Audio-visual Convolutive Blind Source Separation

Qingju Liu, Wenwu Wang, Philip Jackson

Centre for Vision, Speech and Signal Processing
University of Surrey, Guildford, UK
{Q.Liu,W.Wang,P.Jackson}@surrey.ac.uk

***Abstract*—We present a novel method for speech separation from their audio mixtures using the audio-visual coherence. It consists of two stages: in the off-line training process, we use the Gaussian mixture model to characterise statistically the audio-visual coherence with features obtained from the training set; at the separation stage, likelihood maximization is performed on the independent component analysis (ICA)-separated spectral components. To address the permutation and scaling indeterminacies of the frequency-domain blind source separation (BSS), a new sorting and rescaling scheme using the bimodal coherence is proposed. We tested our algorithm on the XM2VTS database, and the results show that our algorithm can address the permutation problem with high accuracy, and mitigate the scaling problem effectively.**

I. Introduction

Looking at the speaker's lips improves the intelligibility of human speech embedded in cocktail party noise [1] due to the contribution of the complementary visual information to the audio signal. The complementarity of visual and audio stimuli is often termed as the audio-visual coherence, which can be statistically approximated using mathematical techniques. Therefore, visual stimuli contain additional information about audio signals, and we can utilize the audio-visual coherence to assist separation of the source signals from their audio mixtures. This is known as audio-visual blind source separation (BSS), a recent development in multi-modal signal processing. Different from traditional BSS, where only audio signals are used [2]–[7], audio-visual BSS incorporates visual information into the separation process.

Wang et al. [9] implemented such a separation system by applying the Bayesian framework to the fused feature observations for both instantaneous and convolutive mixtures of decorrelated sources. Rivet et al. [10] 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 scaling ambiguities in the spectral domain. Casanovas et al. [13] built relationship between synchronous structures on both audio and visual modalities, to detect the audio sources activity and then built the audio models and separated the original soundtrack from only one microphone recording. However, the algorithm proposed in [9] considered a convolutive model with a relatively small number of taps for the mixing filters; the approach in [10] trained the audio-visual coherence with high dimensional audio feature vectors, thus the coherence model was sensitive to outliers. Cross-modal correlation was not exploited in the separation stage in [13], which used spectral masks from a pure audio point of view. The scaling ambiguity problem with the extracted source components is not addressed in [9] or [13].

We have implemented the similar effect in our previous works in [11] and [12] to resolve the spectral indeterminacies. In [11], we combined the Mel-scale frequency coefficients (MFCC) as audio features with some geometric visual features to form the audio-visual space, then we proposed an adapted expectation maximization (AEM) algorithm to train the audio-visual coherence, which was utilised to address the permutation problem. In [12], we changed the audio features with the filterbank analysis, and focused on mitigating the scaling ambiguity.

In this paper, we consider the convolutive model [4]–[10] with the assumption of non-Gaussianity and independence constraints of the sources, which relates to the real room acoustic mixture model. In the off-line training process, the power spectrum of the audio signals is mapped into Mel-scale filterbanks as the audio features; visual features are extracted from the training videos. We synchronize and merge the features to obtain the audio-visual data for the estimation of the parameters of the bimodal coherence characterised by the Gaussian mixture models (GMM). The audio-visual coherence is then applied to address the permutation and scaling indeterminacy in the frequency domain. The main contribution in this paper is the introduction of a new criterion for evaluating the confidence of the audio-visual coherence, which is used to reduce the influence of outliers on the cumulative log-likelihood.

The remainder of the paper is organised as follows. An overview of convolutive BSS is presented in Section II. Section III introduces our detailed training process to obtain the audio-visual coherence. Detailed indeterminacies cancellation algorithm exploiting the audio-visual coherence is presented in Section IV. The simulation results are analysed and discussed in Section V. Finally Section VI concludes the paper.

II. Convolutive Blind Source Separation

For convolutive BSS, the observation at each sensor is a sum of filtered source signals. The speech mixing process for a cocktail party scenario can be approximated with the convolutive model:

$$\begin{aligned} x_p(n) &= \sum_{k=1}^{K} \sum_{m=0}^{+\infty} h_{pk}(m) s_k(n-m) + \xi_p(n), \\ \mathbf{x}(n) &= \mathbf{H} * \mathbf{s}(n) + \boldsymbol{\xi}(n), \end{aligned} \tag{1}$$

where h_{pk} represents the room impulse response filter from source k to sensor p. We denote $\mathbf{x}(n) = [x_1(n), ..., x_P(n)]^T$ as the observation vector at the discrete time index n; $\mathbf{s}(n) = [s_1(n), ..., s_K(n)]^T$ the source vector and $\boldsymbol{\xi}(n)$ the additive noise vector; $\mathbf{H}$ the mixing matrix whose elements are filters h_{pk} and $*$ denotes convolution.

Convolutive BSS aims to find a set of separation filters $\{w_{kp}\}$ that satisfy:

$$\begin{aligned} \hat{s}_k(n) = y_k(n) &= \sum_{p=1}^{P} \sum_{m=0}^{+\infty} w_{kp}(m) x_p(n-m), \\ \hat{\mathbf{s}}(n) = \mathbf{y}(n) &= \mathbf{W} * \mathbf{x}(n), \end{aligned} \tag{2}$$

where $\mathbf{W}$ is the separation matrix whose entry w_{kp} is the impulse response filter from observation p to the estimate of source k.

Convolutive BSS can be directly performed in the time domain [8] by deconvolution, but the computational complexity is very high and sometimes it cannot guarantee the convergence to a global optimum, especially when the mixing filters have long taps. Based on the short-time stationarity of the speech signal and the linear time-invariance of the mixing process, an alternative is to perform BSS in the time-frequency domain by applying the short-time Fourier transform (STFT) to the observations. In each frequency bin f, we get an instantaneous mixing model:

$$\boldsymbol{X}(f,t) = \mathbf{H}(f)\boldsymbol{S}(f,t), \tag{3}$$

where $\boldsymbol{X}(f,t) = [X_1(f,t), ..., X_P(f,t)]^T$ and $\mathbf{H}(f)$ is the Fourier transform of the filter matrix $\mathbf{H}$.

ICA is applied separately in each frequency bin f to obtain the independent outputs $\boldsymbol{Y}(f,t) = [Y_1(f,t), ..., Y_K(f,t)]^T$, assumed to be the source estimates:

$$\boldsymbol{Y}(f,t) = \mathbf{W}(f)\boldsymbol{X}(f,t) = \hat{\boldsymbol{S}}(f,t). \tag{4}$$

However, the ICA algorithms can estimate the sources only up to a permutation matrix $\mathbf{P}(f)$ and a diagonal matrix of gains $\mathbf{D}(f)$:

$$\hat{\boldsymbol{S}}(f,t) = \boldsymbol{Y}(f,t) = \mathbf{P}(f)\mathbf{D}(f)\boldsymbol{S}(f,t). \tag{5}$$

These are the so-called permutation ($\mathbf{P}(f)$) and scaling ($\mathbf{D}(f)$) indeterminacy problems.

For the permutation problem, $Y_k(f,t)$ may correspond to different source signals at different frequency bins. Many algorithms have been proposed, with [9]–[11] or (most of the available algorithms are) without [5]–[7] the visual information. The methods in [9]–[11] use audio-visual coherence maximization to the alignment of the spectral components, the approach in [5] utilizes the continuity of the spectral components while [6] employs beamforming theory and [7] is a combination of the previous two algorithms. As for the scaling problem, $Y_k(f,t)$ is amplified with different scales at different frequency bins. The problem is addressed in [10] from the model variance point of view, [12] mitigates this problem for a high noise environment by the estimation of the audio spectrum distribution, and [7] uses a minimum distortion principle.

III. GMM Training

In the off-line training process, we use a GMM to approximate the joint probability of the audio-visual data $\mathbf{u_T}(t)$ extracted from the audio-visual stimuli used for training (denoted as $\mathbf{T}$).

$$p_{AV}(\mathbf{u_T}(t)) = \sum_{i=1}^{I} \gamma_i \mathcal{N}(\mathbf{u_T}(t) \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \tag{6}$$

where γ_i is the weighting parameter, $\boldsymbol{\mu}_i$ is the mean vector, $\boldsymbol{\Sigma}_i$ is the covariance matrix of the i-th kernel, and each kernel of this mixture represents one cluster of the audio-visual data modeled by a joint Gaussian distribution. To model the audio-visual correlation for each speaker, first we need to extract the audio-visual features, described as follows.

A. Feature Extraction

We exploit the non-linear resolution of the human auditory system across an audio spectrum using the Mel-scale filterbank analysis. We denote $\mathcal{F}_l$ as the group of the frequency bins spanned by the l-th filter. The mono power spectrum is mapped into these filters to abtain the L-dimensional audio feature $\mathbf{a_T}(t) = [a_{\mathbf{T}1}(t), ..., a_{\mathbf{T}L}(t)]^T$ for statistical training, where

$$a_{\mathbf{T}l}(t) = \log \sum_{f \in \mathcal{F}_l} b_l(f) |S_{\mathbf{T}}(f,t)|^2, \tag{7}$$

and $b_l(f)$ is the magnitude of the l-th filter while $S_{\mathbf{T}}(f,t)$ is the spectral component of the training audio. Figure 1 shows a typical speech signal and its audio features after the Mel-scale filterbank analysis.

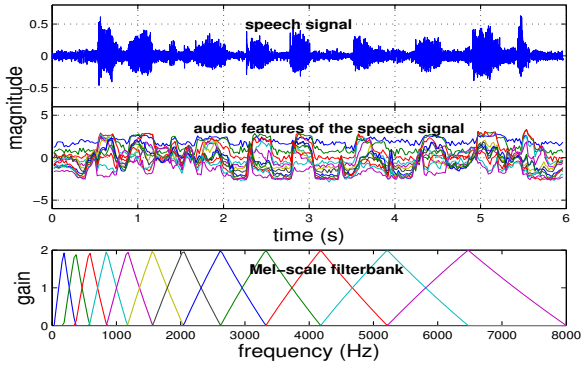

Fig. 1. The speech signal and its audio features obtained by the filterbank analysis.

For the visual features, first we crop an area from the video to get the gross mouth region. We then use snakes [14] to detect the mouth contour. Snakes, also called active contour model, is an energy mimization process to delineate an object outline. Then we relocate the mouth centre and extract a 64×96 mouth region based on the contour. Then the fast block discrete cosine transform (DCT) is employed on the mouth region to compress the image. Finally principal component analysis (PCA) is applied to the DCT data to get the visual feature vector $\mathbf{v_T}(t)$. In the experiment, we used 3 principal

components as visual features, which took up to 64.8% total variance. Figure 2 shows the detailed visual feature extraction process.

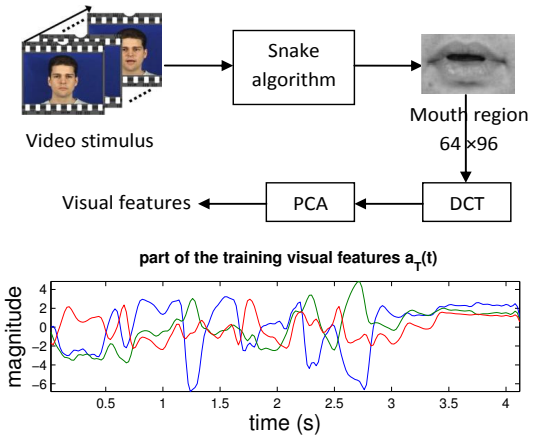

Fig. 2. The visual feature extraction process and part of the training visual features $\mathbf{a_T}(t)$.

B. Feature Fusion

Using the audio feature vector $\mathbf{a_T}(t) = [a_{\mathbf{T}1}(t), \cdots, a_{\mathbf{T}L}(t)]^T$ obtained by the L Mel-scale filterbank analysis, we synchronize and concatenate the visual feature vector with each element of the audio feature vector to form L sets of audio-visual vector $\mathbf{u_T}l(t) = [\mathbf{v_T}(t); a_{\mathbf{T}l}(t)]$. The objective of the training process is to obtain the parameter set $\lambda_{li} = \{\gamma_{li}, \boldsymbol{\mu}_{li}, \boldsymbol{\Sigma}_{li}\}$ associated with each $\mathbf{u_T}l(t)$.

After independent GMM training, $L \times I$ parameter sets $\{\lambda_{li}\}$ are estimated by the expectation maximization algorithm for each speaker.

IV. Indeterminacies Cancellation Algorithm

The indeterminacies cancellation algorithm is based on coherence maximization. Suppose the separation succeeds without any permutation or scaling ambiguity, then $y_k(n) = s_k(n)$, $y_k(n)$ will have maximum coherence with its corresponding video signal $v_k(t)$. Treating the frequency bin group $f \in \mathcal{F}_l$ as a whole, we can maximize the following criterion in the frequency domain to address the indeterminacies:

$$[\hat{\mathbf{P}}(\mathcal{F}_l), \hat{\mathbf{D}}(\mathcal{F}_l)] = \underset{\mathbf{P}(\mathcal{F}_l), \mathbf{D}(\mathcal{F}_l)}{\arg\max} \sum_t \sum_{k=1}^{K} \log p_{AV}(\mathbf{u}_{kl}(t)), \quad (8)$$

where $\mathbf{u}_{kl}(t) = [\mathbf{v}_k(t); a_{kl}(t)]$ is the audio-visual feature, $\mathbf{v}_k(t)$ is the visual feature associated with the k-th speaker at time frame t, and $a_{kl}(t)$ is the audio feature extracted from the k-th source estimate corresponding to the l-th filterbank.

To estimate $s_1(n)$ from the observations, we can get the separation vector $\mathbf{p}(\mathcal{F}_l)$ and the scale parameter $\alpha(\mathcal{F}_l)$ by maximizing:

$$[\hat{\mathbf{p}}(\mathcal{F}_l), \hat{\alpha}(\mathcal{F}_l)] = \underset{\mathbf{p}(\mathcal{F}_l), \alpha(\mathcal{F}_l)}{\arg\max} \sum_t \log p_{AV}(\mathbf{u}_{1l}(t)). \quad (9)$$

A. Permutation Indeterminacy Cancellation

In equation (8), the direct summation of log-likelihood is very sensitive to outliers. It happens that one outlier may change the total summation greatly and result in a wrong decision. To deal with this problem, we propose a new sorting scheme. For convenience, we present an example of the 2×2 case:

1. Extract the visual features $\mathbf{v}_1(t)$ and $\mathbf{v}_2(t)$ from the video signal associated with the two speakers.
2. Extract the audio features $\mathbf{a}_1(t)$ and $\mathbf{a}_2(t)$ from $Y_1(f,t)$ and $Y_2(f,t)$ respectively.
3. Get the audio-visual data $\mathbf{u}_{kl}(t) = [\mathbf{v}_k(t); a_{kl}(t)]$ and $\mathbf{u}_{k^\dagger l}(t) = [\mathbf{v}_k(t); a_{k^\dagger l}(t)]$, (where $k = 1, 2$, $l = 1, ..., L$, and $\dagger$ denotes the permutation version, $1^\dagger = 2$, $2^\dagger = 1$).
4. Calculate the audio-visual probability $p_{AV}(\mathbf{u}_{kl}(t))$ and $p_{AV}(\mathbf{u}_{k^\dagger l}(t))$ based on the GMM model in equation (6) and the parameter set $\{\lambda_{il}\}_k$ associated with each speaker that has been estimated in the training stage.
5. Define a criterion:

$$\mathcal{J}(l,t) \overset{\text{def}}{=} \begin{cases} 1, & \sum_k \log p_{AV}(\mathbf{u}_{kl}(t)) > \sum_k \log p_{AV}(\mathbf{u}_{k^\dagger l}(t)) \\ 0, & \sum_k \log p_{AV}(\mathbf{u}_{kl}(t)) < \sum_k \log p_{AV}(\mathbf{u}_{k^\dagger l}(t)) \end{cases}.$$

6. If $\sum_{t=1}^{T} \mathcal{J}(l,t) > T/2$, do nothing; otherwise, swap the rows of $\mathbf{W}(f)$ (i.e. $\mathbf{W}(f) \longleftarrow \left[\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right] \mathbf{W}(f)$), and $\mathbf{Y}(f,t)$ for $f \in \mathcal{F}_l$.

In step 5, we have used a new criterion instead of the cumulative log-likelihood, to avoid the influence of outliers as in [11], which is equivalent to majority voting over time frames. For the sake of accuracy, we can iterate steps 2 to 6.

B. Scaling Indeterminacy Cancellation

Suppose we are now interested in addressing the scaling ambiguity of source estimate $y_1(n)$. If $Y_1^\ddagger(f,t) = \alpha(\mathcal{F}_l) Y_1(f,t)$ is the exact copy of the source speech $S_1(f,t)$ for $f \in \mathcal{F}_l$ without any scaling amplification, i.e., $Y_1^\ddagger(f,t) = S_1(f,t)$, for $f \in \mathcal{F}_l$, then this combines with equation (7) to give

$$\sum_{t=1}^{T} a_l^\ddagger(t) = 2T \log |\alpha(\mathcal{F}_l)| + \sum_{t=1}^{T} a_l(t). \quad (10)$$

Therefore we can calculate $\mathcal{F}_l$ spanned by each filter:

$$\alpha(\mathcal{F}_l) = \exp\left\{ \left(\sum_{t=1}^{T} a_l^\ddagger(t) - \sum_{t=1}^{T} a_l(t) \right) / (2T) \right\}. \quad (11)$$

$\sum_{t=1}^{T} a_l(t)$ is straightforward to calculate, and the priority is on the estimation of $\sum_{t=1}^{T} a_l^\ddagger(t)$ from the given visual vector $\mathbf{v}(t)$. First we need to get the marginal probability density of the visual feature corresponding to each filterbank l:

$$p_V(\mathbf{v}(t) \mid l) = \sum_{i=1}^{I} \gamma_{liV} \mathcal{N}(\mathbf{v}(t) \mid \boldsymbol{\mu}_{liV}, \boldsymbol{\Sigma}_{liV}), \quad (12)$$

where γ_{liV} is the weighting parameter, $\boldsymbol{\mu}_{liV}$ is the mean vector, $\boldsymbol{\Sigma}_{liV}$ is the covariance matrix of the visual data, then

$a_l^{\ddagger}(t)$ can be estimated as:

$$a_l^{\ddagger}(t) = \sum_{i}^{I} \beta_{li}(t)\mu_{liA}, \tag{13}$$

where μ_{liA} is the mean parameter of the audio feature $a_l(t)$ for the i-th kernel, and

$$\beta_{li}(t) = \frac{\gamma_{li} p_V(\mathbf{v}(t) \mid i)}{\sum_j \gamma_{lj} p_V(\mathbf{v}(t) \mid j)}.$$

Then from equation (11) we can estimate $\alpha(\mathcal{F}_l)$. In such a way, we get L scale parameters, and each one affects the frequency bins spanned by one filter.

However, adjacent $\mathcal{F}_l$ overlap with each other, and we cannot define two scale parameters for an overlapped frequency bin. To solve this problem, we smooth between the L scale parameters with linear interpolation, as shown in Figure 3.

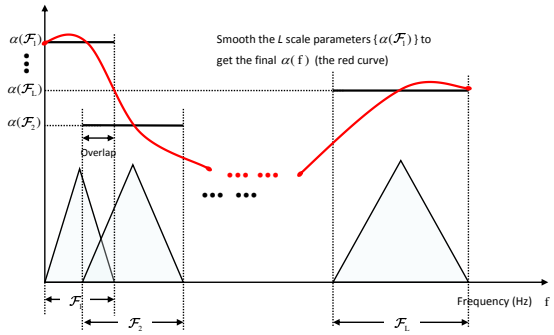

Fig. 3. Smooth between $\alpha(\mathcal{F}_l)$ to obtain $\alpha(f)$.

V. Experimental Results

We tested the proposed algorithm on the XM2VTS [15] multi-modal database. The frontal face videos were captured at 25 fps and the speech signals were continuous sentences of words and digits recorded at 32 kHz. For each speaker, there are 24 recordings repeating 3 sentences. We trained the audio-visual coherence model of one target speaker with audio-visual speech lasting for 41 seconds. The audio was downsampled to 16 kHz, and the 32 ms (512 samples) Hamming window with 12 ms overlap was used in the STFT. Audio features were extracted from 24 Mel-scaled filter banks. We chose the first 3 principal components from the video as the visual features, and they were upsampled to 50 Hz to be synchronized with the audio features.

To test the performance of the permutation cancellation algorithm, the speech signal from the target speaker and another interference speech signal randomly chosen from 96 audio signals by 4 other speakers were transformed into the time-frequency domain by the STFT. We swapped the spectral components of consecutive frequency bins of a filter. Then we calculated the accuracy rate of the permutation cancellation with different frame numbers T. In Figure 4, each

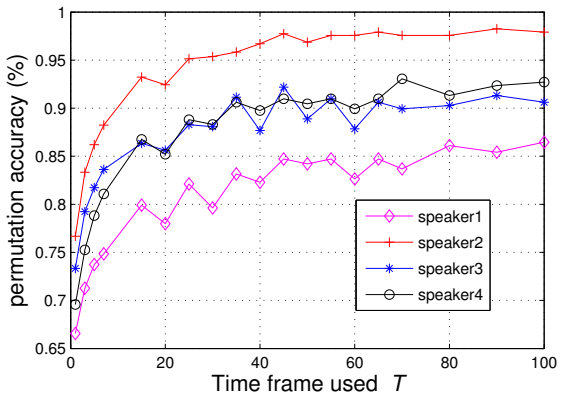

Fig. 4. The permutation accuracy with different time frames.

curve represents the comparison of the target speaker with an interference speaker. The result is an average of 24 mixtures.

To test the scaling cancellation, we amplified the spectral components from the target speaker with different scaling parameters $d(f)$ at different frequency bins. If the scaling problem was solved successfully, we should get $\alpha(f)$ that satisfies $\alpha(f) \cdot d(f) = 1$. Figure 5 shows the real and estimated scaling factors $\alpha(f)$ with the algorithm described in section 3. The solid curve (estimated $\alpha(f) \cdot d(f)$) in the lower part is close to 1, so we have mitigated the scaling problem greatly.

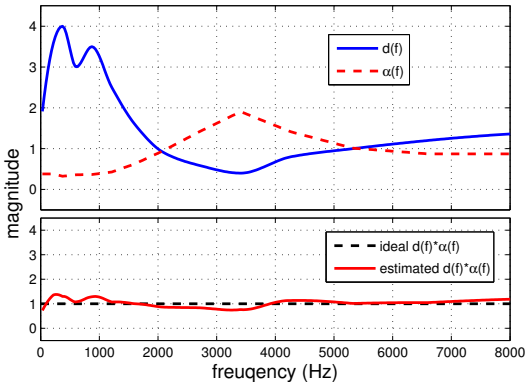

Fig. 5. The scaling cancellation results.

We then tested the algorithm with convolutive mixtures synthesized on a computer. The filters $\{h_{pk}\}$ were generated by the system utilizing the impulse response measurements of a conference room [16] with various positions of the microphones and the speakers. Two audio signals with each lasting 4 s were convolved with the filters to generate the mixtures.

We used the signal to interference ratio (SIR) at different signal to noise ratios (SNRs) as a criterion to evaluate the performance of our bimodal BSS algorithm. Based on the

extraction of $s_1(n)$ from the observations, we have

$$\text{SIR} = 10\log\frac{\sum_n \parallel \sum_{p=1}^{P} w_{1p}(n) * h_{p1}(n) * s_1(n) \parallel}{\sum_n \parallel \hat{s}_1(n) - \sum_{p=1}^{P} w_{1p}(n) * h_{p1}(n) * s_1(n) \parallel}. \quad (14)$$

The degradation of convolutive BSS performance is maily caused by the permutation problem. From the upper half of table I, we found that after applying the permuation cancellation algorithm with the sorting scheme in section IV-*A*, SIRs of the source signal were improved greatly in a wide range of noise levels, with the highest point at about 20dB. In this table, the input SNR is the ratio between the audio signals and gaussian noise, i.e. energy$(s_1 + s2)$/energy$(noise)$, and the input SIR is the ratio between a target signal and the interference, i.e. energy(s_1)/energy$(s_2 + noise)$. The output SIR was caculated by equation (14). In the lower half of table I, after permuation indeterminacy cancellation, the scaling ambiguity cancellation algorithm was applied to the realigned spectral components, which improved the performance in high noise environment.

TABLE I
OUTPUT SIR (dB) COMPARISON.

Evaluation of Permutation Indeterminacy Cancellation						
Input SNR		10	15	20	25	30
Input SIR		-1.42	-0.91	-0.74	-0.68	-0.66
Output SIR	before sorting	4.02	6.40	8.81	13.41	13.24
	after sorting	5.29	9.28	12.97	14.66	14.8
Evaluation of Scaling Indeterminacy Cancellation						
Input SNR		4	6	8	10	12
Input SIR		-3.13	-2.38	-1.82	-1.42	-1.15
Output SIR	before scaling	0.66	2.07	3.83	5.29	6.77
	after scaling	2.07	3.71	4.55	5.82	6.80

VI. CONCLUSION

We have presented a new audio-visual convolutive BSS system. In this system, we have combined the audio features with visual features to form an audio-visual feature space. A new sorting scheme exploiting the audio-visual coherence to solve the permutation indeterminacy problem has also been presented. We also provided a new method to estimate the power spectrum to mitigate the scaling ambiguity. Our algorithm has been tested on the XM2VTS database and has shown good performance.

ACKNOWLEDGMENT

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).

REFERENCES

[1] Schwartz, J.L., Berthommier, F. and Savariaux, C., "Seeing to Hear Better: Evidence For Early Audio-visual Interactions in Speech Identification.", *Cognition*, vol. 93, pp. B69–78, Sep. 2004.

[2] 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.

[3] Cardoso, J.F., Souloumiac, A.: "Blind Beamforming for Non-Gaussian Signals." *IEEE Proc.-F,* vol. 140, no. 6, pp. 362–370, 1993.

[4] Comon, P., "Independent Component Analysis, a New Concept?", *Signal Process.*, vol. 36, pp. 287–314, 1994.

[5] Anemüller, J., Kollmeier, B.: "Amplitude Modulation Decorrelation for Convolutive Blind Source Separation." In: *Proc. ICA*, pp. 215–220, 2000.

[6] 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.

[7] Sawada, H., Mukai, R., Araki, S., and Makino, S., "A Robust and Precise Method for Solving the Permutation Problem of Frequency-Domain Blind Source Separation", *IEEE Trans. Speech Audio Process.*, vol. 12, pp. 530–538, 2004.

[8] 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.

[9] Wang, W., Cosker, D., Hicks, Y., Sanei, S. and Chambers, J., "Video Assisted Speech Source Separation", in *Proc. IEEE ICASSP*, pp. 425–428, 2005.

[10] Rivet, B., Girin, L. and Jutten, C., "Mixing Audiovisual Speech Processing and Blind Source Separation for the Extraction of Speech Signals from Convolutive Mixtures", *IEEE Trans. Audio Speech Lang. Process.*, vol. 15, pp. 96–108, 2007.

[11] Liu, Q., Wang, W. and Jackson, P., "Use of Bimodal Coherence to Resolve Spectral Indeterminacy in Convolutive BSS", in *Proc. LVA/ICA* 2010.

[12] Liu, Q., Wang, W. and Jackson, P., "Bimodal Coherence based Scale Ambiguity Cancellation for Target Speech Extraction and Enhancement", in *Proc. Interspeech* 2010.

[13] Casanovas, A.L., Monaci, Gianluca., Vandergheynst, P. and Gribonval, R., "Blind Audiovisual Source Separation Using Overcomplete Dictionaries", in *Proc. IEEE ICASSP*, 2008.

[14] Kass, M., Witkin, A. and Terzopoulos, D., "Snakes: Active Contour Models", *International Journal of Computer Vision*, pp. 321–331, 1988.

[15] Messer, K., Matas, J., Kittler, J., Luettin, J., Maitre, G.: XM2VTSDB: The Extended M2VTS Database. In: AVBPA, 1999. [Online] Available: http://www.ee.surrey.ac.uk/CVSSP/xm2vtsdb/

[16] Westner, A.: Room Impulse Responses, 1998. [Online] Available: http://alumni.media.mit.edu/~westner/papers/ica99/node2.html