Use of Bimodal Coherence to Resolve Spectral Indeterminacy in Convolutive BSS

Qingju Liu, Wenwu Wang, and Philip Jackson

Centre for Vision, Speech and Signal Processing,
Faculty of Engineering and Physical Sciences,
University of Surrey, Guildford, GU2 7XH, United Kingdom
{Q.Liu,W.Wang,P.Jackson}@surrey.ac.uk

Abstract. Recent studies show that visual information contained in visual speech can be helpful for the performance enhancement of audio-only blind source separation (BSS) algorithms. Such information is exploited through the statistical characterisation of the coherence between the audio and visual speech using, e.g. a Gaussian mixture model (GMM). In this paper, we present two new contributions. An adapted expectation maximization (AEM) algorithm is proposed in the training process to model the audio-visual coherence upon the extracted features. The coherence is exploited to solve the permutation problem in the frequency domain using a new sorting scheme. We test our algorithm on the XM2VTS multimodal database. The experimental results show that our proposed algorithm outperforms traditional audio-only BSS.

Key words: convolutive blind source separation (BSS), audio-visual coherence, Gaussian mixture model (GMM), feature extraction and fusion, adapted expectation maximization (AEM), indeterminacy

1 Introduction

Human speech perception is essentially bimodal as speech is perceived by auditory and visual senses. In traditional blind source separation (BSS) for auditory mixtures, only audio signals are considered. With the independence assumption, many algorithms have been proposed, e.g. [1]-[4]. The use of visual stimuli in BSS represents a recent development in multi-modal signal processing. Sodoyer et al. [5] addressed the separation problem for an instantaneous stationary mixture of decorrelated sources, with no further assumptions on independence or non-Gaussianity. Wang et al. [6] implemented a similar idea by applying the Bayesian framework to the fused feature observations for both instantaneous and convolutive mixtures. Rivet et al. [7] proposed a new statistical tool utilizing the log-Rayleigh distribution for modeling the audio-visual coherence, and then used the coherence to address the permutation and scale ambiguities in the spectral

This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) (Grant number EP/H012842/1) and the MOD University Defence Research Centre on Signal Processing (UDRC).

domain. However, the algorithm proposed in [5] used simple visual stimuli with only plosive consonants and vowels and worked for only instantaneous mixtures; the method in [6] considered a convolutive model with a relatively small number of taps for the mixing filters; the approach in [7] trained the audio-visual coherence with high dimensional audio feature vectors, thus the coherence model was sensitive to outliers.

In this paper, we consider the convolutive model [6]-[11] with the assumption of non-Gaussianity and independence constraints of the sources. We synchronize and merge the modified Mel-frequency cepstrum coefficients (MFCCs) as audio features and some geometric-type features from the video stream to obtain the audio-visual features for the estimation of the parameters of the bimodal coherence. A GMM model is trained on the audio-visual features using the adapted expectation maximization (AEM) algorithm that considers the different influences of the audio features on the model. The audio-visual coherence is then applied to address the permutation indeterminacy in the frequency domain based on an iterative sorting scheme. The remainder of the paper is organised as follows. An overview of convolutive BSS is presented in Section 2. Section 3 introduces our bimodal feature extraction and fusion method. Detailed indeterminacies cancellation algorithm exploiting the audio-visual coherence is presented in Section 4. The simulation results are analyzed and discussed in Section 5. Finally Section 6 concludes the paper.

2 BSS for Convolutive Mixtures

BSS aims to recover sources from their mixtures without any or with little prior knowledge about the sources or the mixing process. Consider the convolutive model:

$$x_p(n) = \sum_{k=1}^{K} \sum_{m=0}^{+\infty} h_{pk}(m) s_k(n-m) + \xi_p(n), \qquad \text{for } p = 1, ..., P \tag{1}$$

or in matrix form: $\mathbf{x}(n) = \mathbf{H}(n) * \mathbf{s}(n) + \boldsymbol{\xi}(n)$, where $\mathbf{x}(n) = [x_1(n), ..., x_P(n)]^T$ are P observations obtained from K sources $\mathbf{s}(n) = [s_1(n), ..., s_K(n)]^T$ and $*$ denotes a convolution; $\mathbf{H}(n)$ is the mixing matrix whose entry $h_{pk}(n)$ represents the impulse response from source k to sensor p; $\boldsymbol{\xi}(n)$ is the additive noise vector; n is the discrete time index. The objective of convolutive BSS is to find a set of separation filters $\{w_{kp}(n)\}$ that satisfy:

$$\hat{s}_k(n) = y_k(n) = \sum_{p=1}^{P} \sum_{m=0}^{+\infty} w_{kp}(m) x_p(n-m). \qquad \text{for } k = 1, ..., K \tag{2}$$

The matrix form of the separation process is $\hat{\mathbf{s}}(n) = \mathbf{y}(n) = \mathbf{W}(n) * \mathbf{x}(n)$ where $\mathbf{W}(n)$ is the separation matrix whose entries are the impulse responses $w_{kp}(n)$.

Convolutive BSS is often performed in the frequency domain as depicted by the upper dashed box in Fig.1. After applying the short-time Fourier transform (STFT) to the observations, the convolutive mixture in the time domain is transformed to a set of instantaneous mixtures in the frequency domain. Then ICA is

applied to the spectral components $\mathbf{X}(f,t) = [X_1(f,t), ..., X_P(f,t)]^T$ in each frequency bin f to obtain the independent outputs $\mathbf{Y}(f,t) = [Y_1(f,t), ..., Y_K(f,t)]^T$, and t is the time-frame index. In matrix form $\mathbf{Y}(f,t) = \mathbf{W}(f)\mathbf{X}(f,t) = \hat{\mathbf{S}}(f,t)$, where $\mathbf{W}(f)$ is the separation filter, assumed to be linear time-invariant (LTI). It would be ideal if we could exactly recover the original sources that $\hat{\mathbf{S}}(f,t) = \mathbf{Y}(f,t) = \mathbf{S}(f,t)$. However, the ICA algorithms can estimate the sources only up to a permutation matrix $\mathbf{P}(f)$ and a diagonal matrix of gains $\mathbf{D}(f)$:

$$\hat{\mathbf{S}}(f,t) = \mathbf{Y}(f,t) = \mathbf{P}(f)\mathbf{D}(f)\mathbf{S}(f,t). \tag{3}$$

The permutation and scale ambiguities at each frequency bin present severe problems when reconstructing the separated sources in the time domain:

1. Recovered signal $Y_k(f,t)$ may not correspond to the same source $s_k(n)$ at some frequency bins, caused by $\mathbf{P}(f_i) \neq \mathbf{P}(f_j), i \neq j$.....**permutation indeterminacy**
2. Spectral components of $Y_k(f,t)$ coming from $s_k(n)$ are amplified at different frequency bins, caused by $\mathbf{D}(f_i) \neq \mathbf{D}(f_j), i \neq j$.....**scale indeterminacy**

To solve the permutation ambiguity, there are traditionally two methods. The first method is based on the continuity of adjacent bins, also known as the correlation approach [8]. The other method uses beam-forming theory [3] such as directional pattern estimation for permutation alignment [9]. As for the scale ambiguity, the minimum distortion principle can be applied to reduce the influence of a scale factor [10].

Previous algorithms using only audio streams have some drawbacks. The correlation approach may lead to continuous alignment errors; the beam-forming approach requires prior knowledge about the microphone array arrangement and some constraints on the spacing of microphones; the scale ambiguity is not perfectly solved. To solve the indeterminacies, as motivated by the work in [6] and [7], we use the bimodal coherence of audio-visual features described in detail in the next section (shown in Fig.1).

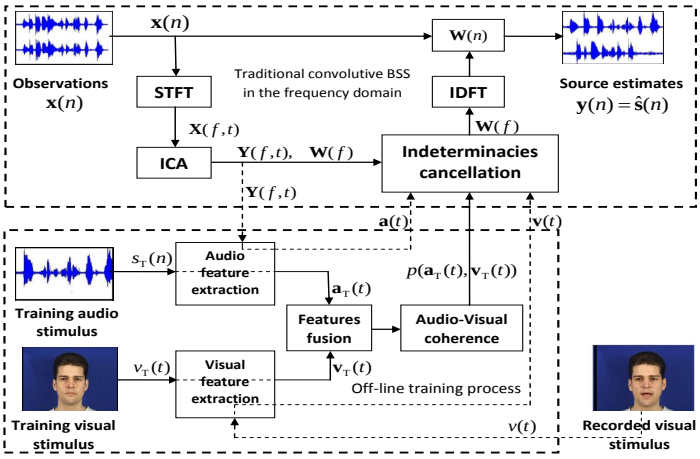

Fig.1. Flow of a typical audio-visual BSS system.

3 Feature Extraction and Fusion

3.1 Audio & Visual Feature Extraction

We take the Mel-frequency cepstrum coefficients (MFCCs) as audio features as in [6] with some modifications. The MFCCs exploit the non-linear resolution of the human auditory system across an audio spectrum, which are the Discrete Cosine Transform (DCT) results of the logarithm of the short term power spectrum on a Mel-scale frequency. To avoid inverse DFT to $\mathbf{Y}(f,t)$ in the separation process described in Section 4, we replace the first component of MFCCs with the logarithmic power of spectral data. We obtain the modified L-dimensional MFCCs $\mathbf{a_T}(t) = [\text{logE}(t), c_1(t)..., c_{L-1}(t)]^T$ (Fig.2). For simplicity, we denote the audio feature vector as $\mathbf{a_T}(t) = [a_{\mathbf{T}1}(t), ..., a_{\mathbf{T}L}(t)]^T$.

Unlike the appearance-based visual features used in [6], we use the same front geometric visual features as in [5][7]: the lip width (LW) and height (LH) from the internal labial contour. Fig.3 shows the method for obtaining the 2-dimensional visual feature vector $\mathbf{v_T}(t) = [\text{LW}(t), \text{LH}(t)]^T$.

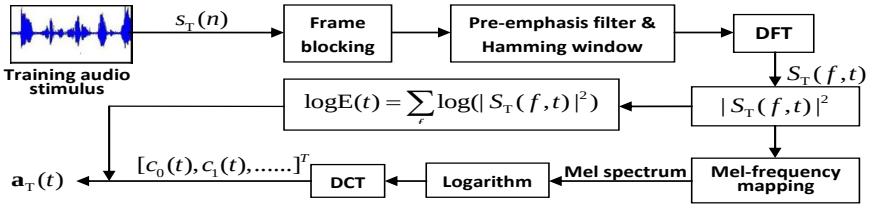

Fig.2. Audio feature extraction in the training process.

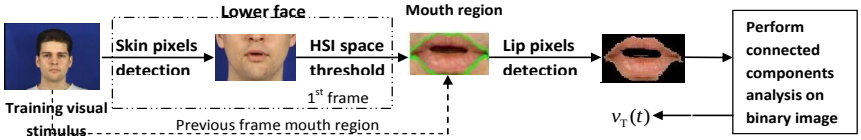

Fig.3. Visual feature extraction in the training process.

3.2 Feature-Level Fusion

We concatenate the audio and visual features after synchronization to get the $(L+2)$-dimensional audio-visual vector $\mathbf{u_T}(t) = [\mathbf{a_T}(t); \mathbf{v_T}(t)]$, which will be used for training. The audio-visual coherence can be statistically characterized as a GMM with I kernels:

$$p_{AV}(\mathbf{u_T}(t)) = \sum_{i=1}^{I} \gamma_i p_G(\mathbf{u_T}(t) \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \tag{4}$$

where γ_i is the weighting parameter, $\boldsymbol{\mu}_i$ is the mean vector and $\boldsymbol{\Sigma}_i$ is the covariance matrix of the i-th kernel. Every kernel of this mixture represents one cluster of the audio-visual data modeled by a joint Gaussian distribution:

$$p_G(\mathbf{u_T}(t) \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \frac{\exp\{-\frac{1}{2}(\mathbf{u_T}(t) - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i{}^{-1}(\mathbf{u_T}(t) - \boldsymbol{\mu}_i)\}}{\sqrt{(2\pi)^{L+2} \mid \boldsymbol{\Sigma}_i \mid}}. \tag{5}$$

We denote $\lambda_i = \{\gamma_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\}$ as the parameter set, and it can be estimated by the expectation maximization (EM) algorithm. In the traditional EM training process, all the components of the training data are treated equally whatever their magnitudes. Therefore if we train the data $\mathbf{u}_\dagger(t) = [\mathbf{a_T}(t); \mathbf{v_T}(t)]$ and $\mathbf{u}_\ddagger(t) = [2a_{\mathbf{T}1}(t), a_{\mathbf{T}2}(t), ...a_{\mathbf{T}L}(t), \mathbf{v_T}(t)^T]^T$ respectively, we get two joint distributions $p_{AV\dagger}(\cdot)$ and $p_{AV\ddagger}(\cdot)$ with two sets of parameters $\{\lambda_{i\dagger}\}$ and $\{\lambda_{i\ddagger}\}$. However, these joint distributions are identical:

$$p_{AV\dagger}(\mathbf{u}_\dagger(t)) = p_{AV\ddagger}(\mathbf{u}_\ddagger(t)). \tag{6}$$

Thus the influence of $a_{\mathbf{T}1}(t)$ on the final probability does not change even its magnitude is doubled. Nevertheless, some components of the audio vector with large magnitudes are actually more informative about the audio-visual coherence than the remaining components (consider, for instance, the case of lossy compression of audio and images using DCT where small components can be discarded). For example, the first component of the audio vector ($a_{\mathbf{T}1}(t)$) should play a more dominant role in affecting the probability $p_{AV}(\mathbf{u_T}(t))$ than the last one. Also, the components of the audio vector having very small magnitudes are likely to be affected by noise. Therefore, considering these factors, we propose an adapted expectation maximization (AEM) algorithm.

I. Initialize the parameter set $\{\lambda_i\}$ with the K-means algorithm.
II. Run the following iterative process:

i. Compute the influence parameters $\beta_i(\cdot)$ of $\mathbf{u_T}(t)$ for $i = 1, ...I$.

$$\beta_i(\mathbf{u_T}(t)) = 1 - \frac{\parallel \mathbf{u_T}(t) - \boldsymbol{\mu}_i \parallel}{\sum_{j=1}^{I} \parallel \mathbf{u_T}(t) - \boldsymbol{\mu}_j \parallel}, \tag{7}$$

where $\parallel \cdot \parallel$ denotes the squared Euclidean distance.

ii. Calculate the probability of each cluster given $\mathbf{u_T}(t)$.

$$p_i(\mathbf{u_T}(t)) = \frac{\gamma_i p_G(\mathbf{u_T}(t) \mid \boldsymbol{\mu}_i, \Sigma_i)\beta_i(\mathbf{u_T}(t))}{\sum_{j=1}^{I} \gamma_j p_G(\mathbf{u_T}(t) \mid \boldsymbol{\mu}_j, \Sigma_j)\beta_j(\mathbf{u_T}(t))}. \tag{8}$$

iii. Update the parameter set $\{\lambda_i\}$:

$$\boldsymbol{\mu}_i = \frac{\sum_t \mathbf{u_T}(t) p_i(\mathbf{u_T}(t))}{\sum_t p_i(\mathbf{u_T}(t))}, \gamma_i = \frac{\sum_t p_i(\mathbf{u_T}(t))}{\sum_t}, \Sigma_i = \frac{\sum_t (\mathbf{u_T}(t) - \boldsymbol{\mu}_i)^2 p_i(\mathbf{u_T}(t))}{\sum_t p_i(\mathbf{u_T}(t))}. \tag{9}$$

---

4 Resolution of Spectral Indeterminacy

As $y_k(n)$ is the estimate of $s_k(n)$, $y_k(n)$ will have maximum coherence with the corresponding video signal $v_k(t)$. Therefore we can maximize the following criterion in the frequency domain to address the indeterminacies as mentioned in Section 2:

$$[\hat{\mathbf{P}}(f), \hat{\mathbf{D}}(f)] = \arg\max_{\mathbf{P}(f), \mathbf{D}(f)} \sum_t \sum_{k=1}^{K} p_{AV}(\mathbf{u}_k(t)), \tag{10}$$

where $\mathbf{u}_k(t) = [\mathbf{a}_k(t); \mathbf{v}_k(t)]$ is the audio-visual feature extracted from the profile $\hat{S}_k(\cdot, t) = Y_k(\cdot, t)$ of the k-th source estimate and the recorded video associated

with the k-th speaker at time-frame t. If we are just interested in an estimate of $s_1(n)$ from the observations, we can get the first separation vector $\mathbf{p}(f)$ (note it is not the separation matrix) and the scale parameter $\alpha(f)$ by maximizing:

$$[\hat{\mathbf{p}}(f), \hat{\alpha}(f)] = \arg\max_{\mathbf{p}(f),\alpha(f)} \sum_t p_{AV}(\mathbf{u}_1(t)). \tag{11}$$

Since the permutation problem is the main factor in the degradation of the recovered sources, we focus on permutation indeterminacy cancellation for a two-source and two-mixture case detailed as follows. Suppose there are $2M$ frequency bins in total. Based on the symmetry, we will only consider the positive M bins. Denote $\mathbf{v}_1(t)$ as the visual feature that we have extracted from the recorded video signal associated with the target speaker. Generate an intermediate variable $Y_1^\dagger(f,t)$ spanning the same frequency and time-frame space as $Y_1(f,t)$ (or $Y_2(f,t)$). Initialize $\mathbf{P}(f)$ with identity matrices for $f = f_1, ..., f_M$.

I. Test which profile, $Y_1(\cdot,t)$ or $Y_2(\cdot,t)$, is more coherent with $\mathbf{v}_1(t)$.

1. For $f = f_1, ..., f_M$, let $Y_1^\dagger(f,\cdot) = Y_2(f,\cdot)$.
2. Extract the audio feature $\mathbf{a}_1(t)$ and $\mathbf{a}_1^\dagger(t)$ from $Y_1(\cdot,t)$ and $Y_1^\dagger(\cdot,t)$. Let $\mathbf{u}_1(t) = [\mathbf{a}_1(t); \mathbf{v}_1(t)]$, $\mathbf{u}_1^\dagger(t) = [\mathbf{a}_1^\dagger(t); \mathbf{v}_1(t)]$, and then calculate the audio-visual probability $p_{AV}(\mathbf{u}_1(t))$ and $p_{AV}(\mathbf{u}_1^\dagger(t))$ respectively based on the GMM model in equation (4) and the parameter set λ_i that has been estimated with the AEM algorithm.
3. If $\sum_t p_{AV}(\mathbf{u}_1(t)) > \sum_t p_{AV}(\mathbf{u}_1^\dagger(t))$, do nothing; otherwise, swap the rows of $\mathbf{P}(f)$ (i.e. $\mathbf{P}(f) \longleftarrow \left[\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right] \mathbf{P}(f)$), $\mathbf{W}(f)$ and $\mathbf{Y}(f,\cdot)$ for $f = f_1, ..., f_M$.

II. Equally divided M bins into 2 parts.

II.i. 1. For $f = f_1, ...f_{M/2}$, $Y_1^\dagger(f,\cdot) = Y_2(f,\cdot)$; for the remaining bins, $Y_1^\dagger(f,\cdot) = Y_1(f,\cdot)$.
2. Extract $\mathbf{u}_1(t)$ and $\mathbf{u}_1^\dagger(t)$, and then calculate $p_{AV}(\mathbf{u}_1(t))$ and $p_{AV}(\mathbf{u}_1^\dagger(t))$.
3. If $\sum_t p_{AV}(\mathbf{u}_1(t)) > \sum_t p_{AV}(\mathbf{u}_1^\dagger(t))$ do nothing; otherwise, update $\mathbf{P}(f)$, $\mathbf{W}(f)$ and $\mathbf{Y}(f,\cdot)$ as in step **I** for $f = f_1, ...f_{M/2}$.

II.ii. Repeat the replacement, calculation, comparison and update as in step **II.i** for $f = f_{M/2+1}, ...f_M$.

III. Divide M bins into 4 (8, 16,...) parts, and continue the progressive scheme.

This scheme can reach a high resolution, which is determined by the number of partitions at the final division, and the larger the number, the higher the resolution. However, most permutations occur continuously in practical situations, therefore even if we stop running the algorithm at a very 'coarse' resolution, the permutation ambiguity can still be well reduced.

The scale indeterminacy can be addressed by some gradient algorithms [4]. However, estimating the gradient of $\sum_t p_{AV}(\mathbf{u}_1(t))$ is computationally demanding, and it remains an issue in our future work.

5 Experimental Results

The proposed method was tested on the XM2VTS [12] multi-modal database, in which the speech data were recorded 4 times at approximately one month

intervals, with continuous sentences of words and digits in mono, 16 bit, 32 kHz, PCM wave files, and the frontal face videos were captured at 25 fps.

We trained the audio-visual coherence model of the target speaker with concatenated audio and visual speech signals lasting for approximately 40 seconds. The audio was downsampled to 16 kHz, and the 32 ms (512 points) Hamming window with 12 ms overlap was applied in STFT. 5-dimensional ($L = 5$) MFCCs as audio features were extracted from 24 mel-scaled filter banks. The visual features were upsampled to 50 Hz to be synchronized with the audio features. Thus the audio-visual data were 7-dimensional. For simplicity, we only used 5 ($I = 5$) kernels to approximate the audio-visual coherence.

The algorithm was tested on convolutive mixtures synthesized on computer. The filters $\{h_{pk}(n)\}$ were generated by the system utilizing the impulse response measurements of a conference room [13] with various positions of the microphones and the speakers. We resampled those filters and used the beginning 256 measurements (the reverberation time was 16 ms) as the final mixing filters. Two audio signals with each lasting 4 s were convolved with the filters to generate the mixtures, and Gaussian white noise (GWN) was added to both mixtures.

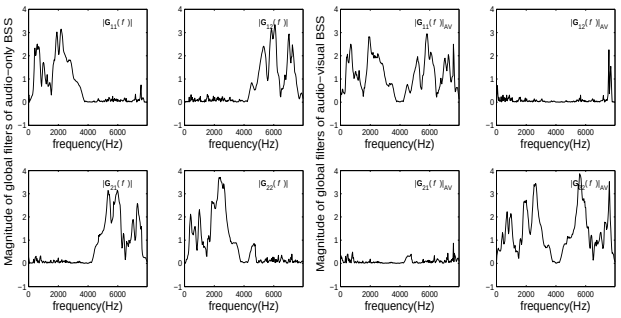

Fig.4. Global filters comparison.

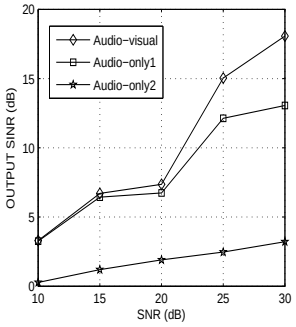

Fig.5. SINR comparison.

We use the global filters $\mathbf{G}(f) = \begin{bmatrix} \mathbf{G}_{11}(f) & \mathbf{G}_{12}(f) \\ \mathbf{G}_{21}(f) & \mathbf{G}_{22}(f) \end{bmatrix} = \mathbf{W}(f)\mathbf{H}(f)$ in the frequency domain and signal to interference and noise ratio (SINR) at different signal to noise ratios (SNRs) as criteria to evaluate the performance of our bimodal BSS algorithm. Suppose $s_1(n)$ is the target source, then

$$\mathrm{SINR} = 10 \log \frac{P_{s_1}}{P_{\hat{s}_1 - s_1}} = 10 \log \frac{\sum_n \| \sum_{p=1}^{P} w_{1p}(n) * h_{p1}(n) * s_1(n) \|}{\sum_n \| \hat{s}_1(n) - \sum_{p=1}^{P} w_{1p}(n) * h_{p1}(n) * s_1(n) \|}. \quad (12)$$

Fig.4 is the comparison of the global filters between the frequency-domain audio-only BSS using the correlation method [8] (left half) and audio-visual BSS (right half). It shows that our algorithm has corrected the permutation ambiguity at most frequency bins, while the permutation ambiguities in a large number of bins has not been resolved with the correlation method [8]. Fig.5 shows the SINR over different input SNRs. The SINR is calculated over a total of 100 independent runs with different convolutive filters. In the figure, Audio-only1 and Audio-only2 are two algorithms using only audio signals. Audio-only1 is a frequency-domain BSS algorithm, exploiting the correlation method [8]. Audio-only2 is a time-domain fast fixed-point BSS algorithm based on a

convolutive sphering process [11], and when the order of the mixing filters is high (e.g., 256 in our simulation), it may not converge. However, the ICA technique degrades in adverse conditions, and as a result the improvement of bimodal BSS disappears at low SNRs. When it is noise-free, the SINR of bimodal BSS is 21.9dB.

6 Conclusions

We have presented a new audio-visual convolutive BSS system. In this system, we have used the modified MFCCs as audio features, which were combined with geometric visual features to form an audio-visual feature space. An adapted EM algorithm is then proposed to exploit the different influences of the audio features on the statistically modeling of the audio-visual coherence. A new sorting scheme exploiting the audio-visual coherence to solve the spectral indeterminacy problem has also been presented. Our algorithm has been tested on the XM2VTS database and has shown improved performance over audio-only BSS systems. In the future, we will consider using some dynamic features in video as well, instead of just static features. In addition, we will increase the number of kernels to improve the accuracy of the audio-visual model.

References

1. Jutten, C., Herault, J.: Blind Separation of Sources, Part I: An Adaptive Algorithm Based on Neuromimetic Architecture. Signal Process. vol. 24, no. 1, pp. 1–10 (1991)
2. Comon, P.: Independent Component Analysis, a New Concept?. Signal Process., vol. 36, no. 3, pp. 287–314 (1994)
3. Cardoso, J.F., Souloumiac, A.: Blind Beamforming for Non-Gaussian Signals. IEEE Proc.-F, vol. 140, no. 6, pp. 362–370 (1993)
4. Hyvärinen, A., Karhunen, J. Oja. E.: Independent Component Analysis. John Wiley & Sons, New York (2001)
5. Sodoyer, D., Schwartz, J.L., Girin, L., Klinkisch, J., Jutten, C.: Separation of Audio-Visual Speech Sources: a New Approach Exploiting the Audio-Visual Coherence of Speech Stimuli. EURASIP J. Appl. Signal Process., vol. 11, pp. 1165–1173 (2002)
6. Wang, W., Cosker, D., Hicks, Y., Sanei, S., Chambers, J.: Video Assisted Speech Source Separation. In: Proc. IEEE ICASSP, pp. 425–428 (2005)
7. Rivet, B., Girin, L., Jutten, C.: Mixing Audiovisual Apeech Processing and Blind Source Separation for the Extraction of Speech Signals from Convolutive Mixtures. IEEE Trans. Audio Speech Lang. Process., vol. 15, no. 1, pp. 96–108 (2009)
8. Anemüller, J., Kollmeier, B.: Ampltude Modulation Decorrelation for Convolutive Blind Source Separation. In: Proc. ICA, pp. 215–220 (2000)
9. Ikram, M.Z., Morgan, D.R.: A Beamforming Approach to Permutation Alignment for Multichannel Frequency-Domain Blind Speech Separation. In: Proc. IEEE ICASSP, pp. 881–884 (2002)
10. Matsuoka, K., Nakashima, S.: Minimal Distortion Principle for Blind Source Separation. In: Proc. ICA, pp.722–727 (2001)
11. Thomas, J., Deville, Y., Hosseini, S.: Time-Domain Fast Fixed-Point Algorithms for Convolutive ICA. IEEE Signal Process. Lett. vol. 13, no. 4, pp. 228–231 (2006)
12. Messer, K., Matas, J., Kittler, J., Luettin, J., Maitre, G.: XM2VTSDB: The Extended M2VTS Database. In: AVBPA (1999). `http://www.ee.surrey.ac.uk/CVSSP/xm2vtsdb/`
13. Westner, A.: Room Impulse Responses (1998). `http://alumni.media.mit.edu/~westner/papers/ica99/node2.html`