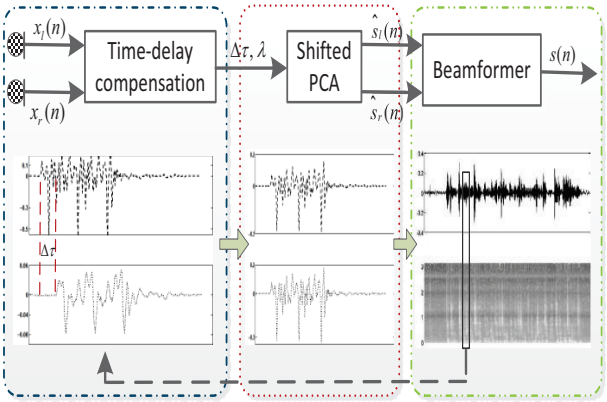

Fig. 1. A brief illustration of this dual-channel speech enhancement system. The time-delay compensation and shifted PCA are two preprocessors to prepare dual-channel signals for the beamformer.

attenuation so as to simplify analysis, it can be attained as

$$x_i(n) = a_i s(n-\tau_i) + v_i(n), i \in \{l, r\}, \tag{1}$$

where a_i denote the attenuation factors, τ_i are time factors from the sound source to the two acoustic sensors, $v_i(n)$ are the interferences, respectively. Define ITD $\Delta\tau$ as

$$\Delta\tau = \tau_r - \tau_l. \tag{2}$$

Before speech enhancement, we should conduct some preprocesses like enframing and windowing, and Hanning window is used here. Therefore, the relationship between binaural signals using TDC can be given by

$$W \odot x_l(n-\Delta\tau) = \lambda W \odot x_r(n) + \Delta v, \tag{3}$$

where W, λ and Δv denote the window function, attenuation difference and the disparity of received noises, respectively, so $\odot$ represents element-wise multiplication. In fact, Δv is also the error of TDC, and our goal is to make binaural signals similar as much as possible. From the standpoint of noises, Eq.(3) can be rewritten as

$$\Delta v = W \odot x_l(n-\Delta\tau) - \lambda W \odot x_r(n).$$

In most cases, binaural signals are preprocessed by normalization, thus Δv is thought as zero-mean Gaussian noise. Hereby the variance of Δv can be defined as

$$y = ||W \odot x_l(n-\Delta\tau) - \lambda W \odot x_r(n)||^2. \tag{4}$$

Therefore, the parameters λ and $\Delta\tau$ can be calculated by maximum likelihood estimation as follows

$$\frac{\partial y}{\partial \lambda} = \frac{\partial}{\partial \lambda}||W \odot x_l(n-\Delta\tau) - \lambda W \odot x_r(n)||^2.$$

Set this partial derivative to zero and λ, namely IID, can be easily solved as

$$\widetilde{\lambda} = \frac{\sum_N W^2(n) x_r(n) x_l(n-\Delta\tau)}{\sum_N W^2(n) x_r^2(n)}, \tag{5}$$

where N denotes the length of window. In terms of time-delay $\Delta\tau$, it is difficult to compute λ from $\partial y/\partial\Delta\tau$ directly, but transformed into the frequency domain instead, and Eq.(4) will be replaced by

$$Y(e^{j\omega}) = ||\boldsymbol{X}_l(e^{j\omega})e^{-j\omega\Delta\tau} - \lambda \boldsymbol{X}_r(e^{j\omega})||^2, \tag{6}$$

where $Y(e^{j\omega})$ and $\boldsymbol{X}(e^{j\omega})$ are the Fourier transforms of variance and binaural signals processed by window function, respectively, i.e. $\mathscr{F}\{W \odot x_r(n)\} = \boldsymbol{X}_r(e^{j\omega})$, $\mathscr{F}\{W \odot x_l(n-\Delta\tau)\} = \boldsymbol{X}_l(e^{j\omega})e^{-j\omega\Delta\tau}$. Therefore, if

$$\boldsymbol{A}(e^{j\omega}) = \boldsymbol{X}_l(e^{j\omega})e^{-j\omega\Delta\tau} - \lambda \boldsymbol{X}_r(e^{j\omega}),$$

then $\partial Y(e^{j\omega})/\partial\Delta\tau$ can be formulated as

$$\begin{aligned}\frac{\partial Y(e^{j\omega})}{\partial\Delta\tau} &= \frac{\partial}{\partial\Delta\tau}\left(\boldsymbol{A}^*(e^{j\omega})\boldsymbol{A}(e^{j\omega})\right) \\ &= \frac{\partial \boldsymbol{A}(e^{j\omega})}{\partial\Delta\tau} \cdot \frac{\partial Y(e^{j\omega})}{\partial \boldsymbol{A}(e^{j\omega})} \\ &= -j2\omega \boldsymbol{X}_l^*(e^{j\omega})\boldsymbol{A}(e^{j\omega})e^{-j\omega\Delta\tau}.\end{aligned} \tag{7}$$

Let $\partial Y(e^{j\omega})/\partial\Delta\tau$ be zero, for $j\omega$ and $e^{-j\omega\Delta\tau}$ are not equal to zero, it will be obtained

$$\boldsymbol{X}_l^*(e^{j\omega})\left(\boldsymbol{X}_l(e^{j\omega})e^{-j\omega\Delta\tau} - \lambda \boldsymbol{X}_r(e^{j\omega})\right) = 0, \tag{8}$$

where $*$ indicates the complex conjugate. Then taking Eq.(8) back to the time domain using Inverse Discrete Fourier Transform (IDFT), it can be shown as

$$\begin{aligned}\delta(n-\Delta\tau) &= R(n) \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{\lambda \boldsymbol{X}_l^*(e^{j\omega})\boldsymbol{X}_r(e^{j\omega})}{\boldsymbol{X}_l^*(e^{j\omega})\boldsymbol{X}_l(e^{j\omega})} \cdot e^{j\omega n} d\omega,\end{aligned} \tag{9}$$

where $R(n)$ is another version of GCC function. Thereout, $\Delta\tau$ can be estimated as

$$\widetilde{\Delta\tau} = arg\max_n R(n). \tag{10}$$

As a consequence, $\widetilde{\Delta\tau}$ is the optimal time-delay based on Minimum Mean Square Error (MMSE) criterion. From Eq. 9, we observe that this time-delay estimate is equivalent to the Roth weighting for GCC [14] [15], yet the two approaches begin with different horizons, and we can reduce the fluctuation of time-delays [13].

III. Shifted PCA

PCA is a widely-used method in PAE, and the channel-based PAE plays an important role in spatial audio analysis-synthesis. Given that both primary and ambient components are directional and diffuse, PAE aims to separate the primary components (i.e. the signals mainly containing source speech) from the ambient components based on perceptual spatial features, which can be characterized by the aforementioned inter-channel time-delay and inter-channel intensity difference. It can extract the desired source from the directional interferences and provide for the playback systems [16]. According to PAE, let $\boldsymbol{s}_l, \boldsymbol{s}_r$ and $\boldsymbol{v}_l, \boldsymbol{v}_r$ be the primary and ambient components in the two channels, respectively. We assume that

$$\boldsymbol{s}_l = \lambda \boldsymbol{s}_r, \boldsymbol{s}_i \perp \boldsymbol{v}_j, \boldsymbol{v}_l \perp \boldsymbol{v}_r, \forall i, j \in \{l, r\},$$

where $\perp$ represents that two signals are unrelated. In general, the primary and ambient components in the binaural signals are correlated and uncorrelated, respectively. Since both the correlated primary components are the counterparts of source speech, we can use the magnitude panned $\boldsymbol{s}_l = \lambda \boldsymbol{s}_r$ for analysis, and λ is the primary panning factor, i.e. IID estimated by $\widetilde{\lambda}$. PCA is involved in PAE by decomposing the covariance matrix of the input audio into its eigenvectors and eigenvalues. He *et*

al. presented a shifted PCA [17], where the extracted primary components are evaluated by

$$\begin{aligned}\hat{s}_l(n) &= \frac{1}{1+\lambda^2}\Big(x_l(n)+\lambda x_r(n+\Delta\tau)\Big),\\ \hat{s}_r(n) &= \frac{\lambda}{1+\lambda^2}\Big(x_l(n-\Delta\tau)+\lambda x_r(n)\Big),\end{aligned} \tag{11}$$

where $\Delta\tau$ denotes the interaural time-delay estimated by $\widetilde{\Delta\tau}$ in Sect.II. When $\Delta\tau = 0$ (The source speech propagates from the vertical median plane of two sensors), the shifted PCA is degenerated to naive PCA. When $\Delta\tau < 0$ (The source speech propagates from the right surface of two sensors), the extracted primary components are weighted sum of advanced left channel signal and delayed right channel signal. When $\Delta\tau > 0$ (The source speech propagates from the left surface of two sensors), the extracted primary components are weighted sum of advanced right channel signal and delayed left channel signal. In the frequency domain, the shifted PCA is given by

$$\begin{aligned}\hat{S}_l(\omega) &= \frac{1}{1+\lambda^2}\Big(X_l(\omega)+\lambda X_r(\omega)e^{j\omega\Delta\tau}\Big),\\ \hat{S}_r(\omega) &= \frac{\lambda}{1+\lambda^2}\Big(X_l(\omega)e^{-j\omega\Delta\tau}+\lambda X_r(\omega)\Big).\end{aligned} \tag{12}$$

Referring to Schwarz *et al*.'s work [18] by introducing a frequency-domain filter weight $G(\omega)$ on the binaural signals, we can extend this shifted PCA to adapt to the reverberant speech shown as

$$\begin{aligned}\hat{S}_l(\omega) &= \frac{1}{1+\lambda^2}\Big(X_l(\omega)+G_r(\omega)X_r(\omega)e^{j\omega\Delta\tau}\Big),\\ \hat{S}_r(\omega) &= \frac{\lambda}{1+\lambda^2}\Big(G_l(\omega)X_l(\omega)e^{-j\omega\Delta\tau}+\lambda X_r(\omega)\Big),\end{aligned} \tag{13}$$

where the $G_i(\omega), i\in\{l,r\}$ is evaluated by

$$G_i(\omega) = \max\Big(1-\mu\frac{1}{S\hat{N}R_i}, G_{\min}\Big), i\in\{l,r\}, \tag{14}$$

where the μ and $G_{\min}$ denote overestimate factor of SNR and spectral floor, respectively. And the frequency-dependent $S\hat{N}R_i$ is estimated by

$$S\hat{N}R_l = \frac{|X_l^2(\omega)e^{j\omega\Delta\tau}|}{|X_l^2(\omega)e^{j\omega\Delta\tau}-X_r^2(\omega)|}, S\hat{N}R_r = \frac{|X_r^2(\omega)|}{|X_l^2(\omega)e^{j\omega\Delta\tau}-X_r^2(\omega)|}. \tag{15}$$

IV. Beamformer with Post-filtering

Beamforming is one of the most well-known techniques in multichannel speech enhancement. In this section, an effective adaptive beamformer with post-filtering is introduced to enhance aforehand obtained dual-channel speech by shifted PCA, and the framework is shown in Fig.2.

Based on TDC and shifted PCA, the received speech can be modeled as $\hat{\boldsymbol{s}}_i = \boldsymbol{s}+\boldsymbol{v}_i, i\in\{l,r\}$, where the noise components $\boldsymbol{v}_i$ have been reduced to some extent. Simmer *et al*. [19] have demonstrated the multichannel Wiener filter could be expressed by a classical MVDR beamformer followed by a single-channel wiener post-filter, i.e. the optimum weight

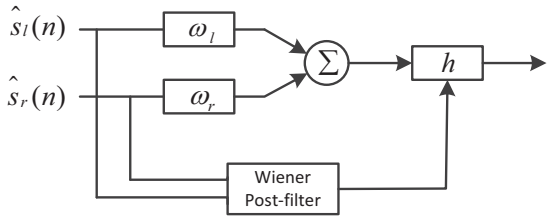

Fig. 2. Dual-channel adaptive beamformer with post-filter.

vector of Fig. 2 to process the noisy $\hat{\boldsymbol{s}}_i$ to best match desired $\boldsymbol{s}$, and is given by

$$\boldsymbol{w}_{opt} = \underbrace{\frac{\psi_{ss}}{\psi_{ss}+\psi_{vv}}}_{\text{Wiener filter}} \cdot \underbrace{\frac{\boldsymbol{\Psi}_{\hat{s}\hat{s}}^{-1}\boldsymbol{\alpha}}{\boldsymbol{\alpha}^H\boldsymbol{\Psi}_{\hat{s}\hat{s}}^{-1}\boldsymbol{\alpha}}}_{\text{MVDR beamformer}}. \tag{16}$$

We can see that the vector of optimal filter coefficients $\boldsymbol{w}_{opt}$ is factorized into two parts, that is

$$h = \frac{\psi_{ss}}{\psi_{ss}+\psi_{vv}}, \tag{17}$$

which is the transfer function of single-channel Wiener post-filter calculated whereafter, and

$$\boldsymbol{w}_{\text{MVDR}} = \frac{\boldsymbol{\Psi}_{\hat{s}\hat{s}}^{-1}\boldsymbol{\alpha}}{\boldsymbol{\alpha}^H\boldsymbol{\Psi}_{\hat{s}\hat{s}}^{-1}\boldsymbol{\alpha}}, \tag{18}$$

which is the system response of MVDR beamformer by MMSE. In Eq.(16), $\boldsymbol{\alpha}$ is the directional vector of source speech, which is only measured by the relative position versus receivers. Then the post-filter term should be considered based on the assumptions about correlation/uncorrelation, the auto- and cross-spectral densities on dual-channel can be easily deduced as

$$\psi_{\hat{s}_i\hat{s}_i} = \psi_{ss}+\psi_{vv}, \quad \psi_{\hat{s}_i\hat{s}_j} = \psi_{ss}, \tag{19}$$

where the terms on the left of equations can be estimated by standard recursive power spectral density, e.g.

$$\psi_{\hat{s}_i\hat{s}_j} = \beta\psi'_{\hat{s}_i\hat{s}_j} + (1-\beta)\hat{s}_i\hat{s}_j^*, \tag{20}$$

where $\psi_{\hat{s}_i\hat{s}_j}$ and $\psi'_{\hat{s}_i\hat{s}_j}$ are the spectral estimates for the current and previous frames, respectively, and $(\cdot)^*$ denotes the complex conjugate operator. β is called memory factor close to unity, and is given by $exp(-D/\tau_0 f_s)$, where D is the filterbank decimation factor, f_s is the sampling frequency, and τ_0 is the decay time constant.

V. Experiments and Discussions

In this section, the performances of this dual-channel speech enhancement algorithm are evaluated in a real scenario. The head-related impulse responses (HRIRs) in CIPIC database are used in experiments which are measured by the U. C. Davis CIPIC Interface Laboratory for 45 different subjects, where the distance from the speaker to dummy head is $1m$ and $azimuth \in [-80^\circ, 80^\circ], elevation \in [-45^\circ, 230.625^\circ]$ [20]. The source speech sentences are taken from the IEEE database corpus [21], which (about 7-12 words) are phonetically balanced with relatively low word-context predictability. The interference sources are derived from NOISEX database including

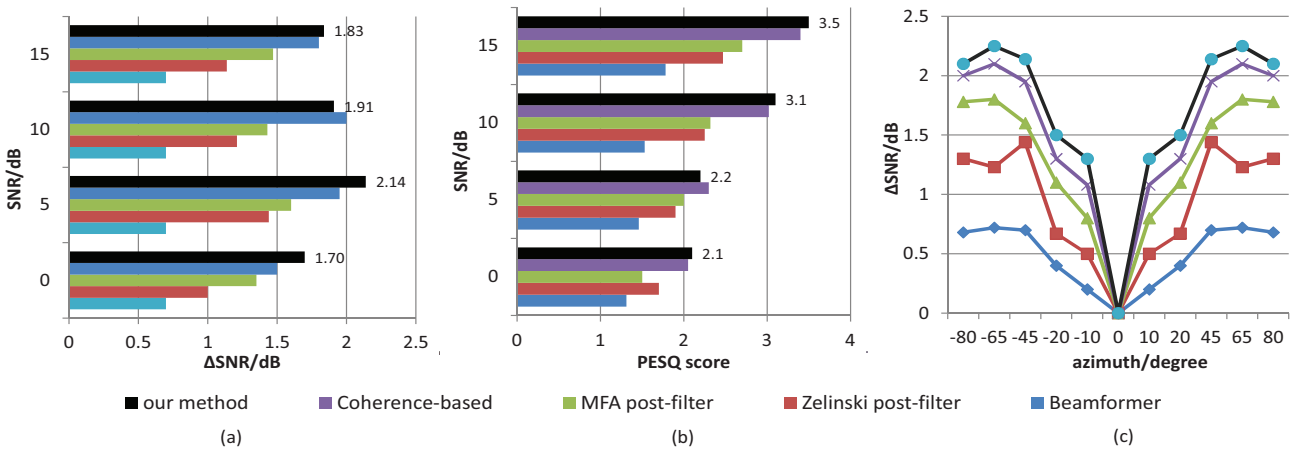

Fig. 3. Performance comparison of different methods: (a) ΔSNR, (b) Perceptual evaluation of speech quality (PESQ) in terms of SNR, (c) ΔSNR in various horizontal azimuths.

babble and factory noises [22]. The sampling frequency is $16kHz$ and frame length is about $100ms$. The overestimate factor μ is set to 1.5, the spectral floor $G_{\min}$ is 0.12. Some state-of-the-art methods are compared in experiments involving Beamformer [4], Zelinski post-filter [6], MFA post-filter [9], Coherence-based [8]. In these compared approaches, [4] denotes the base-line, [6] is a basic representative of beamforming with post-filter only using ITD, [9] corresponds to the up-to-date version of [6], and [8] means other mainstreams.

In order to assess the speech enhancement, first we set the source speech right ahead $(0°, 0°)$ and interference at $(45°, 0°)$. Fig.3(a) shows the SNR ($\log_{10}\frac{E(s^2(n)+v^2(n))}{E(v^2(n))}$ dB) enhancement, from which it can be seen that our method achieves the most prominent reduction of noise level generally, particularly with low $SNRs$. Then the Coherence-based follows, and the base-line beamformer is the most limited, because Coherence-based algorithms are more adaptive to process coherent noise.

In terms of enhanced speech quality, we have compared the PESQ [23] of these methods objectively. As Fig.3(b) illustrates, this method has not obtained obvious superiority compared with Coherence-based, since post-filtering would slightly deteriorate the speech intelligibility measures by over-estimating noise spectral. What's more, this method also outperforms the other three, which means TDC and shifted PCA have almost made up the defects of post-filter. The contribution should be attributed to exacting primary extractions by effective binaural cues estimates.

Additionally, when the interference is put in different horizontal directions with $SNR = 5dB$ to research the influence by azimuth, we have found that the larger angular difference between desired source signal and interference, the more ΔSNR enhanced in general as Fig.3(c) describes, which accords with the fact that more spacial distance means little listening jamming. Besides, the roughly symmetrical enhanced curves indicate that the existing methods often achieve same effectiveness for left/right plane.

Finally, we further extend these comparisons in reverberant environments to validate the availability of our algorithm to the real applications. In this context, the HRTFs are revised by the roomsim toolbox[1]. Some crude results are illustrated

[1]Available at: http://media.paisley.ac.uk/ campbell/Roomsim/.

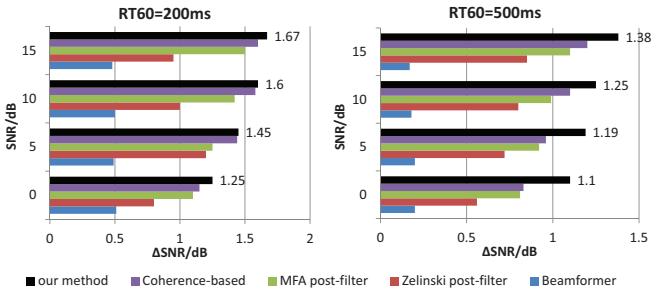

Fig. 4. SNR enhancement results of different methods with RT60.

in Fig. 4. We can see that when $RT60 = 200ms$, the ΔSNR of our method does not exceed others' too much, while this superiority is amplified saliently when $RT60 = 500ms$. This is mainly because the extended shifted PCA can be adaptive to the reverberation to some degree compared to other approaches. Practically speaking, with regard to reverberation, certain attached dereverberation strategies would be more effective.

VI. Conclusions

This paper presents a dual-channel beamformer with post-filter for speech enhancement, which can be applied into practical scenarios. The TDC algorithm is used to estimate binaural cues (i.e. ITD and IID), which are employed in the shifted PCA to extract primary components. Also we have made a improvement for the shifted PCA such as to adapt to reverberation in the frequency domain. Then the inputs of beamformer would avoid magnitude incoherence by weighting according to IID. Experiments turn out that the proposed processes are helpful to beamformer for both ΔSNR and PESQ, particularly this method shows some reverberation robustness. The TDC and shifted PCA can indeed do noise reduction to some extent, since the latter could realize primary components extraction from ambient components. In addition, it seems that post-filter would degrade the intelligibility of speech, whereas the proposed two modules offset the defect well.

VII. Acknowledgement

This work is supported by Huawei Innovation Research Program, National Natural Science Foundation of China (NSFC, No. 60875050, 60675025, 61340046), National High Technology Research and Development Program of China

(863 Program, No. 2006AA04Z247), Science and Technology Innovation Commission of Shenzhen Municipality (No. 201005280682A, No. JCYJ20120614152234873), Specialized Research Fund for the Doctoral Program of Higher Education (No. 20130001110011).

REFERENCES

[1] Yi Hu and Philipos C Loizou, "A comparative intelligibility study of single-microphone noise reduction algorithms," *J. Acoust. Soc. Amer.*, vol. 122, no. 3, pp. 1777–1786, 2007.

[2] Sharon Gannot, David Burshtein, and Ehud Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Trans. Signal Process.*, vol. 49, no. 8, pp. 1614–1626, 2001.

[3] Thomas Lotter and Peter Vary, "Dual-channel speech enhancement by superdirective beamforming," *EURASIP Journal on Applied Signal Processing*, vol. 2006, pp. 175–175, 2006.

[4] Otis Lamont Frost III, "An algorithm for linearly constrained adaptive array processing," *Proceedings of the IEEE*, vol. 60, no. 8, pp. 926–935, 1972.

[5] Lloyd J Griffiths and Charles W Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. Antennas, Propag.*, vol. 30, no. 1, pp. 27–34, 1982.

[6] Rainer Zelinski, "A microphone array with adaptive post-filtering for noise reduction in reverberant rooms," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 1988, pp. 2578–2581.

[7] Iain A McCowan and Hervé Bourlard, "Microphone array post-filter based on noise field coherence," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 709–716, 2003.

[8] Nima Yousefian and Philipos C Loizou, "A dual-microphone speech enhancement algorithm based on the coherence function," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 20, no. 2, pp. 599–609, 2012.

[9] Masoumeh Azarpour, Gerald Enzner, and Rainer Martin, "Adaptive binaural noise reduction based on matched-filter equalization and post-filtering," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2013, pp. 7259–7263.

[10] Yu Takahashi, Tomoya Takatani, Keiichi Osako, Hiroshi Saruwatari, and Kiyohiro Shikano, "Blind spatial subtraction array for speech enhancement in noisy environment," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 17, no. 4, pp. 650–664, 2009.

[11] Hong Liu, Zhuo Fu, and Xiaofei Li, "A two-layer probabilistic model based on time-delay compensation for binaural sound localization," in *IEEE Int. Conf. on Robotics and Automation (ICRA)*, 2013, pp. 2705–2712.

[12] Hong Liu, Jie Zhang, and Zhuo Fu, "A new hierarchical binaural sound source localization method based on interaural matching filter," in *IEEE Int. Conf. on Robotics and Automation (ICRA)*, 2014, pp. 1598–1605.

[13] Hong Liu and Jie Zhang, "A novel binaural sound source localization model based on time-delay compensation and interaural coherence," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2014, pp. 1438–1442.

[14] Charles Knapp and G Clifford Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 4, pp. 320–327, 1976.

[15] Peter R Roth, "Effective measurements using digital signal analysis," *IEEE Spectrum*, vol. 8, no. 4, pp. 62–70, 1971.

[16] Michael M Goodwin and Jean-Marc Jot, "Primary-ambient signal decomposition and vector-based localization for spatial audio coding and enhancement," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2007, vol. 1, pp. I–9.

[17] Gan Woon-Seng He, Jianjun and Tan En-Leng, "A study on the time-frequency domain primaryambient extraction for stereo audio signals," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2014, pp. 2868–2872.

[18] Andreas Schwarz, Klaus Reindl, and Walter Kellermann, "A two-channel reverberation suppression scheme based on blind signal separation and wiener filtering," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2012, pp. 113–116.

[19] Claude Marro, Yannick Mahieux, and Klaus Uwe Simmer, "Analysis of noise reduction and dereverberation techniques based on microphone arrays with postfiltering," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 3, pp. 240–259, 1998.

[20] V Ralph Algazi, Richard O Duda, Dennis M Thompson, and Carlos Avendano, "The cipic hrtf database," in *IEEE Workshop Appl. Signal Process. Audio, Acoust. (WASPAA)*, 2001, pp. 99–102.

[21] EH Rothauser, WD Chapman, N Guttman, KS Nordby, HR Silbiger, GE Urbanek, and M Weinstock, "Ieee recommended practice for speech quality measurements," *IEEE Transactions on Audio Electroacoust*, vol. 17, no. 3, pp. 225–246, 1969.

[22] Andrew Varga and Herman JM Steeneken, "Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech communication*, vol. 12, no. 3, pp. 247–251, 1993.

[23] 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 *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2001, vol. 2, pp. 749–752.