Speech Enhancement via Combination of Wiener Filter and Blind Source Separation

Hongmei Hu[1,4], Jalil Taghia[2], Jinqiu Sang[1], Jalal Taghia[3], Nasser Mohammadiha[2], Masoumeh Azarpour[3], Rajyalakshmi Dokku[3], Shouyan Wang[1], Mark E Lutman[1] and Stefan Bleeck[1]

[1] Institute of Sound and Vibration Research, University of Southampton, Southampton, UK
[2] School of Electrical Engineering, Royal Institute of Technology, Stockholm, Sweden
[3] Institute of Communication Acoustics, Ruhr-Universitaet, Bochum, Germany
[4] Department of Testing and Control, Jiangsu University, Zhenjiang, China

hongmei.hu@soton.ac.uk

Abstract. Automatic speech recognition (ASR) often fails in acoustically noisy environments. Aimed to improve speech recognition scores of an ASR in a real-life like acoustical environment, a speech pre-processing system is proposed in this paper, which consists of several stages: First, a convolutive blind source separation (BSS) is applied to the spectrogram of the signals that are pre-processed by binaural Wiener filtering (BWF). Secondly, the target speech is detected by an ASR system recognition rate based on a Hidden Markov Model (HMM). To evaluate the performance of the proposed algorithm, the signal-to-interference ratio (SIR), the improvement signal-to-noise ratio (ISNR) and the speech recognition rates of the output signals were calculated using the signal corpus of the CHiME database. The results show an improvement in SIR and ISNR, but no obvious improvement of speech recognition scores. Improvements for future research are suggested.

Keywords: ASR, BWF, BSS

1 Introduction

In real life scenarios, speech must often be recognized in noisy environments, where the target speech is contaminated by both noise and interference speech. This scenario is often referred to as "cocktail party effect" [1]. Applications that require speech recognition (teleconferencing, automatic speech recognition (ASR), hearing aids, etc) do not work well in such environments. Traditional speech enhancement algorithms often work only in narrowly specified conditions or with specific noise statistics. Algorithms exist and work well, when the background noise is stationary and non-speech [2-5], however, these algorithms often fail when competing speakers are present. A possible solution to this problem is to use source separation algorithm like beamform-

ing [6, 7]. Beamforming algorithms make use of a microphone array to form a beam towards the target signal. However, beamforming algorithms require a-priory knowledge about the acoustic environment and the sources involved, or a large number of sensors are required for good performance. Another algorithm for source separation is blind source separation (BSS) [8, 9]. In BSS source signals are estimated only based on the information of signals observed at each input channel. BSS thus requires no a-priory knowledge and furthermore, requires only a small number of microphones.

In this paper, we describe a system that was developed to deal with a given real-life like acoustical environment that was defined in the PASCAL 'CHiME' challenge [10]. The task of this challenge is to automatically recognize spoken commands in an acoustically clustered environment that was recorded via KEMAR in a living room and a kitchen and is contained in the CHiME domestic audio corpus. Various real-life noise sources include for example competing speakers, a television set, a washing machine, closing doors, a child hitting sticks and many more. In this paper, a novel source separation speech enhancement system is proposed that combines binaural Wiener filtering (BWF) pre-processing and corrected rate based target speech selection. We aimed to develop a flexible approach that combines the strengths of BSS techniques with noise reduction algorithms.

2 Methodologies

Fig. **1** shows the overall framework of the proposed system. It consists of three parts: BWF, BSS and target speech selection.

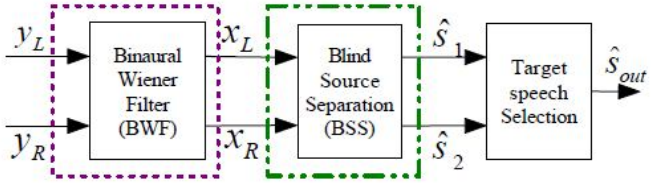

Fig. 1. Binaural Wiener filter preprocessed BSS noise reduction framework

First, noise reduction, based on BWF is performed. After applying single channel noise power spectral density (PSD) estimation to the mixture in each channel individually, BWF is used to remove some types of noise signals which are neither speech like nor too non-stationary. Secondly, the pre-processed binaural signals are transferred into the time-frequency domain and complex-valued independent component analysis (ICA) is applied to each frequency bin. In the end, we will have two separated signals. One of them is the target source and the other interference. Hence, in our model, both of the separated signals are fed into an ASR system provided by the CHiME challenge. Finally, the signal with highest recognition rate is selected as the target speech for the CHiME challenge.

2.1 Binaural Noise Reduction

As a first step of the BSS, two channel Wiener filtering is performed in order to suppress the interfering noise signals as effectively as possible while keeping the binaural cues of target speech intact. Fig.2 shows the workflow of this binaural noise reduction algorithm. The algorithm mainly consists of two parts: single channel noise power estimation and a two channels Wiener filter.

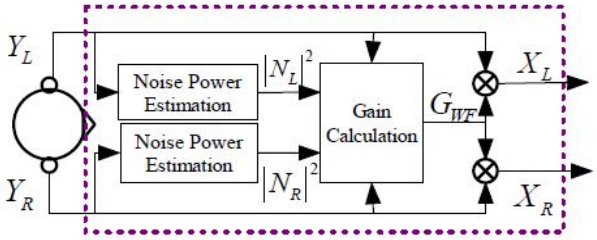

Fig.2. Binaural noise reduction scheme

2.1.1 Noise Power estimation

Binaural noise power estimators usually assume certain conditions, for instance knowledge about the direction of the target source. In our situation, this information is unknown, therefore we chose to use a single channel noise power estimator and apply it to the left and the right channels individually. In a second step, a noise PSD estimation algorithm based on the minimum mean square error (MMSE) by Hendriks et al. [11] was used to estimate the noise PSD of noisy speech in each channel individually. This specific noise estimator has been shown to estimate noise power robustly, and it can track non-stationary noises with reasonably low mean estimation error and low estimation error variance [2].

2.1.2 Two channel Wiener filter

Beamforming [6, 7], post-filtering [4, 5] and multi-channel Wiener filtering [3] are often used in multi-channel noise reduction. However, beamforming [6, 7] requires a robust estimation of the direction of the target speech. In the given sound material that was recorded in highly reverberated environment (e.g., 300 ms reverberation), this was difficult to estimate correctly and therefore we decided against using beamforming. Post-filtering [4, 5] on the other hand, assumes that background noises in each channel are only weakly correlated and not directional. In the given sound material noise sources were often directional, and therefore post-filtering was not applied. As a promising alternative, BWF was chosen as pre-processing in order to keep some binaural cues in the two channels for the following BSS processing.

In binaural processing, the observed signals, $Y_L(k,l)$ and $Y_R(k,l)$, in the k^{th} frequency bin and the l^{th} frame in the left and right channels can be written as

$$Y_L(k,l) = X_L(k,l) + N_L(k,l) \qquad (1)$$
$$Y_R(k,l) = X_R(k,l) + N_R(k,l)$$

where k and l, denote the frequency bin index and the frame index respectively; $X_i(k,l)$ and $N_i(k,l)$, (i =L, R), are the short-time Fourier transform (STFT) of the speech and noise signals. The speech signals represent the mixture of both the target and the interfering speech. The noise signals represent background interfering noise signals (e.g. washing machine, door closing, child hitting sticks, etc). The single channel noise estimation [11] is robust at estimating PSD of the background noise signals. Two channel Wiener filtering based on a-priori SNR estimation is implemented [12] for its capability of reducing "musical noise" [3]. Its gain function is

$$G_{WF}(k,l) = \frac{\xi(k,l)}{1+\xi(k,l)} \qquad (2)$$

where $\xi(k,l)$ is the a priori SNR as defined in [13]. The a-priori SNR $\xi(k,l)$ is updated in a decision-directed scheme by

$$\xi(k,l) = \alpha \frac{\left|X_L(k,l-1)\right|^2 + \left|X_R(k,l-1)\right|^2}{\left|N_L(k,l-1)\right|^2 + \left|N_R(k,l-1)\right|^2} + (1-\alpha) \times \max\left[\gamma(k,l) - 1, 0\right] \qquad (3)$$

where α (0< α <1) is a 'forgetting' factor and $\gamma(k,l)$ is the a posterior SNR as defined in [13].

2.2 Blind source separation

Supposing N sources and $M\,(M \geq N)$ microphones, given the source vector $\mathbf{s}(n) = \left[s_1(n),\ldots,s_N(n)\right]^T$, and the observed vector $\mathbf{x}(n) = \left[x_1(n),\ldots,x_M(n)\right]^T$, the mixing channels can be modeled by finite impulse response (FIR) filters of length P. The convolutive mixing process is formulated as

$$\mathbf{x}(n) = \mathbf{h}(n) * \mathbf{s}(n) = \sum_{p=0}^{P-1} h(p)s(n-p) \qquad (4)$$

where $\mathbf{h}(n)$ is a sequence of $M \times N$ matrices containing the impulse response of mixing channels. For separation, we use FIR unmixing filters of length L and obtain estimated source signal vector $\hat{\mathbf{s}}(n) = \left[\hat{s}_1(n),\ldots,\hat{s}_N(n)\right]^T$ by

$$\hat{\mathbf{s}}(n) = \mathbf{w}(n) * \mathbf{x}(n) = \sum_{l=0}^{L-1} w(l)x(n-l) \qquad (5)$$

Here, $\mathbf{w}(n)$ is obtained by a frequency-domain BSS approach.

2.2.1 Frequency-domain BSS

After transforming the signals to the time–frequency domain using blockwise L–point STFT, the convolution becomes a multiplication

$$\begin{aligned}\mathbf{X}(m,f) &= \mathbf{H}(f)\mathbf{S}(m,f) \\ \hat{\mathbf{S}}(m,f) &= \mathbf{W}(f)\mathbf{X}(m,f)\end{aligned} \tag{6}$$

where m is a decimated version of the time index n, $\mathbf{X}(m,f)$, $\hat{\mathbf{S}}(m,f)$, $\mathbf{H}(f)$ and $\mathbf{W}(f)$ are the STFTs of $\mathbf{x}(n)$, $\hat{\mathbf{s}}(n)$, $\mathbf{h}(n)$ and $\mathbf{w}(n)$, respectively, and $f \in \left[f_0,\ldots,f_{L/2} \right]$ is the frequency.

In the frequency domain, it is possible to separate each frequency bin independently using complex-valued instantaneous BSS algorithms such as FastICA [14, 15]. However, there are scaling and permutation ambiguities in each frequency bin. This is expressed as

$$\hat{\mathbf{S}}(m,f) = \mathbf{W}(f)\mathbf{X}(m,f) = \mathbf{\Lambda}(f)\mathbf{D}(f)\mathbf{S}(m,f) \tag{7}$$

where $\mathbf{D}(f)$ is a permutation matrix and $\mathbf{\Lambda}(f)$ a scaling matrix at frequency f. It is necessary to correct the scaling and permutation ambiguities before transforming the signals back to the time domain.

The scaling ambiguity can be resolved by using the minimal distortion principle [8] as

$$\mathbf{W}_s(f) = diag\left(\mathbf{W}_P^{-1}(f)\right) \cdot \mathbf{W}_P(f) \tag{8}$$

where $\mathbf{W}_P(f)$ is $\mathbf{W}(f)$ after permutation correction, $\mathbf{W}_s(f)$ is the one after scaling correction, $(\cdot)^{-1}$ denotes inversion of a square matrix or pseudo inversion of a rectangular matrix: $diag(\cdot)$ retains only the main diagonal components of the matrix.

Finally, the unmixing network $\mathbf{w}(n)$ is obtained by inverse Fourier transforming $\mathbf{W}_s(f)$, and the estimated source $\hat{\mathbf{s}}(n)$ is obtained by filtering $\mathbf{x}(n)$ through $\mathbf{w}(n)$. The workflow of the frequency-domain BSS is shown in Fig.3.

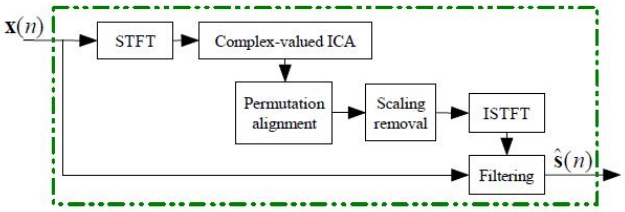

Fig.3. The workflow of frequency-domain BSS.

2.2.2 Permutation alignment

The inter-frequency dependence of speech sources can be exploited in order to align the permutations across all frequency bins. The correlation between separated signal envelopes is commonly used as a measure of inter-frequency dependence. However, this dependence is only clearly exhibited among a small set of frequencies. Another inter-frequency dependence measure is the correlation between power ratios of separated signals which exhibits a clearer inter-frequency dependence among all frequencies [9].

The $M \times N$ mixing network at frequency f can be estimated from the separation network by

$$\mathbf{A}(f) = \mathbf{W}^{-1}(f) = \left[a_1(f), \ldots a_N(f)\right] \tag{9}$$

where $a_i(f)$ is the i^{th} column vector of $\mathbf{A}(f)$. The observed signal can be decomposed to

$$\mathbf{X}(m,f) = \sum_{i=1}^{N} a_i(f)\hat{S}_i(m,f) \tag{10}$$

where $\hat{S}_i(m,f)$ is the ith component of $\hat{\mathbf{S}}(m,f)$ i.e., $\hat{\mathbf{S}}(m,f) = \left[\hat{S}_1(m,f), \ldots, \hat{S}_N(m,f)\right]^T$.

A power ratio measure is calculated to represent the dominance of the i^{th} separated signal in the observations at frequency f. It is defined as

$$\zeta_i^f(m) = \frac{\| a_i(f)Y_i(m,f) \|^2}{\sum_{k=1}^{N} \| a_k(f)Y_k(m,f) \|^2} \tag{11}$$

where the denominator is the total power of the observed signal $\mathbf{X}(m,f)$, and the numerator is the power of the i^{th} separated signal. Being in the range [0, 1], equation (11) is close to 1 when the i^{th} separated signal is dominant, and close to 0 when others are dominant. The power ratio measure exhibits the signal activity due to the sparseness of the speech signals. The correlation coefficient of signal power ratios can be

used for measuring inter-frequency dependence and solving the permutation problem. The normalized bin-wise correlation coefficient between two power ratio sequences $\zeta_i^{f_1}(m)$ and $\zeta_j^{f_2}(m)$ is defined as

$$\rho\left(\zeta_i^{f_1},\zeta_j^{f_2}\right)=\frac{r_{ij}(f_1,f_2)-\mu_i(f_1)\mu_j(f_2)}{\sigma_i(f_1)\sigma_j(f_2)} \tag{12}$$

where i and j are indices of two separated channels, f_1 and f_2 are two frequencies, $r_{ij}(f_1,f_2)=E\left\{\zeta_i^{f_1},\zeta_j^{f_2}\right\}$, $\mu_i(f)=E\left\{\zeta_i^{f}\right\}$, $\sigma_i(f)=\sqrt{E\left\{\left(\zeta_i^{f}\right)^2\right\}-\mu_i^2(f)}$ are, the correlation, mean, and standard deviation at time m. respectively. $E\{\cdot\}$ denotes the expectation value. Being in the range of [-1, 1], equation (12) equals equation (11) when the two sequences are identical. In general, equation (12) tends to be high if output channels i and j originate from the same source and low if they represent different sources. This property will be used for aligning the permutation.

We employed a procedure first to perform a rough global optimization followed by a fine local optimization [16]. In the global optimization, the centroid for each source is calculated as the average of the power ratio sequences with the current permutations by using a k-means clustering algorithm. Then the current permutations are optimized to maximize the correlation coefficients between power ratio sequences and the current centroids. This procedure is repeated until it converges. A local optimization is performed in order to achieve a better permutation alignment and to maximize the score values over a set of selected frequencies. In our system, adjacent and harmonic frequencies are thus considered. The fine local optimization is performed for one selected frequency f at a time, and repeated until no improvement is found for any frequency f.

2.3 Summary of the proposed algorithm

In summary, our system is described by the following steps:

Step 1: single channel noise power estimation is applied to each channel to estimate the PSD of noise.

Step 2: the estimated noise PSD of each channel is used in decision directed approach to derive a priori SNR.

Step 3: binaural Wiener filter is applied to the magnitude spectrum of noisy speech in both channels and two enhanced speech signals are derived.

Step 4: enhanced signals are transformed to the frequency domain by STFT.

Step 5: a complex-valued ICA is applied on each frequency bin.

Step 6: the permutation alignment is solved along frequency bins.

Step 7: the separated signals are transformed into time-domain by using ISTFT.

Step 8: selecting the target speech from the two separated sources by choosing the output with the higher recognition rate based on the provided ASR.

3 Results

The algorithm was evaluated on the test database provided by the CHiME challenge [17]. In the test set the target sound has been convolved with binaural room impulse responses (BRIR) and mixed with binaural recordings from the CHiME domestic audio corpus. The BRIR was measured at a position 2 meters in front of a KEMAR dummy. The temporal placement of the Grid utterances within the 20 hours of CHiME data has been controlled in a manner which produces mixtures at 6 different SNRS (-6, -3, 0, 3, 6, 9 dB) giving 3,600 test utterances in total, sampled at 16 kHz.

3.2 Evaluations and results

The performance of the proposed algorithm was evaluated by three measures: the signal-to-interference ratio (SIR), the improvement signal-to-noise ratio (ISNR) and finally the speech recognition rates. The SIR for the source i was calculated by

$$SIR_i = 10\log_{10}\frac{\sum_{j=1}^{M}\sum_t\left[s_j^{(i)}(t)+es_j^{(i)}(t)\right]^2}{\sum_{j=1}^{M}\sum_t\left[ei_j^{(i)}(t)\right]^2} \tag{13}$$

where $es_j^{(i)}(t)$ and $ei_j^{(i)}(t)$ represent filtering distortion, and interference, respectively. These two distinct errors are obtained by decomposing the estimated source i to the j^{th} channel, $\hat{s}_j^{(i)}(t)$, into:

$$\hat{s}_j^{(i)}(t) = s_j^{(i)}(t) + es_j^{(i)}(t) + ei_j^{(i)}(t) \tag{14}$$

Roughly, $es_j^{(i)}(t)$ stands for the distance between the estimated source $\hat{s}_j^{(i)}(t)$ and the filtered version of the source, $ei_j^{(i)}(t)$ is the quantity of other sources present in the estimated source [18]. The SNR is calculated in the same way as [19], and ISNR is the difference between the SNR of the enhanced speech, $SNR_{enhanced}$, and that of the noisy speech , SNR_{noisy}, as

$$ISNR = SNR_{enhanced} - SNR_{noisy} \tag{15}$$

Fig.4 shows results measured in SIR and ΔSNR. There are modest improvements in ISNR and SIR when the algorithm is evaluated using the CHiME test database. To further evaluate the proposed algorithm, a "standard" Hidden Markov Model (HMM)-based speech recognizer and a scoring tool in the PASCAL CHiME Challenge [10] was used to get the speech recognition rate. Table 1 shows the recognition rates of the enhanced speech.

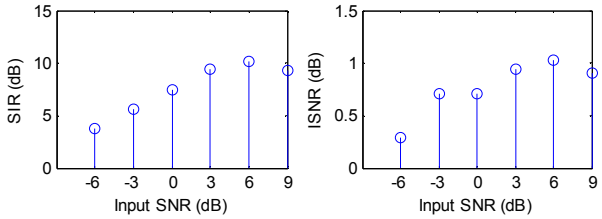

Fig.4. The SIR and ISNR for the whole CHiME test database.

Although there are some improvement in SNR and ISNR, Table 1 shows no obvious improvement for the ASR correct rates for all the conditions. The reason for this might be that the features for ASR training are extracted from clean speech in CHiME challenge, so one promising way to improve the recognition performance for this challenge might be to extract more robust features from the noisy speech and the enhanced speech in the future for this real-life like environments.

Table 1. Speech recognition rates in %

SNR / Speech	-6dB	-3dB	0dB	3dB	6dB	9dB
Baseline	30.3	35.4	49.5	62.9	75	82.4
Enhanced	31	36.2	50.9	63.3	75.3	82.5

4 Conclusions

A novel system was proposed to improve the ASR performance in real-life like acoustical environments using both noise reduction and source separation. First input signals are pre-processed by noise reduction based on noise estimation and BWF. Then a time-frequency domain independent component analysis (ICA) is applied on the spectrogram to implement blind source separation (BSS). To choose the target speech candidate with highest speech recognition rate, both of the separated signals are fed into an ASR system. The signal with higher recognition rate is finally selected as the target speech. The SIR the ISNR and the speech recognition rates were calculated to evaluate the algorithm. The proposed algorithm showed a positive effect on SIR and SNR, but no obvious improvement was found for the speech recognition corrected rates for the enhanced speech when using the CHiME database. The reason for the non-improvement in speech recognition rates might be that the ASR used in the CHiME challenge is trained with clean speech, in the future, one of our work can be focus on the extraction of the robust features in noisy environments, another promising way is that noise environment classification could be applied and specific noise reduction strategies could be chosen for different noise scenarios.

Acknowledgements

This work was funded by the European Commission within the Marie Curie ITN AUDIS, grant PITNGA-2008-214699. The authors performed this work while doing their group research at University of Southampton. The authors also appreciate Dr. Ning Ma in the University of Sheffield to help us setting up the CHiME ASR system.

References

1. A. S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound*, MIT Press, 1990.
2. J. Taghia, J. Taghia, N. Mohammadiha *et al.*, "An evaluation of noise power spectral density estimation algorithms in adverse acoustic environments," in ICASSP 2011, Prague, Czech Republic, 2011.
3. P. Scalart, and J. V. Filho, "Speech enhancement based on a priori signal to noise estimation." pp. 629-632 vol. 2.
4. I. Cohen, "Multichannel post-filtering in nonstationary noise environments," *Signal Processing, IEEE Transactions on,* vol. 52, no. 5, pp. 1149-1160, 2004.
5. C. Zheng, Y. Zhou, X. Hu *et al.*, "Two-Chennel Post-filtering Based on Adaptive Smoothing and Noise Properties," in ICASSP 2011, Prague, Czech Republic, 2011.
6. A. Sylvain, D. Patrick, and S. Philippe, "Modal Analysis Based Beamforming for Nearfield or Farfield Speaker Localization in Robotics." pp. 866-871.
7. J. M. Valin, F. Michaud, and J. Rouat, "Robust 3D Localization and Tracking of Sound Sources Using Beamforming and Particle Filtering." pp. IV-IV.
8. K. Matsuoka, "Minimal distortion principle for blind source separation." pp. 2138-2143.
9. H. Sawada, S. Araki, and S. Makino, "Measuring Dependence of Bin-wise Separated Signals for Permutation Alignment in Frequency-domain BSS." pp. 3247-3250.
10. "CHiME," http://www.dcs.shef.ac.uk/spandh/chime/PCC/introduction.html.
11. R. C. Hendriks, R. Heusdens, and J. Jensen, "MMSE Based Noise PSD Tracking with Low Complexity," in ICASSP, Dallas, TX, 2010, pp. p.4266-4269.
12. J. Li, S. Sakamoto, S. Hongo *et al.*, "Two-stage binaural speech enhancement with Wiener filter for high-quality speech communication," *Speech Communication,* vol. (In Press), 2010.
13. Y. Ephraim, and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *Acoustics, Speech and Signal Processing, IEEE Transactions on,* vol. 32, no. 6, pp. 1109-1121, 1984.
14. A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, New York: John Wiley & Sons, 2011.
15. E. Bingham, and A. Hyvarinen, "A fast fixed-point algorithm for independent component analysis of complex valued signals," *Int J Neural Syst,* vol. 10, no. 1, pp. 1-8, Feb, 2000.
16. H. Sawada, S. Araki, and S. Makino, "Underdetermined Convolutive Blind Source Separation via Frequency Bin-Wise Clustering and Permutation Alignment," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 19, no. 3, pp. 516-527, 2011.
17. H. Christensen, J. Barker, N. Ma *et al.*, "The CHiME corpus: a resource and a challenge for Computational Hearing in Multisource Environments," in Interspeech 2011, Makuhari, Japan, 2010.
18. E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 14, no. 4, pp. 1462-1469, 2006.
19. P. C. Loizou, *Speech Enhancement: Theory and Practive*: CRC Press, 2007.