INTEGRATING BINAURAL CUES AND BLIND SOURCE SEPARATION METHOD FOR SEPARATING REVERBERANT SPEECH MIXTURES

Atiyeh Alinaghi, Wenwu Wang and Philip JB Jackson

Centre for Vision, Speech and Signal Processing (CVSSP)
Department of Electronic Engineering (FEPS)
University of Surrey, Guildford GU2 7XH, UK

ABSTRACT

This paper presents a new method for reverberant speech separation, based on the combination of binaural cues and blind source separation (BSS) for the automatic classification of the time-frequency (T-F) units of the speech mixture spectrogram. The main idea is to model interaural phase difference, interaural level difference and frequency bin-wise mixing vectors by Gaussian mixture models for each source and then evaluate that model at each T-F point and assign the units with high probability to that source. The model parameters and the assigned regions are refined iteratively using the Expectation-Maximization (EM) algorithm. The proposed method also addresses the permutation problem of the frequency domain BSS by initializing the mixing vectors for each frequency channel. The EM algorithm starts with binaural cues and after a few iterations the estimated probabilistic mask is used to initialize and re-estimate the mixing vector model parameters. We performed experiments on speech mixtures, and showed an average of about 0.8 dB improvement in signal-to-distortion (SDR) over the binaural-only baseline.

Index Terms— EM algorithm, interaural phase difference, interaural level difference, blind source separation, mixing vectors

1. INTRODUCTION

In real environments, speech signals are usually collected together with other speakers' voice and background noise which can degrade the performance of automatic speech recognition (ASR) systems. Therefore, it is important to separate speech signals in recorded mixtures prior to further processing. One approach is blind source separation methods (BSS), such as independent component analysis (ICA) [1]. Although they show promising results in acoustically dry (anechoic) and overdetermined situations, their performance is limited in reverberant environments, especially for under-determined cases. One solution is to work in the frequency domain where the reverberant convolutive mixtures are transformed to the complex weighted product of the source spectrograms in each frequency bin [2]. However, the permutation alignment of the sources across frequency bins is still an issue in spite of different proposed solutions [3]. Another approach to solve the cocktail party problem, where speech signals are mixed, is Computational Auditory Scene Analysis (CASA) [4] which is inspired by the human auditory system and exploits monaural and binaural cues such as pitch, interaural level difference (ILD) and interaural phase difference (IPD). An important advantage of this method is that the number of sources can equal or exceed the number of microphones, which is usually two.

In this paper, we propose a new method for separating reverberant speech mixtures by classifying the T-F units of their spectrograms into different sources, basd on the integration of the ILD and IPD cues as in [5], and the mixing vectors estimated by a BSS algorithm in e.g. [6]. In both methods probability distribution functions are applied to model the ILD, IPD and **h** statistically which can be evaluated at each T-F point of the spectrogram. Then the parameters of each source model are re-estimated according to the T-F regions that are most likely to be dominated by that source. Once the model parameters have been updated, the probability of each T-F point dominated by a specific source will be refined by the EM algorithm to improve the results.

In section 2, the binaural cues are modeled. Bin-wise classification using the mixing vectors estimated by BSS is discussed in section 3. Section 4 explains the EM algorithm to maximize the combined log likelihood and estimate the model parameters of all the three cues, while solving the permutation problem of the frequency domain BSS . The experimental setup and results are in section 5, and finally section 6 contains the conclusions.

2. CLASSIFICATION OF TIME-FREQUENCY UNITS BASED ON BINAURAL CUES

In stereo recordings there are two microphones representing right and left ears, and so two mixtures are available, *l(n)* and *r(n)*, where n is the discrete time index. Each recording is the

Thanks to CVSSP for funding A. Alinaghi, and to B. Shinn-Cunningham for providing us with BRIRs. Special thanks to Michael Mandel for sharing his code and helping with its related issues.

combination of filtered source signals with additive or reverberant noise. It is found [5] that a reverberant noise model works for both cases:

$$l(n) = \sum_{i=1}^{N} s_i(n) * h_{il}(n) * n_l(n),$$
$$r(n) = \sum_{i=1}^{N} s_i(n) * h_{ir}(n) * n_r(n), \quad (1)$$

where N, known as *a priori*, is the number of sources, $s_i(n)$, $h_{il}(n)$ and $h_{ir}(n)$ are the ith source signal and the room impulse responses from source i to the left and right ears, respectively. $n_l(n)$ and $n_r(n)$ are the background noise. The spectrogram of each signal can be computed using the short time Fourier transform (STFT). The interaural spectrogram, i.e. the ratio of the left and right spectrograms, is formed:

$$\frac{L(\omega,t)}{R(\omega,t)} = 10^{\alpha(\omega,t)/20} e^{j\phi(\omega,t)} \quad (2)$$

where $L(\omega,t)$ and $R(\omega,t)$ are the transformed left and right signals at each frequency ω and time frame t, respectively. At each T-F point (ω,t), two observations are available, $\alpha(\omega,t)$, i.e. the ILD, and $\phi(\omega,t)$, i.e. the IPD. Since all the measured phases are wrapped to the range $(-\pi,\pi]$, they cannot be mapped to their corresponding interaural time differences (ITD) uniquely. In other words the targets with greater azimuths may be considered as being from smaller angles due to spatial aliasing. In order to avoid this ambiguity, a top-down process is suggested in [5] where the equally spaced ITDs corresponding to azimuths from -90^o to 90^o are mapped to the corresponding IPDs without ambiguity. Then the difference between the observed IPDs and the predicted IPDs gives the phase residuals $\hat{\phi}(\omega,t;\tau) = \arg(e^{j\phi(\omega,t)}e^{-j\omega\tau(\omega)})$ that can be modeled by a normal distribution for each candidate ITD, τ. The ILDs are modeled by a Gaussian distribution.

Therefore, the main task is to find the model parameters, i.e. the mean and variances, that best fit the observations (α and $\hat{\phi}$). The parameters that maximize the log likelihood for the given observation, can be estimated using the EM algorithm:

$$L(\Theta) = \sum_{\omega,t} \log p(\hat{\phi}(\omega,t;\tau), \alpha(\omega,t)|\Theta) \quad (3)$$

$$= \sum_{\omega,t} \log \sum_{i,\tau} [\psi_{i,\tau} . \mathcal{N}(\hat{\phi}(\omega,t;\tau)|\xi_{i,\tau}(\omega), \sigma^2_{i,\tau}(\omega)) \cdot \mathcal{N}(\alpha(\omega,t)|\mu_i(\omega), \eta_i^2(\omega))] \quad (4)$$

where $\xi_{i,\tau}, \sigma^2_{i,\tau}, \mu_i$ and η_i^2 are the mean and variance of the IPD residuals and the ILDs, respectively. Equation (4) represents the Gaussian mixture model with one Gaussian distribution for each source i and each azimuth (corresponding to each τ). Therefore, there are *N* (number of sources)$\times N_\tau$ (number of equally spaced ITDs) Gaussian distributions being mixed by the mixing weight $\psi_{i,\tau}$ which can be initialized by the PHAT histogram [7].

3. BIN-WISE CLASSIFICATION BY MIXING VECTORS ESTIMATION

In this method, instead of taking the ratio of the left and right spectrograms, the two measured signals are put together to form a new data whose elements are 2 dimensional vectors (the number of sensors). Moreover, assuming the sparseness of audio signals, at each T-F unit only one source is dominant and hence the STFT of observations at each T-F unit can be represented as:

$$\mathbf{x}(\omega,t) = \sum_{j=1}^{N} \mathbf{h}_j \mathbf{s}_j(\omega,t) \approx \mathbf{h}_j \mathbf{s}_j(\omega,t) \quad (5)$$

where $\mathbf{x}(\omega,t) = [L(\omega,t), R(\omega,t)]^T$ and $\mathbf{h}_j = [h_{jl}, h_{jr}]^T$. Then each observation vector is normalized to remove the effect of the source amplitude. The filter coefficients, $\mathbf{h}_k$, also known as the mixing matrices in BSS methods, are modeled as a complex Gaussian density function, evaluated for each observation [6].

$$p(\mathbf{x}|\mathbf{a}_i, \gamma_i) = \frac{1}{(\pi\gamma_i^2)^2} \exp\left(-\frac{||\mathbf{x} - (\mathbf{a}_i^H\mathbf{x}).\mathbf{a}_i||^2}{\gamma_i^2}\right) \quad (6)$$

where $\mathbf{a}_i$ is the centroid with unit norm $||\mathbf{a}_i||^2 = 1$, and γ_i^2 is the variance. The orthogonal projection of each observation $\mathbf{x}$ onto the subspace spanned by $\mathbf{a}_i$ can be estimated by $(\mathbf{a}_i^H\mathbf{x}).\mathbf{a}_i$. Therefore, the minimum distance between the point $\mathbf{x}$ and the subspace is $||\mathbf{x} - (\mathbf{a}_i^H\mathbf{x}).\mathbf{a}_i||$ which represents the probability of that point belonging to the ith class. In other words, the probability of each T-F unit coming from source i can be estimated for $i = 1...N$ to find out which source is dominant in that unit. Since the order of the recovered sources at each frequency bin is not necessarily the same as others, the permutation alignment is needed before transforming the signals to the time domain [6].

4. INTEGRATION OF BINAURAL CUES AND MODELED ROOM IMPULSE RESPONSE WITH EM ALGORITHM

To improve the reliability of allocating each T-F unit to a specific source, we propose to combine the above two approaches. Accordingly, three different observations are exploited $\{\hat{\phi}(\omega,t;\tau), \alpha(\omega,t)$ and $\mathbf{x}(\omega,t)\}$ with parameters $\hat{\Theta}$:

$$\hat{\Theta} = \{\xi_{i,\tau}(\omega), \sigma_{i,\tau}(\omega), \mu_i(\omega), \eta_i(\omega), \mathbf{a}_k(\omega), \gamma_k(\omega), \psi_{i,\tau}(\omega)\} \quad (7)$$

where $\xi_{i,\tau}, \sigma^2_{i,\tau}, \mu_i, \eta_i^2, \mathbf{a}_i$, and γ_i^2 are the mean and variance of the IPDs, the ILDs and the mixing vectors, respectively. However, the probabilistic classification in this BSS method is performed for each frequency bin separately and therefore the permutation alignment over the frequency bins is still a problem, as shown by different source index k in parameters (7). Although [6] introduced a method based on a posteriori

probability and showed that it works well, it is computationally expensive. Therefore, we propose an alternative approach using the information from the binaural cues.

4.1. Solving the permutation problem

As mentioned in section 4, the three different features of each T-F point can be combined to give more reliable information about the dominant source at each unit. However, the permutation problem of bin-wise classification should be solved before estimating overall probabilties:

$$L(\hat{\Theta}) = \max_{\theta} \sum_{\omega,t} \log p(\phi(\omega,t;\tau), \alpha(\omega,t), \mathbf{x}(\omega,t)|\Theta) \quad (8)$$

Since the EM algorithm can be initialized either from the E-step or the M-step and also there is usually no prior information about the mixing filters, we propose to initialize the mask first and then estimate the initial values of $\mathbf{a}_i(\omega)$ and $\gamma_i(\omega)$ based on the masked spectrogram. In order to initialize the mask properly, we applied IPD and/or ILD cues with initialized parameters to estimate the mask and let the program run for one iteration with no BSS contribution.

4.2. EM Algorithm

In the E-step, given the estimated parameters, Θ_s at M-step and the observations, assuming the statistical independence [5], the probability that each T-F unit, (ω, t), is dominated by source i at time delay τ is calculated as:

$$\begin{aligned}\nu_{i,\tau}(\omega,t) \propto \psi_{i,\tau}(\omega).\mathcal{N}(\hat{\phi}(\omega,t;\tau)|\xi_{i,\tau}(\omega), \sigma^2_{i,\tau}(\omega)).\\ \mathcal{N}(\alpha(\omega,t)|\mu_i(\omega), \eta_i^2(\omega)).\\ \mathcal{N}(\mathbf{x}(\omega,t)|\mathbf{a}_i(\omega), \gamma_i^2(\omega))\end{aligned} \quad (9)$$

where $\nu_{i,\tau}(\omega,t)$ is the occupation likelihood. In the M-step, the IPD residual parameters $(\xi_{i,\tau}(\omega), \sigma^2_{i,\tau}(\omega))$, and the ILD parameters $(\mu_i(\omega), \eta_i^2(\omega))$ are re-estimated for each source and the time delay using the estimated occupation likelihood $\nu_{i,\tau}(\omega,t)$ in the E-step and the observations, as explained in [5]. For the first iteration, we set $\mathcal{N}(\mathbf{x}(\omega,t)|\mathbf{a}_i(\omega), \gamma_i^2(\omega)) = 1$ in equation (9) to remove the effect of the BSS contribution. Once the masking $M_i(\omega,t) \equiv \sum_\tau \nu_{i,\tau}$ is obtained after one iteration based on only the information of binaural cues, the parameters of the mixing vectors, $(\mathbf{a}_i(\omega), \gamma_i^2(\omega))$, can be estimated from the next M-step without the permutation problem akin to [6].

$$\mathbf{R}_i(\omega) = \sum_{t,\tau} \nu_{i,\tau}(\omega,t).\mathbf{x}(\omega,t)\mathbf{x}^H(\omega,t) \quad (10)$$

$$\gamma_i^2(\omega) = \frac{\sum_{t,\tau} \nu_{i,\tau}(\omega,t).||\mathbf{x} - (\mathbf{a}_i^H\mathbf{x}).\mathbf{a}_i||^2}{\sum_{t,\tau} \nu_{i,\tau}(\omega,t)} \quad (11)$$

$$\psi_{i,\tau}(\omega) = \frac{1}{T}\sum_t \nu_{i,\tau}(\omega,t) \quad (12)$$

where T is the number of all time frames and optimum $\mathbf{a}_i$ is the eigenvector corresponding to the maximum eigenvalue of $\mathbf{R}_i$. Since the source order is known in $\nu_{i,\tau}$, the permutation problem is circumvented.

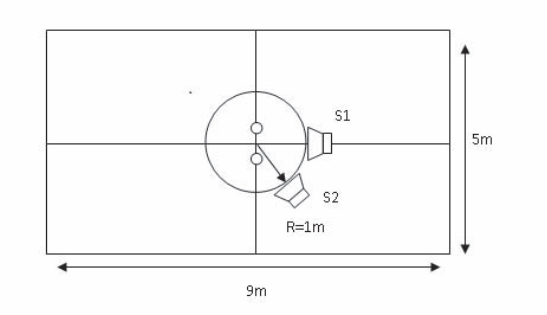

Fig. 1. Set-up configuration: source 1 is placed in front of the microphones and source 2 at θ^o clockwise to the central line of microphones pair.

5. EXPERIMENTS AND RESULTS

Similar to [5], 15 utterances with the length of approximately 3 s were chosen randomly from the TIMIT dataset and then shortened to 2.5 s for consistency with the silence at the end of the signals being removed. Moreover, all of them were normalized to have the same root mean square (RMS) amplitude before convolving with the room impulse responses (RIRs). The binaural RIRs (BRIRs) include a head related transfer function (HRTF) [8] with reverberation time of $T_{60} = 565$ ms. Since it was desirable to test the effect of reverberation time on the improvement of the proposed method, different RIRs were simulated with a similar configuration to [8] but without HRTFs as those applied by [5] were no longer available. Different T_{60} and different azimuths from 30^o to 75^o with step of 15^o were chosen for each set up. The two microphones were positioned 17 cm apart (similar to the size of human head diameter) at the center of the room. To generate the mixtures, 15 pairs from those 15 selected utterances were chosen. The target source was placed at 0^o and the interferer at 60^o, both of them at 1 m from the microphones.

The complexity of each model in [5] is represented by the mode, which ranges from the simplest mode where no ILD cues are used, Θ_{00}, to the most complicated one where the parameters of binaural cues are frequency dependent, $\Theta_{\Omega,\Omega}$. The concept of a garbage source is introduced to reduce the effect of reverberation. We studied the performance of our proposed method under all these modes. We also studied a novel mode Θ_{11}^G with frequency independent cues and garbage source (denoted by the superscript G).

The performance of binaural cues without and with the BSS contribution is evaluated based on the signal-to-distortion ratio (SDR) [9]. We applied an FIR Wiener filter to the estimated signal with the target signal as reference. Therefore, any energy in the estimated signal that could be explained by a filtered version of the target signal was considered as the target signal. Any remaining energy was considered as distortion.

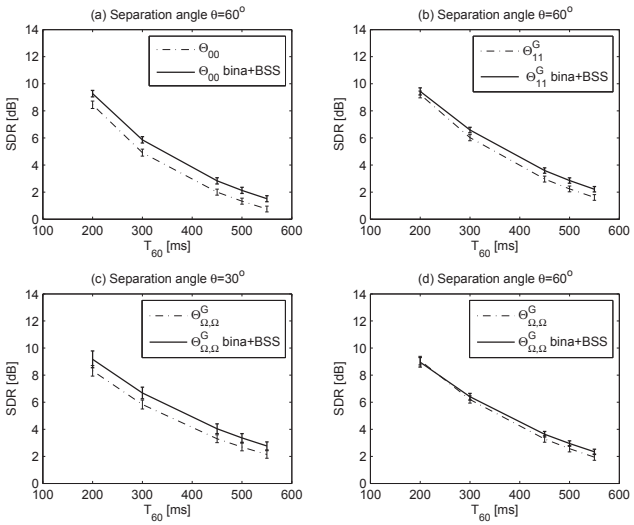

Fig. 2. The SDR improvement by the proposed method (bina+BSS) denothed by solid lines, over the method in [5] denothed by dot-dashed lines at different T_{60}s for (*a*) Θ_{00}, $\theta = 60^o$, (*b*) Θ^G_{11}, $\theta = 60^o$, (*c*) $\Theta^G_{\Omega,\Omega}$, $\theta = 30^o$, and (*d*) $\Theta^G_{\Omega,\Omega}$, $\theta = 60^o$.

As shown in figure 2 and table 1, the performance of the proposed algorithm (bina+BSS) is consistently better than the algorithm in [5] in which only the binaural cues are used. It can be seen that the improvement is quite considerable at simpler modes, but still exists even for the most complex mode $\Theta^G_{\Omega\Omega}$. Figure 2 also illustrates that the improvement becomes more significant as the reverberation time, T_{60} increases. This can be explained by the fact that the ILD and so its contribution reduces at higher reverberation and so the mixing vectors provide more distinct information, having more effect on the results.

6. CONCLUSION

This paper has presented a method to combine binaural cues and BSS approaches to classify the T-F units in the spectrogram of the mixtures. The proposed method improves the SDR of the separated signals consistently compared to a similar method with only binaural cues. Although the results are for the mixtures of two speakers, future work can be extended for more sources with only two microphones.

7. REFERENCES

[1] A. Hyvarinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural Network*, vol. 13, pp. 411–430, March 2000.

[2] S. Makino, H. Sawada, R. Mukai, and S. Araki, "Blind source separation of convolutive mixtures of speech in frequency domain," *IEICE Transactions*, vol. E88, no. 7, pp. 1640–1655, July 2005.

[3] H. Sawada, R. Mukai, S. Araki, and S.Makino, "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 5, pp. 530–538, September 2004.

[4] D. L. Wang and G. J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*, Wiley IEEE press, 2006.

[5] M. I. Mandel, R. J. Weiss, and D. P. W. Ellis, "Model-based expectation-maximization source separation and localization," *IEEE Transasctions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 382–394, February 2010.

[6] H. Sawada, S. Araki, and S. Makino, "A two-stage frequency-domain blind source separation method for underdetermined convolutive mixtures," *in IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, October 2007.

[7] P. Aarabi, "Self-localizing dynamic microphone arrays," *IEEE Trans. Syst., Man, Cybern. C*, vol. 32, pp. 474–484, November 2002.

[8] B. Shinn-Cunningham, N. Kopco, and T. Martin, "Localizing nearby sound sources in a classroom: Binaural room impulse responses," *J. Acoust. Soc. Amer.*, vol. 117, pp. 3100–3115, May 2005.

[9] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, July 2006.

Table 1. The SDRs of all the modes in dB for two different separation angles ($60^o, 75^o$) and $T_{60} = 0.3s, 0.45s$, each value is an average over 15 different mixtures.

modes	T_{60}=0.3 s, $\theta = 75^o$		T_{60}=0.3 s, $\theta = 60^o$		T_{60}=0.45 s, $\theta = 60^o$	
$\Theta_{ild,ipd}$	binaural	bina+bss	binaural	bina+bss	binaural	bina+bss
Θ_{00}	6.09	6.64	4.91	5.86	2.00	2.83
Θ_{01}	5.78	6.69	4.71	5.90	1.88	2.85
$\Theta_{0\Omega}$	6.18	6.69	5.29	5.96	2.30	2.92
Θ_{10}	6.12	6.51	5.38	6.12	2.43	3.09
$\Theta_{\Omega 0}$	6.14	6.50	5.38	6.07	2.48	3.12
Θ_{11}	5.73	6.56	5.24	6.16	2.38	3.12
Θ^G_{11}	6.40	**6.80**	6.02	**6.58**	2.96	3.59
$\Theta_{\Omega\Omega}$	4.73	6.59	4.66	6.15	2.36	3.29
$\Theta^G_{\Omega\Omega}$	**6.53**	6.79	**6.20**	6.40	**3.29**	**3.64**