SPATIAL AND COHERENCE CUES BASED TIME-FREQUENCY MASKING FOR BINAURAL REVERBERANT SPEECH SEPARATION

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

Most of the binaural source separation algorithms only consider the dissimilarities between the recorded mixtures such as interaural phase and level differences (IPD, ILD) to classify and assign the time-frequency (T-F) regions of the mixture spectrograms to each source. However, in this paper we show that the coherence between the left and right recordings can provide extra information to label the T-F units from the sources. This also reduces the effect of reverberation which contains random reflections from different directions showing low correlation between the sensors. Our algorithm assigns the T-F regions into original sources based on weighted combination of IPD, ILD, the mixing vector models and the estimated interaural coherence (IC) between the left and right recordings. The binaural room impulse responses measured in four rooms with various acoustic conditions have been used to evaluate the performance of the proposed method which shows an average improvement of more than 2.23 dB in signal-to-distortion ratio (SDR) in room D with $T_{60} = 0.89$ s over the state-of-the-art algorithms.

Index Terms— Precedence effect, binaural cues, blind source separation

1. INTRODUCTION

In real situations where the recording areas are surrounded by reflecting surfaces, the microphones record not only the direct signals from the sources but also the reflections from the walls, ceiling and other materials in the room. These reflections sustain the energy of the signals for a while resulting in reverberation. Although this effect helps in the estimation of distance, it degrades the performance of source separation and localization algorithms by smearing the energy in the time-frequency spectrogram of the recorded signals. Therefore, it has been of great interest to reduce the effect of reverberation on localization and separation algorithms.

The human hearing system tends to give more weight to the first arriving sound and suppress delayed signals due to reflections in a process known as the *precedence effect* [1, 2]. Approaches have been proposed to model this effect with different implementations [3–6]. The main task is to detect the areas in the T-F domain dominated by the direct sound, which are usually identified by a sudden increase of energy in onsets. The reverberant signals are typically diffuse, hence they are not correlated between the sufficiently spaced sensors [7], whereas the direct signals originated from the same source are coherent. Therefore, the interaural coherence (IC) can be employed to recognize the T-F units dominated by the direct signals [4]. In [4] the IC is used to create a binary mask and consider only the T-F units with a high coherence for interaural time difference (ITD) and ILD estimation so ignoring a large proportion of the input signals. In our approach, we however consider all the T-F units and apply the IC as another cue to generate a soft mask and weight the T-F regions.

The baseline separation algorithm is a modified version of our previous work [8] which combines the IPD, ILD and mixing vectors to estimate the likelihood of each T-F unit being dominated by each source. To improve the performance of that algorithm we control the contribution of each cue to the final decision by giving different weights to their log-likelihood. Although this shows improvement over the two state-of-the-art algorithms [9, 10], the reverberation is still the major effect on the performance degradation that needs to be addressed. This motivates us to adopt IC as a new cue to improve our separation algorithm.

The following section introduces the binaural feature extracted from the short-time Fourier transform (STFT) of the left and right signals. These cues are then used to assign the T-F units of the mixtures to the sources as explained in Section 3. The experiments and results are reported in Section 4. Section 5 relates the proposed work to the literature. Section 6 draws the conclusion and provides suggestions for the future work.

2. FEATURE EXTRACTION

A binaural recording contains two signals received at the left and right ears, $l(n)$ and $r(n)$, where n is the discrete time

Thanks to CVSSP for funding A. Alinaghi.

index. Each recording is a fusion of filtered source signals with additive or reverberant noise:

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

where N, known *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 with head related transfer function (HRTF), respectively; $n_l(n)$ and $n_r(n)$ are the background noise. The STFT of the left and right signals can be computed and then compared for the estimation of the various binaural cues as explained in the following sections.

2.1. Interaural phase and level difference

The interaural STFT, i.e. the ratio of the left and right STFT, 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, which can be modeled by Gaussian distribution and Gaussian mixture models (GMM), respectively [8].

2.2. Mixing vector estimation

In the sparse domain where only one source is active (or dominates) at each T-F unit, the mixing vector of each source (say ith) to the microphones, $\mathbf{a}_i(\omega)$, can be estimated based on the normalized observation vectors. To do so, the left and right signals at each T-F unit are put together to form the two dimensional vectors, $\mathbf{x}(\omega,t) = [L(\omega,t), R(\omega,t)]^T$, which are then normalized to remove the effect of source variations:

$$\mathbf{x}(\omega,t) \leftarrow \frac{\mathbf{x}(\omega,t)}{\sqrt{|L|^2+|R|^2}} = \frac{\mathbf{a}_i(\omega)}{||\mathbf{a}_i(\omega)||} \cdot \frac{s_i(\omega,t)}{|s_i(\omega,t)|}. \quad (3)$$

The normalized vectors are then divided into N clusters with the centroid of each cluster representing the mixing vector of the corresponding source [10].

2.3. Interaural Coherence

In addition to IPD and ILD which are measures of dissimilarities between the left and right signals, the similarity between them can also be measured by interaural coherence (IC) which is usually estimated from the normalized cross-correlation function in the time domain [4]. This is an implementation of the precedence effect to identify the T-F units dominated by direct signals and give more weight to them. However, as we work in the frequency domain, the coherence between the two signals l and r is defined as :

$$\Gamma_{l,r}(\omega,t) = \frac{\Phi_{l,r}(\omega,t)}{\sqrt{\Phi_{l,l}(\omega,t) \cdot \Phi_{r,r}(\omega,t)}} \quad (4)$$

where $\Phi_{l,l}(\omega)$ and $\Phi_{r,r}(\omega)$ represent the auto-power spectral densities (APSD) of l and r, respectively. $\Phi_{l,r}(\omega)$ is cross-power spectral density (CPSD) of the two time-aligned input channels. These densities are calculated by means of a recursive periodogram approach as introduced in [5]:

$$\hat{\Phi}_{l,l}(\omega,t) = \beta\hat{\Phi}_{l,l}(\omega,t-1) + (1-\beta)|L(\omega,t)|^2 \quad (5)$$
$$\hat{\Phi}_{l,r}(\omega,t) = \beta\hat{\Phi}_{l,r}(\omega,t-1) + (1-\beta)L(\omega,t) \cdot R^*(\omega,t) \quad (6)$$

with the smoothing factor $0 \leq \beta \leq 1$ which is set to 0.5 in our experiment to achieve moderate smoothing. $\hat{\Phi}_{r,r}(\omega,t)$ can be estimated similar to $\hat{\Phi}_{l,l}(\omega,t)$ using the $|R(\omega,t)|^2$. This coherence $\Gamma_{l,r}(\omega,t)$ will be almost 1 for T-F units dominated by the source positioned at 0° with coherent left and right recordings, while it will reduce for T-F regions containing more energy from random reverberations and other sources in different azimuths due to time delays. For sources at other azimuths the left signal is shifted by estimated ITD to compensate for the time delay (time-alignment) and a high IC will be obtained at units dominated by that source. Therefore, it can be considered as a soft mask which gives more weight to the target signal and reduces the energy from reverberation and interfering sources in other positions.

3. PROBABILISTIC T-F ASSIGNMENT WITH EM ALGORITHM

According to the preceding section, four different features can be extracted at each T-F unit, $\alpha(\omega,t)$, i.e. the ILD, $\phi(\omega,t)$, i.e. the IPD, $\mathbf{x}(\omega,t)$, i.e. the observation vector, and $\Gamma(\omega,t)$, i.e. the IC. The three former cues can be represented using parametric models such as Gaussian distribution with the parameters being estimated based on the maximum likelihood criterion:

$$L(\hat{\Theta}) = \max_{\Theta} \sum_{\omega,t} \log p(\phi(\omega,t), \alpha(\omega,t), \mathbf{x}(\omega,t)|\Theta), \quad (7)$$

$$\hat{\Theta} = \{\xi_{i,\tau}(\omega), \sigma_{i,\tau}(\omega), \mu_i(\omega), \eta_i(\omega), \mathbf{a}_i(\omega), \gamma_i(\omega), \psi_{i,\tau}\}$$

and $\xi_{i,\tau}, \sigma^2_{i,\tau}$, μ_i, η^2_i, $\mathbf{a}_i$, and γ^2_i are the mean and variance of the IPDs, the ILDs and the mixing vectors, respectively. $\psi_{i,\tau}$ is the mixture coefficient for the GMM, representing the probability of each source (say i) over $\tau = [-15:15]$ covering the azimuths from -90° to 90°. Once the underlying parameters are estimated using the expectation maximization

(EM) algorithm, the probability of each T-F unit belonging to each source can be calculated:

$$\begin{aligned}\nu_{i,\tau}(\omega,t) \quad \propto \quad & \Gamma_i(\omega,t)\psi_{i,\tau}.\mathcal{N}(\hat{\phi}(\omega,t;\tau)|\xi_{i,\tau}(\omega),\sigma^2_{i,\tau}(\omega)).\\ & \mathcal{N}(\alpha(\omega,t)|\mu_i(\omega),\eta_i^2(\omega)).\\ & \mathcal{N}(\mathbf{x}(\omega,t)|\mathbf{a}_i(\omega),\gamma_i^2(\omega)) \qquad (8)\end{aligned}$$

where $M_i(\omega,t) = \sum_\tau \nu_{i,\tau}$ is the occupation likelihood and applied to the mixture as a soft mask to extract the source signals. The interaural coherence (IC), $\Gamma_i(\omega,t)$, is calculated using the left and right mixtures after being time aligned, $\hat{L}_i$ and $\hat{R}_i$, based on the ITD of the ith source using the PHAT histogram [11] in equations (4), (5) and (6). In the case of target signal at 0° no time alignment is needed.

It is observed that these cues are not equally reliable especially in the presence of reverberation [9]. For example, the IPD cue tends to be more robust in reverberant conditions compared to ILD. Therefore, it is more realistic to adjust the contribution of the cues by giving a different weight to each cue before combining them. As opposed to [8] where the cues are combined with equal weight, we can introduce different weights to the cues to adjust the contribution of the cues in order to improve the probability estimation:

$$\begin{aligned}\log(\nu) \propto W_P\cdot\log\psi p(\hat{\phi}|\xi,\sigma^2) + W_L\cdot\log p(\alpha|\mu,\eta^2) + \\ W_B\cdot\log p(\mathbf{x}|\mathbf{a},\gamma^2) + W_C\cdot\log\Gamma(\omega,t) \qquad (9)\end{aligned}$$

where W_P, W_L, W_B and W_C control the influence of IPD, ILD, basis vector and IC cues, respectively. The optimum weighting coefficients have been estimated empirically based on extensive tests and set typically as $W_P = 0.8$, $W_L = 0.1$, $W_B = 0.5$, and $W_C = 1.0$ in experiments described in the next section. Here, we investigated the weights which are fixed over time and frequency. However, based on Duplex theory [12] it is expected that the ILD cue is more reliable in high frequencies while IPD is more robust in low frequency range. Therefore, introducing frequency dependent weighting to these cues may further improve the performance of the proposed algorithm, which we leave to our future study.

To recover each (say ith) source signal the corresponding occupation likelihood, $M_i(\omega,t)$, is multiplied by the mixture STFT and then transfered back to the time domain using the inverse short time Fourier transform (ISTFT).

4. EXPERIMENTS AND RESULTS

Similar to [9], we chose the TIMIT data set which is a continuous speech corpus containing 6300 utterances spoken by 630 native American English speakers [13]. 15 utterances, spoken by both male and female speakers, were selected randomly with approximately the same length (about 3 s), and then shortened to 2.5 s for consistency. The two common sentences spoken by all speakers (sa1 and sa2) were removed from the selection set to avoid mixtures containing identical word sequences, which would violate the assumption of sparsity and be unlikely from a practical perspective. All the utterances were also normalized to have equal root mean square amplitude.

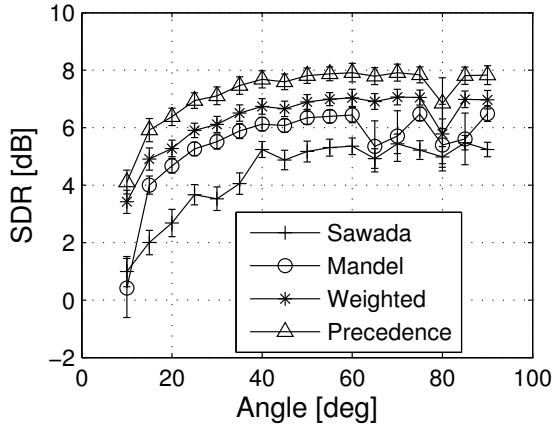
The binaural room impulse responses (BRIR) measured by Hummersone [3] were used to generate the mixtures. These were recorded using a dummy head and torso in four different types of room, named as A, B, C and D at the University of Surrey. Table 1 shows the acoustical properties of the rooms in which the signals were recorded. The HRTF is incorporated in the BRIR which makes the signals similar to what a person would hear in that position. For each T_{60} and configuration, 15 pairs from those 15 selected utterances were chosen in such a way that no signal would be mixed with itself. The mixtures were then generated by simply adding the reverberant target and interferer signals which is equivalent to assuming superposition of their respective sound fields. The target source was always located at the zero azimuth while the interferer's azimuth varied from 10° to 90° with steps of 5°, 1.5 m away from the head (this defines 17 different configurations). This is an ecologically valid approach to investigating the effect of target-interferer angular displacement on the system performance, given that we typically turn to face the target [14]. However, this was not known as a priori to the algorithms as not only the target sources but also the other sources are localized and recovered.

Table 1. Room acoustical properties in initial time delay gap (ITDG), direct-to-reverberant ratio (DRR) and reverberation time T_{60} [3].

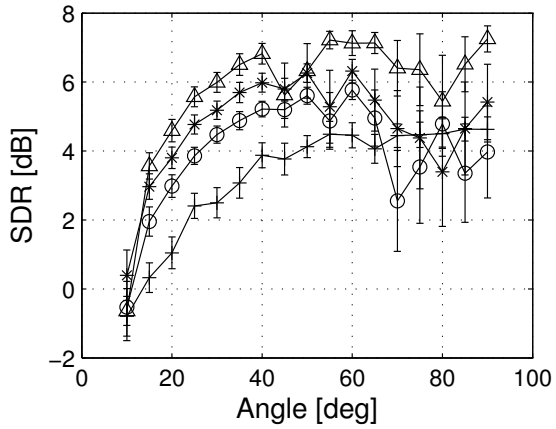
Room	ITDG [ms]	DRR [dB]	T_{60} [s]
A	8.72	6.09	0.32
B	9.66	5.31	0.47
C	11.9	8.82	0.68
D	21.6	6.12	0.89

The performance of the baseline and the proposed algorithms is evaluated based on the signal-to-distortion-ratio (SDR) [15]. Sawada's method [10] exploits only the mixing vectors while Mandel's algorithm [9] is based on IPD and ILD cues. Our approach starts with uniform combination of the cues and evolves to the weighted combination and then incorporates the precedence effect. Table 2 represents the performance results of the mentioned methods in four different rooms with various acoustical properties. It can be seen that the introduction of weighting cues has improved the combined algorithm [8] by about 0.3 dB for all conditions. It is also clear that the incorporation of the precedence effect has boosted the quality of the recovered signals especially in room D with more than 1 dB compared to the adjustable cue technique. This matches up with our expectation as the algorithm is designed to tackle the reverberation effect which is most severe in room D with $T_{60} = 0.89$ s. Overall, we achieved an average of 2.34 and 2.12 dB improvement over [9] and [10]

for determined (2 source) and under-determined (3 source) mixtures in room D, respectively. Figure 1 illustrates the SDRs of the separated targets in room D with the interfering sources positioned at different angular distances. The large variances for three source cases are due to the invalid ITD initialization which will be addressed in our future work.



(a) 2 source mixtures



(b) 3 source mixtures

Fig. 1. SDR of the recovered target source from (a) determined (2src) and (b) underdetermined (3src) mixtures in room D with $T_{60} = 0.89s$ with interfering sources positioned at different azimuth angles.

5. RELATION TO PRIOR WORK

In this algorithm we exploit and combine the cues introduced in [9, 10] in a controlled way by giving a different weight (credit) to each cue and improve the performance of uniform combination of the cues in [8].

The proposed algorithm here can also be considered as a joint source separation and dereverberation as in [16]. However, in [16] it is assumed that the number of sources is less than that of microphones whereas our algorithm is based on binaural recordings with two or more sources generating even-determined or under-determined mixtures.

Table 2. Results of baseline methods, the proposed method without weighting ($W_P = W_L = W_B = W_C = 1$) and with weighting ($W_P = 0.8, W_L = 0.1, W_B = 0.5, W_C = 1$) and with precedence effect ($\beta = 0.5$) for reverberant mixtures, with the average over A, B, C and D in SDR [dB].

Case	Methods	A	B	C	D	Mean
2-Src	Sawada	9.11	6.19	8.63	4.36	7.07
	Mandel	10.14	7.10	9.51	5.42	8.04
	Unweighted	10.65	7.27	9.79	5.93	8.41
	Weighted	10.80	7.61	10.05	6.31	8.69
	Precedence	**10.81**	**8.13**	**10.24**	**7.23**	**9.11**
3-Src	Sawada	6.43	4.13	6.03	3.30	4.97
	Mandel	7.81	4.93	7.40	3.97	6.03
	Unweighted	8.31	5.21	7.69	4.20	6.35
	Weighted	8.49	5.52	8.03	4.73	6.69
	Precedence	**8.57**	**5.96**	**8.08**	**5.75**	**7.09**

The precedence effect is also applied and modeled as a low pass filter with an inhibitory gain and time constant to mitigate the reverberant energy from the signal and to improve the separation performance in [17]. However, the model parameters are required to be estimated based on the room properties which are not always known as *a priori*.

The authors in [5] introduce a model-based dereverberation algorithm which employs a simplified binaural coherence model. The model parameters are estimated based on a head diameter of $0.15 - 0.17$ m and known transfer functions between a point source and the microphones and so not applicable for a general condition.

In contrast, we utilized a recursive approach to estimate the coherence between the left and right signals at each T-F unit and calculate a gain function similar to [18] in which the gain is calculated and applied for dereverberation of a single source while we introduced time alignment based on different source ITDs and used the gains for source separation.

6. CONCLUSION AND FUTURE WORK

This paper has proposed a method for fusing various features extracted from binaural recordings to improve the source separation algorithms in reverberant conditions. The contribution of ILD, IPD and mixing vectors is controlled by adjustable weights to give more weight to more reliable cues. Moreover, the precedence effect has been incorporated to the algorithm to mitigate the degrading affect of reverberation. These modifications have boosted the results significantly. For future work we will examine frequency dependent weighting based on Duplex theory. The IC cue can also be fitted to parametric models with parameters being estimated by EM algorithm.

7. REFERENCES

[1] R. Y. Litovskya, H. S. Colburn, W. A. Yost, and S. J. Guzman, "The precedence effect," *J. Acoust. Soc. Amer.*, vol. 106, no. 4, pp. 1633–1654, Oct 1999.

[2] R. M. Stern, D. L. Wang, and G. Brown, *Computational Auditory Scene Analysis*, chapter Binaural Sound Localization, Wiley/IEEE Press, 2006.

[3] C. Hummersone, *A psychoacoustic engineering approach to machine sound source separation in reverberant environments*, Ph.D. thesis, Music and Sound Recording, University of Surrey, UK, 2011.

[4] C. Faller and J. Merimaa, "Sound localization in complex listening situations: Selection of binaural cues based on interaural coherence," *J. Acoust. Soc. Amer.*, vol. 116, no. 5, pp. 3075–3089, November 2004.

[5] M. Jeub, M. Schafer, T. Esch, and P. Vary, "Model-based dereverberation preserving binaural cues," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1732–1745, September 2010.

[6] K. D. Martin, "Echo suppression in a computational model of the precedence effect," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust.*, oct 1997, p. 4 pp.

[7] F. Jacobsen and T. Roisin, "The coherence of reverberant sound fields," *J. Acoust. Soc. Amer.*, vol. 108, no. 1, pp. 204–210, 2000.

[8] A. Alinaghi, W. Wang, and P. J. B. Jackson, "Integrating binaural cues and blind source separation method for separating reverberant speech mixtures," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, May 2011, pp. 209–212.

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

[10] H. Sawada, S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 3, pp. 516–527, March 2011.

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

[12] L. Rayleigh, "On our perception of sound direction," *Philos. Mag.*, vol. 13, no. 74, pp. 214–232, 1907.

[13] J. S. Garofalo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "The DARPA TIMIT acoustic-phonetic continuous speech corpus cdrom," Linguistic Data Consortium, 1993.

[14] H. Kim, J. Kim, K. Komatani, T. Ogata, , and H. G. Okuno, "Target speech detection and separation for humanoid robots in sparse dialogue with noisy home environments," in *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems*, September 2008.

[15] 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.

[16] T. Yoshioka, T. Nakatani, M. Miyoshi, and H.G. Okuno, "Blind separation and dereverberation of speech mixtures by joint optimization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 1, pp. 69 –84, jan. 2011.

[17] C. Hummersone, R. Manson, and T. Brookes, "Dynamic precedence effect modeling for source separation in reverberant environments," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1867–1871, September 2010.

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