Comparison between the Statistical cues in BSS techniques and Binaural cues in CASA approaches for reverberant speech separation

A Alinaghi, P JB Jackson, W Wang

Centre for Vision, Speech and Signal Processing (CVSSP), UK,
e-mail:{A.Alinaghi, P.Jackson, W.Wang}@surrey.ac.uk,

Abstract. Speech source separation has been of great interest for a long time, leading to two major approaches. One of them is based on statistical properties of the signals and mixing process known as blind source separation (BSS). The other approach named as computational auditory scene analysis (CASA) is inspired by human auditory system and exploits monaural and binaural cues. In this paper these two approaches are studied and compared in more depth.

1. Introduction

In real environments the microphones do not only receive the energy from the target source, but also from the reverberation and other interfering sources. Consequently, the recorded signals are mixtures of different sources which degrade the performance of hearing aids, automatic speech recognition (ASR) and many other communication systems. Therefore, it is desired to separate the source signals as an auditory front-end. However, in most cases the source signals and the mixing process are not known, introducing a blind problem.

There have been various methods suggested to perform blind source separation such as independent component analysis (ICA) [14,8] and beamforming [15] which need as many mixtures available as the number of sources and fail when the number of sources exceed the number of sensors. To deal with underdetermined cases, when the number of mixtures are less than that of sources, the signals are transformed into time-frequency (T-F) domain where the speech is sparse and the sources can be separated using T-F masks [17]. In sparse domain, where only one source is dominant, the mixing matrix in ICA algorithm [14] reduces to a mixing vector (MV) which can be estimated by clustering the observation vectors as in [11]. However, the BSS algorithm proposed in [11] degrades in high reverberation.

On the other hand, human auditory system with just two ears has shown great performance for source separation [5,4] which has been studied under the name of computation auditory scene analysis (CASA) [2,16]. It is found that different monaural and binaural cues such as pitch, interaural level difference (ILD) and interaural phase difference (IPD) can be estimated and exploited to generate corresponding T-F masks to identify the T-F units of the mixtures' spectrograms dominated by each source. Although binaural cues applied by [10] have shown significant improvement over other algorithms, their performance is poor where the sources are close to each other.

In this paper, we study the MVs as in [11] and the binaural cues as applied by [10] to investigate the strengths and weakness of each of them. We have found that MVs are estimated based on the mixture models with additive noise whereas the binaural cues are calculated using the mixture models with convolutive noise. Consequently, we have shown that the MVs models are less affected by additive noise while ILD and IPD models are more robust to reverberation. Moreover, we have compared the MV models with ILD and IPD models and presented that MV models are more distinct where the sources are close to each other whereas ILD and IPD models have more overlaps resulting in poor performance when the sources are not well away.

2. Complex Mixing Vector Representation

In BSS approach the noise-free instantaneous mixtures are modeled by $\mathbf{x} = A\mathbf{s}$, where $\mathbf{x}$ is the observation vector, A is the mixing matrix and $\mathbf{s}$ the source vector. Each column of A, $\mathbf{a}_i$, represents the basis vector (mixing vector) from the ith source to the microphones. In the sparse domain where only one source is active with other sources being almost zero, all the columns of the mixing matrix are multiplied by zeros except the one corresponding to the active source. As a result, each observation vector can be considered as a basis vector multiplying the dominant source magnitude. For example in the time domain where $s_1(t)$ is active while other sources are zero the following equation holds.

$$\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = \begin{bmatrix} a_{11} \\ a_{12} \end{bmatrix} s_1(t). \tag{1}$$

Therefore, the points on the scatter plot of $x_1(t)$ versus $x_2(t)$ would be along a line with the direction of $[a_{11}\ a_{12}]^T$.

In real environments such as reverberant rooms, the microphones record not only the direct signal, but also a reflected (i.e. filtered) version of the source along with some added noise:

$$x_1(t) = s_i(t) * h_{i1}(t) + n_1^a(t), \tag{2}$$

$$x_2(t) = s_i(t) * h_{i2}(t) + n_2^a(t), \tag{3}$$

where h_{ik} is the room impulse response from source i to microphone k with n_k^a being the additive noise as $k = 1, 2$. t is the

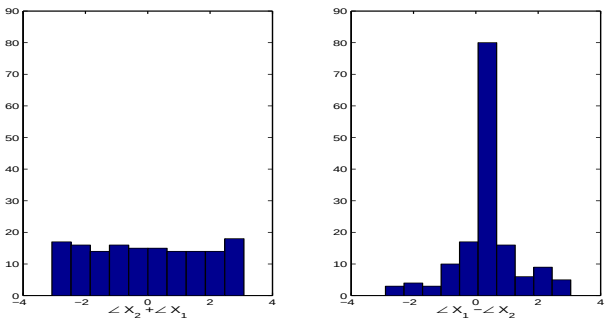


Figure 1. Histograms of $\angle X_1 + \angle X_2$ and $\angle X_1 - \angle X_2$ for T-F samples at frequency 2.35 kHz.

discrete time index and $*$ denotes convolution. In order to simplify the equation, the signals are transformed to T-F domain using short time Fourier transform (STFT):

$$X_1(m,f) \approx S_i(m,f) \cdot H_{i1}(f) + N_1^a(m,f), \quad (4)$$
$$X_2(m,f) \approx S_i(m,f) \cdot H_{i2}(f) + N_2^a(m,f). \quad (5)$$

These two signals are then concatenated to generate complex 2D vectors at each T-F unit, (m,f):

$$\mathbf{x}(m,f) \approx S_i(m,f)\mathbf{a}_i(f) + N^A(m,f), \quad (6)$$

where $\mathbf{x}(m,f) = [X_1(m,f), X_2(m,f)]^T$, is the complex 2D observation vector at each T-F unit, $\mathbf{a}_i(f) = [a_{i1}(f), a_{i2}(f)]^T \approx [H_{i1}(f), H_{i2}(f)]^T$ the basis (mixing) vector for the ith source and $N^A(m,f) = [N_1^a(m,f), N_2^a(m,f)]^T$ is the additive noise that contains background noise. To eliminate the effect of source amplitude variation, the observation vectors are normalized with respect to their magnitudes at each T-F unit,

$$\mathbf{x}(m,f) \leftarrow \frac{\mathbf{x}(m,f)}{\sqrt{(|X_1(m,f)|^2 + |X_2(m,f)|^2}}, \quad (7)$$
$$\mathbf{x}(m,f) \approx \frac{\mathbf{a}_i(f)}{||\mathbf{a}_i(f)||} \cdot \frac{S_i(m,f)}{||S_i(m,f)||}. \quad (8)$$

In the T-F domain with complex $2D$ vectors having amplitude and phase information, it is difficult to illustrate the same property as in 1. However, we will show that the observation vectors have less degrees of freedom after normalization in (8) which removes the effect of source amplitude:

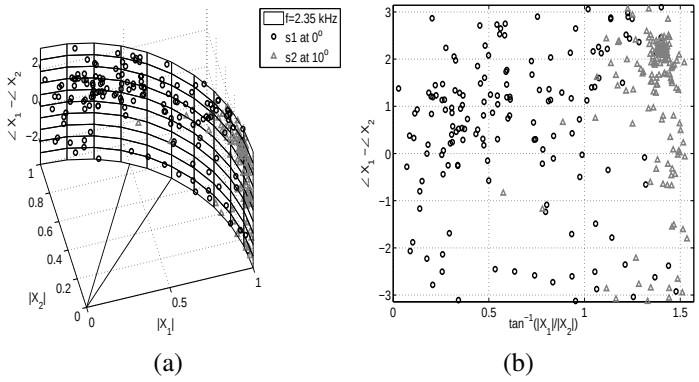
$$\begin{bmatrix} |X_1|e^{j\angle X_1} \\ |X_2|e^{j\angle X_2} \end{bmatrix} = \begin{bmatrix} a_{i1}(f) \\ a_{i2}(f) \end{bmatrix} \cdot \frac{S_i(m,f)}{||S_i(m,f)||}, \quad (9)$$

so at each frequency bin f:

$$|X_1|^2 + |X_2|^2 = 1 \quad (10)$$
$$\angle X_1 + \angle X_2 = \angle a_{i1} + \angle a_{i2} + 2\angle S_i(m), \quad (11)$$
$$\angle X_1 - \angle X_2 = \angle a_{i1} - \angle a_{i2}, \quad (12)$$

with $\angle X_1 + \angle X_2$ and $\angle X_1 - \angle X_2$ having uniform and normal distributions, respectively (see Fig. 1).

Moreover, as shown in (11), $\angle X_1 + \angle X_2$ is time variant due to the random phase of the source signal and therefore cannot



(a) (b)

Figure 2. 2D representation of the observation vectors in frequency channel = 2.35 kHz after normalization and whitening on a (a) unit cylinder wall, and (b) after unwrapping, for two different sources at 0° and 10° azimuths.

be applied to estimate the time-invariant basis vectors. Consequently, the mixing vectors (MV), $\mathbf{a}_i(f)$, which can be evaluated as the main eigenvector of the covariance matrices,

$$R_f = \sum_m^T \mathbf{x}(m,f)\mathbf{x}^H(m,f), \quad (13)$$

will have two degrees of freedom with $||\mathbf{a}_i|| = 1$ and $\angle a_{i1}$ or $\angle a_{i2} = 0$.

This result is consistent with the fact that the covariance matrices are positive-semidefinite and symmetric [3] and so Hermitian in complex domain with all the eigenvalues being real and simple [12]. Consequently, the eigenvectors (mixing vectors) will be like $[r\ cr]^T$ where $r \in \mathbb{R}$ and $c \in \mathbb{C}$, with relative phase and amplitude containing the whole information.

Fig. 2 (a) depicts the normalized and whitened observed samples at frequency channel = 2.35 kHz for two different sources positioned at 0° and 10°. To generate this scatter plot, two random utterances are chosen and convolved with binaural room impulse (BRIR) of room A (see Table 1) for sources at 0° and 10° azimuths independently. It can be seen that due

Table 1. Room acoustical properties in initial time delay gap (ITDG), direct–to-reverberant ratio (DRR) and reverberation time T_{60} [7]. .

Room	Type	ITDG [ms]	DRR [dB]	T_{60} [s]
A	a medium office	8.72	6.09	0.32
D	a large seminar room	21.6	6.12	0.89

to normalization all the points are confined to a unit cylinder which can be unwrapped to a 2D plane as shown in Fig. 2 (b).

3. Closely Spaced Sources

In order to compare the MV models with binaural cues, the model distributions of cues at extreme conditions (e.g. when the sources are close to each other) are illustrated using equal probability contours [13]. To calculate the probability distributions, one needs to estimate the model parameters assuming that the distributions are normal. The mean value of the MV models are estimated based on equation (13). The variance of

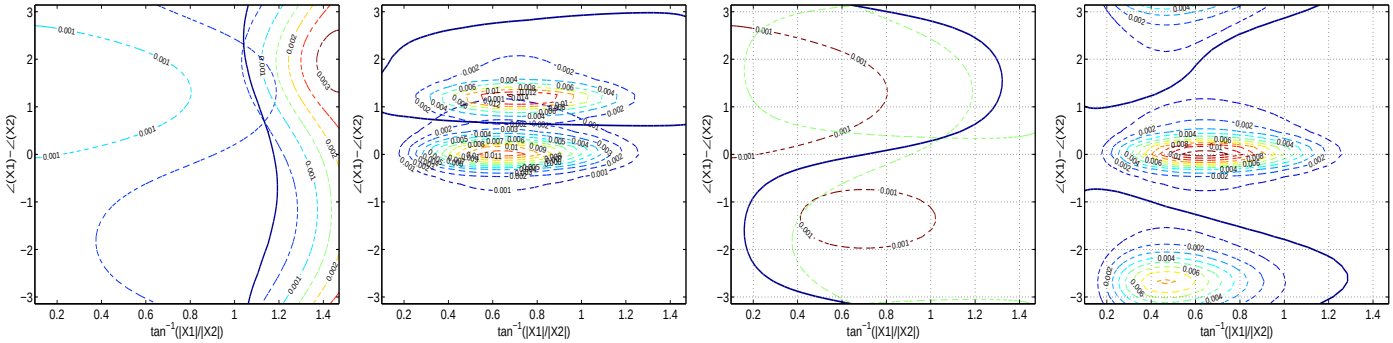


Figure 3. Equal probability contours for sources at 0° and 10° in dashed lines with decision boundary in solid line for (a) mixing vectors and (b) binaural cues in frequency = 2.35 kHz.

Figure 4. Equal probability contours for sources at 0° and 75° in dashed lines with decision boundary in solid line for (a) mixing vectors and (b) binaural cues in frequency = 2.35 kHz.

the MV of each source over time can be calculated according to equation (11) in [1] with posterior probability set to 1 as only one source is considered here. Once the model parameters of the sources at 0° and 10° azimuths have been obtained, $\angle X_1 - \angle X_2$ varies from $-\pi$ to π and $\tan^{-1}\frac{|X_1|}{|X_2|}$ from 0 to $\frac{\pi}{2}$ to cover all possible phase differences and level ratios. Then we set $X_2 = 1$ as reference and set the corresponding $\mathbf{x} = [X_1 X_2]^T$ in equation (6) in [1] to calculate the probabilities and plot the equal probability contours using equation *contour* in Matlab. The results are shown in Fig. 3 (a) which represents distinct distributions where the sources are positioned close to each other. The ILD and IPD model parameters are estimated based on the equations (18)-(22) in [10] with the probability $v_{i,\tau} = 1$. Once the parameters have been achieved, the joint probability of each source is calculated based on equations (7) and (27) in [10], where ϕ varies from $-\pi$ to π and α from 0.1 to 10. The resultant contours are close to each other and have overlaps as represented in Fig. 3 (b) which leads to missassigned T-F units. The decision boundaries draw a line to separate the phase-level values corresponding to sources at 0° and 10°. The MV models are well away from each other, reducing the false assignment of each T-F unit to a source, while binaural cue models show overlap which makes it difficult to decide if the observed phase or level value corresponds to the source at 0° or the source 10°. In other words, MV models are more robust where the sources are positioned near each other.

The same procedure has been followed for sources at 0° and 75° in room A. Fig. 4 represents the equal probability contours in dashed lines and decision boundaries in solid lines for the sources. It can be seen that when the sources are well away from each other with 75° difference in azimuths, binaural cues represent quite distinct source models whereas observation vectors have more overlap.

We also examined how distinct the source models are over the frequency ranges based on the Kullback–Leibler (KL) divergence [6] between the source models for two sources at 0° and 10° or 75° azimuths. As it is shown in Fig. 5 (a), MV based source models are well separated even when the sources are close to each other (10° azimuth) especially in frequency range $2-4$ kHz where ILD and IPD are not very reliable. However, as the sources are positioned away from each other

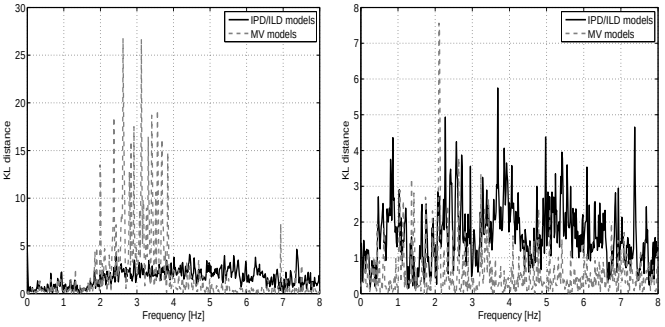


Figure 5. KL distance between the source models based on binaural cues and mixing vectors in room A with $T_{60} = 0.32$ s where one source is at 0° and the second source is at (a) 10° and (b) 80°.

(75° azimuth) the IPD/ILD source models become more distinct compared to those based on MVs (see Fig. 5 (b)). It shows that these models play complementary roles under different positioning.

4. High Reverberation

Next, we examine the effect of two types of noise on the cues. First, speech shaped noise is generated by averaging the spectra of the anechoic recordings of 15 utterances used in the experiments to be added to a clean signal similar to [9]. The clean signal was one of the utterances convolved with anechoic BRIR . The same utterance was also convolved with BRIR of room D (as in Table 1) to introduce convolutive noise. To measure the relative level of this convolutive noise we divided the room D's BRIR at 32 ms, which is also half of the window lengths (64 ms), and zero-padded each remaining part to have two RIRs representing desired early reflections and late noisy reverberation. The two parts were then convolved with the utterance and the relative energy of the signals was measured to be almost 5 dB for room D. Accordingly, we set the additive noise with SNR= 5 dB.

The model parameters of the source were then estimated under three different conditions: 1−anechoic room, 2−in anechoic room with additive noise, and 3−in high reverberant

room, to investigate the effect of additive and convolutive noise. The degradation from the original models is measured based on the KL distance [6] between the pdfs of the noisy observations and those corresponding to the clean anechoic signal.

Table 2. KL-distance between the clean and noisy signal models for three different cues and two types of noise averaged over all frequencies. .

-	MV	IPD	ILD
additive noise	**2.10**	2.70	3.39
convolutive noise	2.31	**2.01**	**3.29**

The results are shown in Table 2. Inspecting the Table 2 it is clear that MV model is more affected by high reverberation with higher KL distance (2.31) compared to (2.10) due to the same level of additive noise. On the other hand, binaural cues, and especially IPD with KL= 2.01, are more robust to reverberation but more sensitive to additive noise with KL= 2.70, playing complementary roles for dealing with different types of noise.

Moreover, we can see that MV and IPD are more reliable compared to ILD with less deviation from the original models, exhibiting smaller KL distances.

5. Conclusion

In this paper the mixing vectors (MV) applied by blind source separation (BSS) techniques have been examined in great detail. We have shown that where the sources are close to each other, the MV models are more distinct compared to binaural cues which overlap in these situations. However, when the sources are positioned away from each other, binaural cues represent distinguishable distributions while MVs show more overlap. Consequently, MVs and binaural cues can play complementary role to enhance the performance of source separation techniques.

In addition, we have examined the influence of additive and convolutive noise on the source models and shown that MVs are more robust to additive noise with less deviation whereas binaural cues and specifically IPD are more robust to reverberation.

Acknowledgements

Thanks to CVSSP for funding Atiyeh Alinaghi. This work was sponsored in part by the EPSRC of the U.K., grant numbers EP/H012842/1 and EP/H050000/1.

References

[1] Atiyeh Alinaghi, Wenwu Wang, and Philip 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)*, pages 209–212, May 2011.

[2] A. S. Bregman. *Auditory Scene Analysis: The Perceptual Organization of Sound*. Massachsettes Institute of Technology, 1994.

[3] W.K. Härdle and L. Simar. *Applied Multivariate Statistical Analysis*. Springer Berlin Heidelberg, 2012.

[4] W. M. Hartmann. How we localize sound. *Physics Today*, Nov 1999.

[5] Monica L. Hawley, Ruth Y. Litovsky, and H. Steven Colburn. Speech intelligibility and localization in a multi-source environment. *J. Acoust. Soc. Amer.*, 105(6):3436–3448, June 1999.

[6] J.R. Hershey and P.A. Olsen. Approximating the kullback leibler divergence between gaussian mixture models. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, volume 4, pages 317–320, 2007.

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

[8] A. Hyvarinen and E. Oja. Independent component analysis: algorithms and applications. *Neural Networks*, 13(4-5):411–430, June 2000.

[9] M. I. Mandel and D. P. W. Ellis. A probability model for interaural phase difference. In *ISCA Workshop Statist. Percept. Audio Process. (SAPA)*, pages 1–6, 2006.

[10] 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.*, 18(2):382–394, February 2010.

[11] Hiroshi Sawada, Shoko Araki, and Shoji Makino. Underdetermined convolutive blind source separation via frequency binwise clustering and permutation alignment. *IEEE Trans. Audio, Speech, Lang. Process.*, 19(3):516–527, March 2011.

[12] D. Serre. *Matrices: Theory and Applications*. Graduate Texts in Mathematics. Springer, 2002.

[13] A. Spanos. *Probability Theory and Statistical Inference: Econometric Modeling with Observational Data*. Cambridge University Press, 1999.

[14] J. V. Stone. *Independent Component Analysis, A tutorial introduction*. Massachsettes Institue of Technology, 2004.

[15] B.D. Van Veen and K.M. Buckley. Beamforming: a versatile approach to spatial filtering. *IEEE ASSP Magazine*, 5(2):4–24, 1988.

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

[17] Ozgur Ylmaz and Scott Rickard. Blind separation of speech mixtures via time-frequency masking. *IEEE Trans. Signal Process.*, 52(7):1830–1847, July 2003.