A DNN-based Post Filter for Geometric Source Separation

Chenghao Chen1, Yi Zhou2,* and Hongqing Liu3

1,2,3School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing, China

*Corresponding author e-mail: zhouy@cqupt.edu.cn

Abstract. In robot audition, the capability of separating or extracting the target speaker signal from the mixture is essential for the auditory information processing. As a powerful method in source separation, the geometric source separation (GSS) algorithm is widely used in many robot audition systems. However, its performance is limited in the real “cocktail party” environments, where the target source is immersed in background noise and interference. In this paper, a deep neural network (DNN)-based post filter is proposed to enhance the output of the GSS method. The proposed DNN-based post filter can suppress the residual background noise and channel leakage without making any assumption about the noise and the leakage as the conventional approach does. The experimental results show that the proposed approach outperforms the conventional method in terms of the short-time objective intelligibility (STOI) and the perceptual evaluation of speech quality (PESQ).

1. Introduction

Our hearing systems allow us to extract one or several desired speech sources from the adverse acoustic environments. To operate like human beings, the robots should be able to not only detect sounds but also separate simultaneous sources from mixtures received by microphone array placed inside a noisy and reverberant enclosure. Usually, there are two effective methods that may achieve this goal: geometric source separation (GSS) and deep neural network (DNN).

Merging the blind source separation with adaptive beamforming [1], geometric source separation (GSS) [2] has been widely used in many robot audition systems [3-4]. Research results have shown that GSS outperforms conventional blind source separation (BSS) as it resolves the permutation ambiguity in the independence criterion by utilizing geometric constraint provided by beamformer [2]. However, its performance is limited in real "cocktail party" situations. In these cases, the target source is immersed in background noise and interference, which is difficult for GSS to deal with. Furthermore, the problem of source leakage may be serious under low source to interference ration (SIR) environment or when reverberation is involved. To overcome these drawbacks, a post filter for GSS was proposed in [4]. The authors aimed at alleviating the interference of stationary noise and constant channel leakage by using two estimators under a suppression rule. However, the post filter performs poorly in non-stationary environments, because of the delay of the noise estimation.

Deep neural network (DNN)-based methods have been successfully applied in monaural sound source separation [5] and noise suppression [6]. Unlike the conventional sound source separation and speech enhancement methods, no linear assumption needs to be made for DNN-based method. A single DNN can be trained for monaural sound source separation and noise suppression...
simultaneously. It has been shown that DNN based methods outperform the conventional approaches in both sound source separation and noise suppression [5, 6].

In this paper, a novel post filter based on DNN for GSS method is proposed to achieve a better separation performance. The whole system is implemented in two steps: in the first step, a GSS method is adopted to get a primary separation result. In the second step, the preliminary separated audio is then enhanced by a well-trained DNN-based post filter. The main advantages of this system reside in the fact that GSS benefits the DNN training by utilizing the geometric information, meanwhile, DNN reduces the impact of non-stationary noise and leakage between channels. Experimental results reveal that the novel post filter outperforms the conventional approaches consistently, even in the environments corrupted by non-stationary noise under low SIR.

The rest of this paper is organized as follows: The overview of the system is presented in Section 2. The GSS algorithm and the proposed post filter are introduced in Section 3. The experimental setting and results are reported in Section 4. Section 5 concludes this paper.

2. System Overview

![Figure 1. Overview of the system.](image)

The whole system is composed of three parts: A microphone array, a GSS module, and a DNN-based post filter as shown in Figure 1.

With four omni-directional microphones, the uniform linear microphone array picks up mixture signal from a noisy environment for the GSS module to process. Without loss of generality, assuming that the interested speaker and the interference speaker are located by a localization algorithm such as generalized cross correlation with phase transform (GCC-PHAT) in [7], we use GSS to get a primary separation result. Figure 2 shows the training and enhancing stage of the DNN-based post filter. Given the output of GSS and reference label of the interested speaker, the DNN is trained to model the complex mapping from the GSS output to the enhanced separation results.

3. The GSS method and proposed post filter

The separation method used in this paper is the geometric source separation (GSS) approach proposed by Parra and Alvino in [2]. As a hybrid algorithm of conventional BSS and adaptive beamforming,
GSS alleviates the drawbacks of BSS and beamforming such as permutation ambiguity and cross-talk problem by introducing “geometric constraint” as a penalty term in optimization criterion. The proposed post filter is a DNN trained to enhance the result of GSS by alleviating the impact of non-stationary background noise and leakage between channels.

3.1. Geometric Source Separation
Considering the non-stationary nature of speech signal, we apply the GSS method in the frequency domain by using the short-time Fourier transform (STFT). The signal received by microphones at time frame $l$ and frequency point $k$ can be expressed as:

$$x(k,l) = A(k)s(k,l) + n(k,l)$$  \hspace{1cm} (1)

where $A(k)$ is the transfer function matrix from the sources to microphones. $s(k,l)$ refers to sources vector consisting of the target source $s_t(k,l)$ and the interference source $s_i(k,l)$ and $n(k,l)$ represent the uncorrelated noise. When the transfer functions are assumed to have unity gain, the elements of $A(k)$ can be described as:

$$a_{ij} = e^{-j2\pi\tau_{ij}}$$  \hspace{1cm} (2)

where $\tau_{ij}$ refers time delay from source $i$ to source $j$ which can be estimated by localization algorithm.

The goal of GSS is to find the separation matrix $W(k)$ that inverts the effect of transfer function matrix $A(k)$ by producing separated source vector $y(k,l)$ such that:

$$y(k,l) = W(k)x(k,l).$$  \hspace{1cm} (3)

As a hybrid approach of beamforming and BSS, GSS algorithm computes the separation matrix $W(k)$ by forming a combined cost function:

$$\lambda_0 J_1(W(k)) + \sum_k J_2(W(k))$$  \hspace{1cm} (4)

where $\lambda_0$ is a scaling constant, such that the two terms are of the same order of magnitude. Cost function $J_1(W(k))$ and $J_2(W(k))$ are deduced from independent criterion and geometric constraint respectively, and can be expressed as:

$$J_1(W(k)) = \sum_{k,l} \alpha(k)\|R_{yy}(k) - \text{diag}[R_{yy}(k)]\|^2,$$  \hspace{1cm} (5)

$$J_2(W(k)) = \|W(k)A(k) - I\|^2$$  \hspace{1cm} (6)

where $\|\cdot\|^2$ refers to the Frobenius norm, defined as $\|M\|^2 = \text{Tr}(MM^H)$. $\alpha(k) = \sum\|R_{\alpha}(k,l)\|^2$ is the power normalized factor designed for faster convergence when using gradient descent. The gradient descent method is then used to updated the separation matrix $W(k)$.

3.2. Proposed DNN-based post filter
Due to the complex working situations of robots, as a powerful method in robot audition, the GSS algorithm may be applied in the adverse acoustic environment. As a hybrid approach of beamforming and BSS, the GSS method shows a certain degree of robustness to noise, for its beamforming part enhances the signals from the desired direction while suppressing those from other directions. Assuming the direction of the target sources is known, source separation can be enhanced by forming individual beams at the target sources separately. However, the GSS method performs poorly in the cocktail party setting, for it cannot suppress the interfering noise from the target direction and the speech quality of GSS output may be deteriorated by channel leakage. To achieve a better performance, a post filter for the GSS was firstly proposed in [4]. Given the output of GSS, they form a combined noise estimator:

$$d(k,l) = d^{\text{stat}}(k,l) + d^{\text{leak}}(k,l)$$  \hspace{1cm} (7)
where $d^{\text{stat}}(k,l)$ is the estimator of the stationary noise computed by Minima Controlled Recursive Average (MCRA) proposed by Cohen[8]. $d^{\text{leak}}(k,l)$ represents leakage from the interference source under the assumption of constant leakage. The combined estimator is then used in a minimum mean-square error (MMSE) based suppression rule to get a spectral gain which can be used to compute the enhanced separation result. However, it often fails to suppress non-stationary noise in the complex acoustic environment, and a trade-off needs to be made to avoid speech distortion due to the complicated interaction between speech and noise [6].

Considering the sophisticated statistical properties of noise and interference, we propose a DNN-based post filter which consists of three hidden layers, one output layer and one input layer as displayed in Figure 3.

Unlike many other DNN-based methods using mixture generated by contaminating the clean target speech utterances with interference utterances, the GSS output signal is used as training data to avoid the possible mismatches. The log-power spectra (LPS) of input data are extracted as the training features. The DNN is adopted here to estimate the LPS feature of the interested source from the contaminated feature. The training stage of the DNN consists of two steps. In the first step, the DNN is pre-trained as a restricted Boltzmann machine (RBM) [9]. The contrastive divergence (CD) [10] are used to train the RBM layer-by-layer. Afterwards, a back-propagation method is applied to fine tune the DNN parameters. The final goal of this step is to minimize the following cost function by using a stochastic gradient descent algorithm,

$$
E = \frac{1}{N} \sum_{n=1}^{N} \left\| \hat{X}_n - \mathbf{Y}_n^{*+\tau} \mathbf{W} + b \right\|_2^2
$$

(8)

where $E$ denotes the mean squared error, $\hat{X}_n$ is the $n$th estimated feature and $X_n$ is the reference label. $\mathbf{Y}_n^{*+\tau}$ refers to the feature of GSS output computed by eqn.(3). $\mathbf{W}$ and $b$ represent the weight and bias parameters to be learned, and $N$ is the batch size.

During learning, no assumption needs to be made about noise and interference. The DNN can achieve separation task and noise suppression task simultaneously by automatically learning the complicated mapping between the contaminated mixture and the separated target speech. In order to get a better generalization capability in separation task, a semi-supervised mode is adopted where the unseen interference is predicted by multiple interference speech in the training stage and is exclude during training. The training data of the interested speaker is provided. In noise suppression task, many different types of noise are used during the training stage to get a good denoise performance in unseen noise condition [11].
4. Experimental Results

In section 4, we conduct simulation experiments using synthetic audio for proposed approach and its baseline methods. To evaluate the separation performance, the short-time objective intelligibility (STOI) [12] and the perceptual evaluation of speech quality (PESQ) [13] are used.

4.1. Setup

![Simulation setting](image)

Figure 4. Simulation setting

To evaluate the proposed post filter and its baseline methods, several experiments were conducted on the simulation setting shown in Figure 4. Two speakers were used as target source and interference source respectively located at 45° and 135°. The uniform linear microphone array contains four omnidirectional microphones spaced by 5cm and was placed at the same height as the speakers. All the audio materials were synthesized using utterances drawn from the GRID corpus [14], which consists of 34 speakers including 18 males and the rest of them are females. Each speaker has 1000 utterances. Before generating the training set and the test set for DNN, we first generated the input audio data of GSS. Speaker 1 in GRID corpus was chosen to be the target speaker while speaker 2 to speaker 33 were selected as interference speakers during training and speaker 34 was left to be the interference speaker in the test stage. Selecting 500 utterances randomly from each speaker, we generated the two-speaker mixture by mixing the target speaker and interference speakers at a range of SIR from -5dB to 5dB with an increment of 5dB. In the training stage, the background noise was randomly drawn from 100 environmental noises [15], and then added to the mixture with the same power as interference, while in the test stage, the unseen noise including f16 noise and the babble noise in were drawn from NOISEX-92 corpus [16]. The f16 noise represents the stationary noise, while the babble noise is a typical representative of non-stationary noise. Given the noisy mixture, the audio data at each microphone was obtained using the image method [17] with a reverberation time of 250ms. The room dimensions were set to 3m x 4m x 2.5m. Used as training and testing data for DNN, the GSS output was obtained straightforwardly by processing the microphone array received data with GSS module.

As for signal analysis, the frame size was set to 512 samples with a overlap of 50% between two frames. A 512-point STFT was applied to each frame. Consisting of one input layer, three hidden layers and one output layer, the DNN adopted in all experiments have the same architecture. The input layer has 1285 (257 x 5, \( \tau = 2 \)) neurons, and each hidden layer has 2048 neurons. The number of neurons for output layer was set to 257, due to the size of STFT. The learning rate was set to 0.01 during the first 20 epochs. To avoid the over-fitting, the learning rate decreased by 10% at each epoch, and training phase was early stopped when there was no further decrement for the cost function in three consecutive epochs.
4.2. Results

Figure 5. Separation Performance (PESQ, STOI) comparison at different SIR levels with babble noise.

Figure 6. Separation Performance (PESQ, STOI) comparison at different SIR levels with f16 noise.

In these experiments, we made comparisons among microphone received data, plain GSS, GSS with conventional post filter in [4], denoted as GSS + post filter, and GSS with our proposed DNN-based post filter, denoted as GSS + DNN. The experiments were conducted under three input SIR levels with two different noise types. Figure 5 (a) and (b) shows PESQ and STOI comparisons of these approaches with babble noise, and Figure 6 (a) and (b) presents the experimental results with f16 noise. Obviously, the GSS method with proposed post filter outperforms those baselines consistently at all SIR levels and noise types. The performance gap between GSS with conventional post filter and GSS with DNN based post filter was up to 0.29 and 1.11 under the measurement of STOI and PESQ respectively.

5. Conclusion

A DNN-based post filter for GSS is proposed in this paper. It demonstrates a better separation performance than conventional post filter in the real cocktail party setting where the mixture of the interested speaker and the interference speaker are corrupted by background noise. Moreover, the proposed post filter models the complex mapping from the GSS output to the reference clean audio directly without making any assumption about the leakage and the noise. However, it should be noted that this study only focuses on the static sources, the post filter for moving source separation should be studied in future.
References

[1] B. Van Veen and K. Buckley, “Beamforming techniques for spatial filtering,” in Digital Signal Processing Handbook. Boca Raton, FL: CRC, 1997.

[2] L. C. Parra and C. V. Alvino, “Geometric source separation: Merging convolutive source separation with geometric beamforming,” IEEE Transactions on Speech and Audio Processing, vol. 10, no. 6, pp. 352–362, 2002.

[3] Grondin, F., L’etourneau, D., Ferland, F., Rousseau, V., Michaud, F.: The ManyEars open framework: Microphone array open software and open hardware system for robotic applications. Autonomous Robots 34(3), 217–32 (2013)

[4] J.-M. Valin, J. Rouat and F. Michaud, Enhanced robot audition based on microphone array source separation with post-filter, in: Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, Sendai, pp. 2133–2128 (2004).

[5] Y. Tu, J. Du, Y. Xu, L. Dai, and C.-H. Lee, “Deep neural network based speech separation for robust speech recognition,” in Proc. Int. Symp. Chinese Spoken Lang. Process., 2014, pp. 532–536. [6] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, “A regression approach to speech enhancement based on deep neural networks,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 23, no. 1, pp. 7–19, Jan. 2015.

[6] C. Knapp, G. Carter, The generalized correlation method for estimation of time delay, IEEE Transaction on Acoustics, Speech, Signal Processing 24 (4) (1976) 320327.

[7] I. Cohen and B. Berdugo, “Speech enhancement for non-stationary noise environments,” Signal Processing, vol. 81, no. 2, pp. 2403–2418, 2001.

[8] Y. Bengio, “Learning deep architectures for AI,” Foundat. And Trends Mach. Learn., Vol. 2, No. 1, pp. 1-127, 2009.

[9] G. E. Hinton, S. Osindero, and Y. Teh, “A fast learning algorithm for deep belief nets,” Neural Comput., vol. 18, no. 7, pp. 1527–1554, 2006.

[10] Y. X. Wang and D. L. Wang, “Towards scaling up classification-based speech separation,” IEEE Trans. Audio, Speech Lang. Process., vol. 21, no. 7, pp. 1381–1390, 2013.

[11] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, “A short-time objective intelligibility measure for time-frequency weighted noisy speech,” in Proc. IEEE Int. Conf. Acoust, Speech, Signal Process., 2010, pp. 4214–4217.

[12] ITU-T, Recommendation P.862, “Perceptual evaluation of speech quality (PESQ): an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs,” International Telecommunication UnionTelecommunication Standardisation Sector, 2001.

[13] M. Cooke, J. Barker, S. Cunningham, and X. Shao, “An audiovisual corpus for speech perception and automatic speech recognition,” Journal of the Acoustical Society of America, Vol. 120, No. 5, pp. 2421-2424, 2006.

[14] G. Hu, 100 nonspeech environmental sounds, 2004 [Online]. Available: http://www.cse.ohio-state.edu/pnl/corpus/HuCorpus.html

[15] A. Varga and H. J. M. 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 Commun., vol. 12, no. 3, pp. 247–251, 1993.

[16] J. B. Allen and D. A. Berkley, “Image method for efficiently simulating small-room acoustics,” J. Acoust. Soc. Amer., vol. 65, pp. 943–950, Apr. 1979.