Use of Bimodal Coherence to Resolve the Permutation Problem in Convolutive BSS

Qingju Liu[1,*], Wenwu Wang[2,*], Philip Jackson[3,*]

Centre for Vision, Speech and Signal Processing, Faculty of Engineering and Physical Sciences, University of Surrey, Guildford, GU2 7XH, United Kingdom

Abstract

Recent studies show that facial 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 three contributions. With the synchronized features, we propose an adapted expectation maximization (AEM) algorithm to model the audio-visual coherence in the off-line training process. To improve the accuracy of this coherence model, we use a frame selection scheme to discard nonstationary features. Then with the coherence maximization technique, we develop a new sorting method to solve the permutation problem in the frequency domain. We test our algorithm on a multimodal speech database composed of different combinations of vowels and consonants. The experimental results show that our

*Corresponding authors. E-mails: {Q.Liu, W.Wang, P.Jackson}@surrey.ac.uk.
Fax: +44 (0) 1483 686031
[1]Tel: +44 (0) 1483 683413
[2]Tel: +44 (0) 1483 686039
[3]Tel: +44 (0) 1483 686044

proposed algorithm outperforms traditional audio-only BSS, which confirms the benefit of using visual speech to assist in separation of the audio.
Keywords: convolutive blind source separation (BSS), audio-visual coherence, Gaussian mixture model (GMM), feature selection and fusion, adapted expectation maximization (AEM), indeterminacy

1. Introduction

Human speech perception is essentially bimodal as speech is perceived by the interactions of auditory and visual sensory processing [1, 2]. Looking at the speaker's lips improves the intelligibility of human speech embedded in cocktail party noise due to the contribution of the complementary visual information [2]. There is a complex non-linear relationship between the auditory and visual streams, usually refered to as the audio-visual coherence or correlation [3]. In feature space, the coherence can be coded by audio-visual atoms or dictionaries [4, 5] with matching pursuit [6] techniques, or characterised statistically with models such as Gaussian mixture models (GMM) [7]. Exploiting these cross-modal interactions, the visual stream has proven a success in improving the robustness to noise in many fields of applications, including automatic speech recognition [8], speaker localization [4, 9], speech enhancement or audio filtering [10, 11], and blind source separation [3, 5, 12–16].

In traditional blind source separation (BSS) for auditory mixtures, typically only audio signals are considered. Under the framework of independent component analysis (ICA) [17], the BSS problems have been extensively studied and many classical algorithms have been proposed for the instantaneous

mixing model such as the "J-H" algorithm [18], JADE [19], Infomax [20], SOBI [21] and FastICA [22] algorithms. For the more complex convolutive mixing model, one can apply either the time domain deconvolution algorithms [23–25], or the frequency domain separation algorithms [12–15, 26–31], which often suffer from the permutation and scaling ambiguity problems.

Considering the bimodal nature of human speech, we could potentially improve the separation of the source signals from their audio mixtures utilizing the audio-visual coherence obtained by the integration of visual speech. This is known as audio-visual or bimodal BSS [3, 5, 12, 13, 15, 16], a recent development in multi-modal signal processing. Sodoyer et al. [3] addressed the separation problem for an instantaneous mixture of decorrelated sources, with no further assumptions on independence or non-Gaussianity. Wang et al. [13] implemented a similar idea by applying the Bayesian framework to the fused feature observations for both instantaneous and convolutive mixtures. Rivet et al. [12] proposed a new statistical tool utilizing the log-Rayleigh distribution for modelling the audio-visual coherence, and then used the coherence to address the permutation and scaling ambiguities in the spectral domain. Casanovas et al. [5] detected temporal audio-visual structures represented by atoms taken from redundant dictionaries, and extracted sources from a soundtrack. Naqvi et al. [16] utilized beamforming in the frequency domain for moving sources in the teleconference-like scenario, incorporating the geometrical model derived on the basis of the beamforming theory.

Despite being promising, these approaches are also limited in some situations. For example, the algorithm proposed in [3] was designed only for instantaneous mixtures. The method in [13] considered a convolutive model

with a relatively small number of taps for the mixing filters. The approach in [12] modelled the audio-visual coherence in a high dimensional feature space, which often results in an over-fitting problem and therefore is sensitive to outliers. Cross-modal correlation was not exploited in the separation stage in [5], where visual information was used only for voice activity detection. In [16], the video provided the position information about the distance and azimuth angles between the moving speakers and the microphone array, however, source separation was still performed in the audio domain.

In this paper, we attempt to address some of these limitations. Motivated by the work in [12, 13], we follow a similar two-stage framework which includes off-line training and online separation. In particular, we consider a convolutive mixing model and address the permutation problem associated with the frequency domain BSS (FD-BSS). In the off-line training stage, we build a model to statistically characterise the audio-visual coherence in the feature space. This coherence is built on the audio-visual features extracted from the target speech. Mel-frequency cepstral coefficients (MFCCs) are used as the audio features, and the lip width and height as visual features, which are synchronised with the audio features on a frame-by-frame basis before statistical training. In the separation stage, coherence maximization is applied for the alignment of the ICA-separated spectral components. Different from [12, 13], however, we have proposed three new techniques to improve the training and separation processes. First, a frame selection scheme is proposed to remove the non-stationary features which consequently improves the robustness and accuracy of the estimation of the audio-visual coherence. Second, the classical expectation maximisation (EM) algorithm is modified

to take into account the different influences of the audio features, resulting in an adapted EM (AEM) algorithm, which further improves the estimation of the joint audio-visual probability. Third, a novel sorting scheme is proposed to address the permutation problem. A preliminary version of this work was presented in [15]. Different from [15], in this paper, we have developed a robust feature selection scheme for audio-visual modelling as mentioned above. In addition, we have further improved the audio feature representation as described in Section 3.1. Moreover, here we have performed systematic evaluations on real recordings, and compared the performance of the proposed method with the state-of-the-art methods.

The remainder of the paper is organised as follows. An overview of traditional frequency domain convolutive BSS and the framework of the proposed audio-visual BSS system are presented in Section 2. Then Section 3 introduces the feature extraction and fusion method for the modelling of the cross-model correlation, including a new frame selection approach and an adapted expectation maximisation algorithm to improve the accuracy of this model. The proposed de-permutation algorithm exploiting the audio-visual coherence is presented in Section 4. The simulation results are analysed and discussed in Section 5, followed by the conclusions.

2. BSS for Convolutive Mixtures

2.1. Convolutive Model

BSS aims to recover sources from their mixtures without any or with little prior knowledge about the sources or the mixing process. Consider a cocktail party scenario, the observation at each sensor is the sum of K filtered source

signals, which can be approximated by 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 $\boldsymbol{\xi}(n) = [\xi_1(n), ..., \xi_P(n)]^T$ the additive noise vector, where T is vector transpose. $\mathbf{H}$ is 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 ($\mathbf{y}(n)$ or $\hat{\mathbf{s}}(n)$ represents the estimated version of $\mathbf{s}(n)$). We consider a time-invariant system where both the mixing filter $\mathbf{H}$ and separation filter $\mathbf{W}$ are assumed to be time-invariant. In practice, a finite impulse response (FIR) filter is used to implement h_{pk} and w_{kp}.

2.2. Frequency domain BSS

Convolutive BSS can be directly performed in the time domain [23–25] by deconvolution, but the computational complexity is high especially when the mixing filters have long taps. Based on the short-time stationarity of the speech signals and the linear time-invariance of the mixing process, an

alternative is to perform convolutive BSS in the frequency domain by applying the short-time Fourier transform (STFT) to the observations. In each frequency bin f, we get an instantaneous mixing model ignoring the noise:

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

where $\boldsymbol{X}(f,t) = [X_1(f,t), ..., X_P(f,t)]^T$ is the observation vector in frequency bin f and time frame t, and $\mathbf{H}(f)$ is the Fourier transform of $\mathbf{H}$.

Then in each frequency bin f, separate ICA [17] algorithms for instantaneous models are applied to obtain the independent outputs $\boldsymbol{Y}(f,t) = [Y_1(f,t), ..., Y_K(f,t)]^T$, assumed to be the source estimates:

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

The technique of the frequency domain BSS is depicted in the upper dashed box of Figure 1. In this paper, a determined system is considered, i.e., $K = P = 2$.

2.3. Scaling and Permutation Indeterminacy

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

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

These are the so-called permutation ($\mathbf{P}(f)$) and scaling ($\mathbf{D}(f)$) ambiguities, which present severe problems when reconstructing the separated sources in the time domain. The scaling ambiguity can be greatly mitigated by the normalization of the separation matrices based on the minimal distortion principle (MDP) [29] **which is also used here**. In this paper, we only

consider the permutation problem, where the order of the recovered source components at each frequency bin may not be consistent with each other (e.g. $\begin{bmatrix} Y_1(f_1,t) \\ Y_2(f_1,t) \end{bmatrix} = \begin{bmatrix} S_1(f_1,t) \\ S_2(f_1,t) \end{bmatrix}, \begin{bmatrix} Y_1(f_2,t) \\ Y_2(f_2,t) \end{bmatrix} = \begin{bmatrix} S_2(f_2,t) \\ S_1(f_2,t) \end{bmatrix}, \begin{bmatrix} Y_1(f_3,t) \\ Y_2(f_3,t) \end{bmatrix} = \begin{bmatrix} S_1(f_3,t) \\ S_2(f_3,t) \end{bmatrix}, \ldots$).

To address the permutation problem, many algorithms [26–28, 30–33] have been proposed. For example, the approach in [26, 30] utilizes the continuity of the spectral components in adjacent frequency channels while [27, 28] use direction of arrival estimation, [31] combines both previous techniques, and [32] utilizes statistical signal models. However, the performance of these algorithms can be degraded by acoustical noise. Information from the video has been shown to be useful for improving the performance of automatic speech recognition systems [8]. The potential of using visual information for audio source separation problems, such as the permutation problem, has not been fully investigated, which motivates this study, as now discussed.

2.4. Overview of the Proposed System

To address the permutation problem in FD-BSS, we use an audio-visual BSS system, shown in Figure 1. As mentioned in Section 1, the system contains two stages: training stage and separation stage. The training stage is shown in the lower dashed box of Figure 1, which includes feature extraction and feature fusion. First, we extract the audio features $\mathbf{a}(t)$ from the training audio data $s(t)$, and some geometric-type features $\mathbf{v}(t)$ from the training video stream $v(t)$. Second, we use the GMM to statistically characterise the audio-visual coherence $p(\mathbf{a}(t), \mathbf{v}(t))$, and then an AEM algorithm is applied to estimate the parameters of the GMM model. The separation stage is shown in the upper dashed box, which is performed in the audio

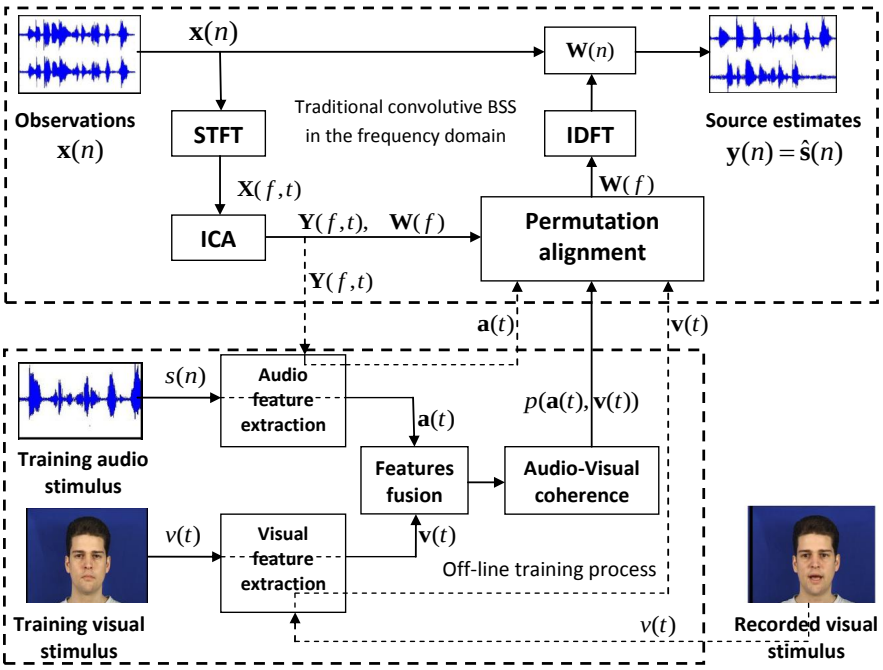


Figure 1: Flow of the proposed audio-visual BSS system. The upper dashed box illustrates the traditional audio-only BSS in the frequency domain. The lower dashed box shows the training process, which aims to estimate the parameters for the audio-visual coherence after audio-visual fusion. Before the reconstruction of the recovered signals in the time domain, we resolve the permutation indeterminacy by audio-visual coherence maximization using the feature vectors $\mathbf{a}(t)$ and $\mathbf{v}(t)$ in dashed lines (with arrows).

domain. To address the permutation problem in the separation stage, we use the information (i.e. the audio-visual joint probability) obtained from the training stage. Specifically, we align the permutations across the frequency bands based on a sorting scheme which iteratively re-estimates the audio-visual coherence probability from the separated source components at each frequency bin based on the trained audio-visual model.

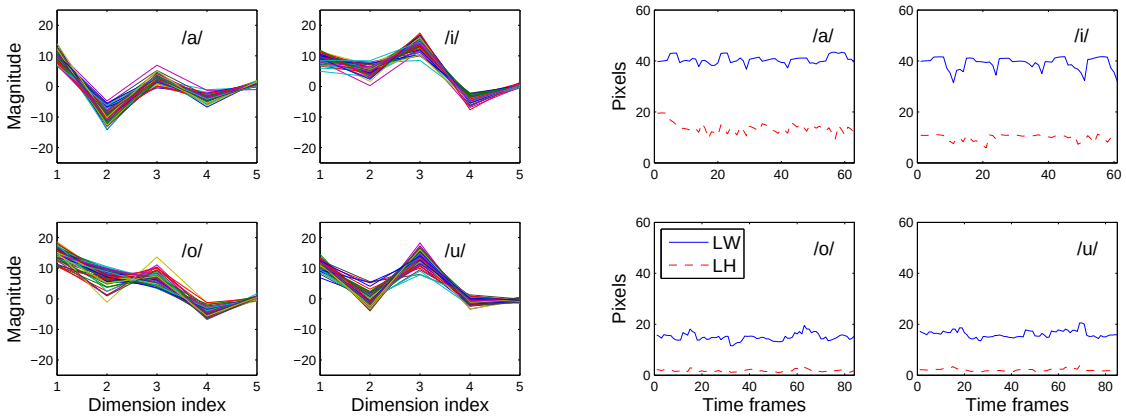
3. Feature Extraction and Fusion

Our algorithm is based on the fact that there is a relationship between the video signal and the corresponding contemporary audio signal, which is the so-called audio-visual coherence. We model the coherence in the feature level, for which the features are extracted from the audio and video data respectively. Since the two types of signals are recorded with different sampling rates and dimensions, we need to extract the features from them separately.

3.1. Extraction of Audio and Visual Features

We take the Mel-frequency cepstral coefficients (MFCCs) as audio features as in [13] 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 time power spectrum on a Mel-frequency scale. We then apply a lifter for the reweighing of cepstral coefficients and obtain a $(L+1)$-dimensional MFCC vector $[c_0(t), c_1(t)..., c_L(t)]^T$, where t is the time frame index, as in equation 3. Compared to our early work in [15], where the first component is the logarithmic power, we remove the first coefficient $c_0(t)$ to avoid the influence of the magnitude, and obtain the L-dimensional audio feature $\mathbf{a}(t) = [c_1(t), ..., c_L(t)]^T$. For simplicity, we denote the training audio feature vector as $\mathbf{a}(t) = [a_1(t), ..., a_L(t)]^T$.

Unlike the appearance-based visual features used in [8, 13], which are sensitive to the lighting conditions, we use the same front geometric visual features as in [3, 12]: the lip width (LW) and height (LH) from the internal labial contour. The geometric features $\mathbf{v}(t) = [\mathrm{LW}(t), \mathrm{LH}(t)]^T$ are



(a) Audio features (b) Visual features

Figure 2: The audio-visual features of four vowels /a/,/i/,/o/,/u/. The vowels are obtained by selecting and concatenating the first several frames in each audio sequence in the data corpus. (a) Each curve corresponds to one time frame, associated to the time frame in the right half plot. We choose $L = 5$, therefore the audio feature is 5-dimensional. (b) The visual features last about 60–80 frames at 50 Hz.

low-dimensional and robust to luminance, which mitigates the over-fitting problem encountered in complex systems with a large number of parameters and greatly reduces the computational complexity. Figure 2 shows the typical features extracted from four vowels: /a/, /i/, /o/ and /u/, which were obtained from the multimodal database, as depicted in Section 5, which is also used in Figures 3 and 4. After the feature extraction process, we concatenate synchronized audio and visual features to build the $(L+2)$-dimensional audio-visual space $\mathbf{u}(t) = [\mathbf{v}(t); \mathbf{a}(t)]$, which will be used for training.

3.2. Robust Feature Frame Selection

If someone utters an isolated speech sound such as /a/, the visual features will likely be stationary with minimal fluctuation. However in the transition periods from one phone to another in fluent speech, the visual parameters fluctuate drastically with a large variance, which can produce ambiguous visual features typical of another speech sound or phone. For instance, in the transition process from /a/ to /b/, several frames of the mouth shape may look like the utterance of /o/. Also, these transitions are not stationary in the audio signal. Therefore, to improve the estimate of audio-visual coherence, we propose a feature frame selection scheme based on the dynamic characteristics [34] of the visual features.

At each time frame centred by the visual feature $\mathbf{v}(t) = [\mathrm{LW}(t), \mathrm{LH}(t)]^T$, we extract a short time period with $2Q+1$ frames, then calculate

$$\gamma_{\mathrm{LW}}(t) = \sigma(\mathrm{LW}(t)) + \alpha_{\mathrm{LW}} \|\mathrm{LW}(t+Q) - \mathrm{LW}(t-Q)\|, \tag{6}$$

where $\sigma(\cdot)$ is the standard deviation and α_{LW} is a weighting coefficient, chosen between 0 and 1. Then we define a Boolean variable to determine the stationarity of this frame

$$\mathcal{F}_{\mathrm{LW}}(t) \stackrel{\mathrm{def}}{=} \begin{cases} 1, \ \gamma_{\mathrm{LW}}(t) < \delta_{\mathrm{LW}} \overline{\mathrm{LW}(t)} \\ 0, \ \text{otherwise} \end{cases} \tag{7}$$

where δ_{LW} is a comparison coefficient, typically chosen as 0.5, and $\overline{\mathrm{LW}(t)}$ is the mean over the $2Q+1$ frames, defined as $\overline{\mathrm{LW}(t)} = \frac{1}{2Q+1} \sum_{q=-Q}^{Q} \mathrm{LW}(t+q)$. Then, we smooth the binary variable between adjacent frames

$$\mathcal{F}^{\mathrm{S}}_{\mathrm{LW}}(t) = \mathcal{F}_{\mathrm{LW}}(t-1) \vee \mathcal{F}_{\mathrm{LW}}(t) \vee \mathcal{F}_{\mathrm{LW}}(t+1), \tag{8}$$

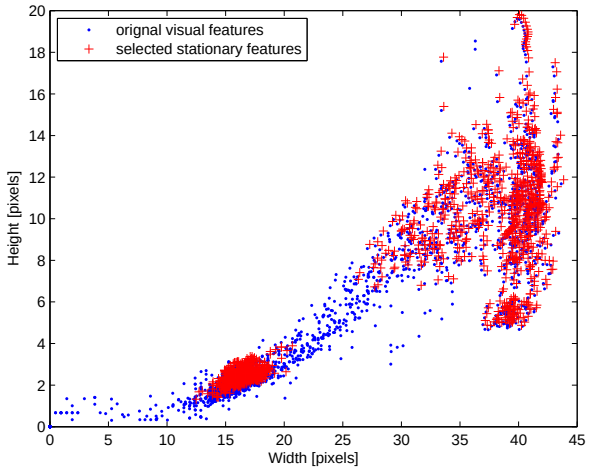


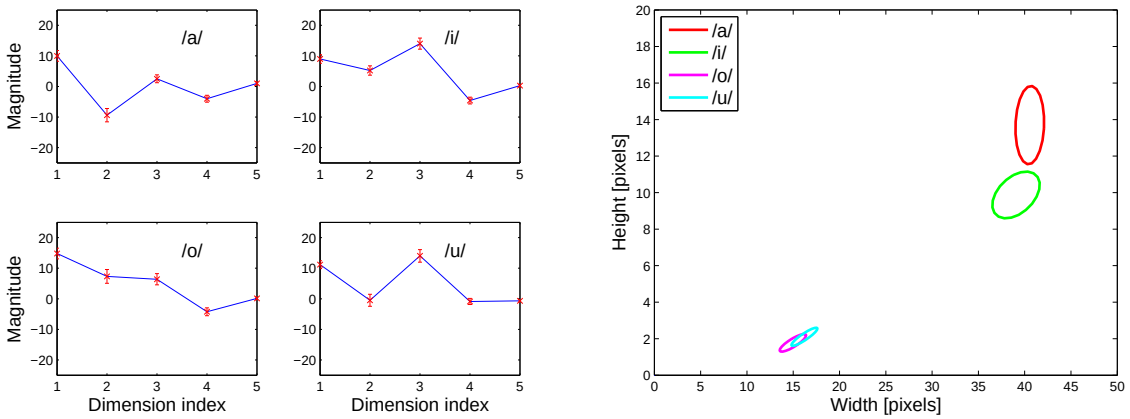
Figure 3: Visual frame selection scheme. Each dot (or cross) represents a two dimensional visual feature vector at one time frame. After frame selection, the transitions from one syllable to another have been removed. In other words, the features are better clustered (denoted by crosses), as compared to the original scatter plot of all the features (denoted by dots).

where $\vee$ denotes disjunction, i.e., a logical OR operator. In the same way, we can determine $\mathcal{F}_{\mathrm{LH}}^{\mathrm{S}}(t)$, and the final decision is

$$\mathcal{F}(t) = \mathcal{F}_{\mathrm{LW}}^{\mathrm{S}}(t) \wedge \mathcal{F}_{\mathrm{LH}}^{\mathrm{S}}(t), \tag{9}$$

where $\wedge$ denotes conjunction, i.e., a logical AND operator.

If $\mathcal{F}(t) = 1$, the frame will be chosen, otherwise it will be discarded. The audio-visual features associated with the selected frames are used in both the training and separation stages. Figure 3 shows an example of the visual features before (dot) and after (cross) selection. The feature selection has essentially removed the redundant unstable features and hence improves the spatial distribution of the clusters of the lip features.



(a) Audio kernels (b) Visual kernels

Figure 4: The audio-visual kernels for four vowels: /a/, /i/, /o/ and /u/. (a) Distributions of audio features described by mean (solid line) and the variance of one standard deviation (bar). (b) The visual kernels show the spacial distributions of the four vowels.

3.3. Feature-Level Fusion

The audio-visual coherence can be statistically characterized by a GMM with I kernels:

$$p_{AV}(\mathbf{u}(t)) = \sum_{i=1}^{I} \gamma_i \mathcal{N}(\mathbf{u}(t) \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \tag{10}$$

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

$$\mathcal{N}(\mathbf{u}(t) \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \frac{\exp\{-\frac{1}{2}(\mathbf{u}(t) - \boldsymbol{\mu}_i)^T {\boldsymbol{\Sigma}_i}^{-1}(\mathbf{u}(t) - \boldsymbol{\mu}_i)\}}{\sqrt{(2\pi)^{L+2} \mid \boldsymbol{\Sigma}_i \mid}}. \tag{11}$$

We denote $\lambda_i = \{\gamma_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\}$ as the parameter set, which 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. 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 ($a_1(t)$) should play a more dominant role in determining the probability $p_{AV}(\mathbf{u}(t))$ 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 $\mathbf{u}(t)$ for $i = 1, ...I$.

$$\beta_i(\mathbf{u}(t)) = 1 - \frac{\parallel \mathbf{u}(t) - \boldsymbol{\mu}_i \parallel}{\sum_{j=1}^{I} \parallel \mathbf{u}(t) - \boldsymbol{\mu}_j \parallel}, \tag{12}$$

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

$$p_i(\mathbf{u}(t)) = \frac{\gamma_i p_G(\mathbf{u}(t) \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)\beta_i(\mathbf{u}(t))}{\sum_{j=1}^{I} \gamma_j p_G(\mathbf{u}(t) \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)\beta_j(\mathbf{u}(t))}. \tag{13}$$

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

$$\begin{aligned} \boldsymbol{\mu}_i = \frac{\sum_t \mathbf{u}(t) p_i(\mathbf{u}(t))}{\sum_t p_i(\mathbf{u}(t))}, \gamma_i = \frac{\sum_t p_i(\mathbf{u}(t))}{\sum_t 1}, \\ \boldsymbol{\Sigma}_i = \frac{\sum_t (\mathbf{u}(t) - \boldsymbol{\mu}_i)(\mathbf{u}(t) - \boldsymbol{\mu}_i)^T p_i(\mathbf{u}(t)) + c\{\boldsymbol{\Sigma}_p\}_i}{\sum_t p_i(\mathbf{u}(t)) + c}, \end{aligned} \tag{14}$$

---

where $\| \cdot \|$ denotes the squared Euclidean distance, and c is a constant chosen to be proportional to the number of samples and $\{\boldsymbol{\Sigma}_p\}_i$ is the penalty

term. Different from the traditional EM algorithm, we have added an influence parameter $\beta_i(\mathbf{u}(t))$ in the expectation step of the AEM algorithm, where $\beta_i(\mathbf{u}(t))$ takes into account the various distance between the sample (i.e. the vector $\mathbf{u}(t)$) to different kernel centres, similar to the idea used in the classical K-means algorithm. Therefore, the different distance between the vector $\mathbf{u}(t)$ and a candidate cluster $\boldsymbol{\mu}_i$ will have a different impact on the probability $p_{AV}(\mathbf{u}(t))$. In addition, unlike our preliminary work in [15], we have added the penalty term $\boldsymbol{\Sigma}_p$ to avoid the covariance matrix becoming singular, which can occur when a kernel converges on one or two sample points (typically outliers). This has a similar effect to a variance floor. Without such a safeguard, the probability would approach infinity in such cases, leading to numerical stability problems in practical implementation. The parameter $\boldsymbol{\Sigma}_p$ can be chosen as a diagonal matrix and the subscript p denotes penalization.

Figure 4 shows the audio-visual kernels for the vowels used in Figure 2. It can be observed that the visual kernels of /o/ and /u/ overlap with each other, but the related audio kernels differ greatly. On the other hand, the audio kernels of /i/ and /u/ look similar, but their visual kernels do not have overlap at all. This implies that the audio and visual features are complementary to each other.

4. Resolution of Permutation Problem

As $y_k(n)$ is the estimate of $s_k(n)$, $y_k(n)$ will have a maximum coherence with the corresponding video signal $v_k(t)$. Therefore we can maximize the following criterion in the frequency domain to address the permutation prob-

lem:

$$\hat{\mathbf{P}}(f) = \arg\max_{\mathbf{P}(f)} \sum_{t} \sum_{k=1}^{K} p_{AV}(\mathbf{u}_k(t)), \tag{15}$$

where $\mathbf{u}_k(t) = [\mathbf{a}_k(t); \mathbf{v}_k(t)]$ is the audio-visual feature extracted from the profile $\hat{S}_k(\cdot, t) = Y_k(\cdot, t)$ and the recorded video associated with the k-th speaker. If we are only interested in an estimate of $s_1(n)$ from the observations, we can get the first separation vector $\mathbf{p}(f)$ by maximizing:

$$\hat{\mathbf{p}}(f) = \arg\max_{\mathbf{p}(f)} \sum_{t} p_{AV}(\mathbf{u}_1(t)). \tag{16}$$

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 two-mixture case (i.e., $K = P = 2$). Assuming the spectral analysis window employs an fast Fourier transform (FFT) size of N samples from the audio mixture signal, based on the symmetry property, we will only need to consider the positive $M = N/2$ bins. We denote $\mathbf{v}_1(t)$ as the visual feature that we have extracted from the recorded video signal associated with the target speaker. We also generate a complementary variable $Y_1^{\dagger}(f, t)$ spanning the same frequency and time-frame space as $Y_1(f, t)$. The proposed sorting scheme for the alignment of the permutation is summarized in the following table, and also shown in Figure 5.

Input: spectral components $\mathbf{Y}(f, t)$, separation matrix $\mathbf{W}(f)$, GMM parameter set $\{\lambda_i\}$, and visual feature $\mathbf{v}_1(t)$.

Output: aligned $\mathbf{W}(f)$ and $\mathbf{Y}(f, t)$.

Initialize the degree of frequency division $I_{\max} = 5$.

For each $i = 1, 2, ..., I_{\max}$, **do**

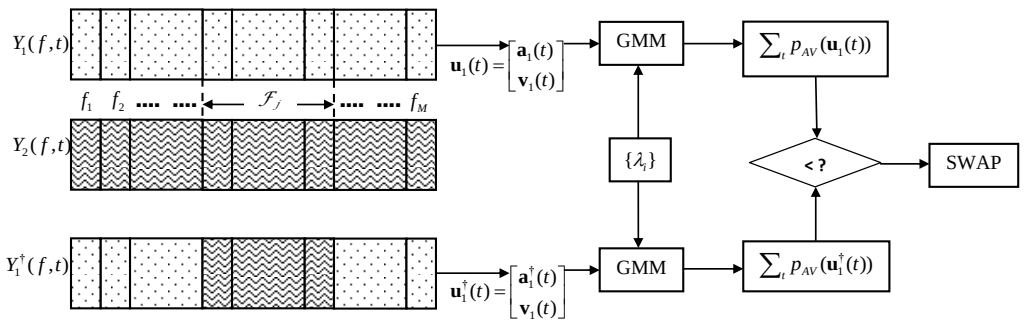


Figure 5: Diagram of the sorting scheme. The j-th loop in the i-th iteration. Spectral components of $Y_1^\dagger(f,t)$ come from $Y_1(f,t)$ (denoted by the dots) except those in the band $\mathcal{F}_j$ which are from $Y_2(f,t)$ (denoted by the curves). We compare the coherence of $Y_1(f,t)$ and $Y_1^\dagger(f,t)$ to determine whether to swap components of $Y_1(f,t)$ and $Y_2(f,t)$ in $\mathcal{F}_j$.

Equally divide M bins into 2^{i-1} parts.

For each $j = 1, ..., 2^{i-1}$, **do**

1) Let the j-th frequency band $\mathcal{F}_j$ span $\{f_{\frac{M}{2^{i-1}}(j-1)+1}, ..., f_{\frac{M}{2^{i-1}}j}\}$. For $f \in \mathcal{F}_j$, let $Y_1^\dagger(f,\cdot) = Y_2(f,\cdot)$; otherwise, $Y_1^\dagger(f,\cdot) = Y_1(f,\cdot)$.
2) Extract the audio feature $\mathbf{a}_1(t)$ and $\mathbf{a}_1^\dagger(t)$ from $Y_1(\cdot,t)$ and $Y_1^\dagger(\cdot,t)$, respectively. Let $\mathbf{u}_1(t) = [\mathbf{a}_1(t); \mathbf{v}_1(t)]$, $\mathbf{u}_1^\dagger(t) = [\mathbf{a}_1^\dagger(t); \mathbf{v}_1(t)]$.
3) 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 (10) and the parameter set $\{\lambda_i\}$ that has been estimated by the AEM algorithm.
4) 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 $\mathbf{W}(f)$ and $\mathbf{Y}(f,\cdot)$ for $f \in \mathcal{F}_j$.

End j

End i

This scheme can reach a high resolution, which is determined by the number of partitions $2^{I_{\max}-1}$ at the final division, and the larger the number $I_{\max}$, the higher the resolution. However, most permutations occur contiguously in practical situations, therefore even if we stop running the algorithm at a very 'coarse' resolution (corresponding to a small $I_{\max}$), the permutation ambiguity can still be substantially reduced (e.g. stop the iteration when the algorithm has divided the positive frequency bins into 16 parts, i.e. $I_{\max} = 5$ and a frequency band of 500Hz).

5. Experimental Results

5.1. Data, Parameter Setup and Performance Metrics

Very similar to the database used in [3, 12], the corpus[4] used in our research contains sequences of "V1-C-V2", where "V1"and "V2" are vowels from /a/, /i/ ,/o/, /u/, and "C" stands for the consonant from /p/, /t/, /k/, /b/, /d/, /g/ or no plosive (in the case of no plosive, the sequences are "V1-V2"). There are 112 combinations recorded twice, one for training and another for testing. The audio sequences are sampled at 16 kHz in mono, 16 bit PCM wave files, while the video sampling rate is 50 Hz and the associated visual features are extracted by a chrome based system with 2496 frames for training and another 2547 frames for testing.

In our experiment, the audio data were obtained by concatenating independent sequences, with each sequence lasting an integer multiple of 20 ms. We concatenated the 112 isolated sequences to obtain approximately

[4]Thanks to Bertrand Rivet in GIPSA-Lab for providing us with this multimodal database.

50 seconds audio for training. In the same way, we chose the beginning 400 frames (approximately 8 s) of the testing data to demonstrate our algorithm. 128 ms sliding Hamming window (2048 taps) with 108 ms overlap (20 ms step size, to be synchronised with the visual features) was applied in STFT. 5-dimensional ($L = 5$) MFCCs as audio features were computed from 24 Mel-scaled filter banks, thus the audio-visual feature was 7-dimensional. We set the FFT size as 2048 (i.e., M=1024). In the frame selection process, we reserved 62% of features by assigning $\delta_{\mathrm{LW}} = 0.35, \delta_{\mathrm{LH}} = 0.5$, $Q = 2$, and $\alpha_{\mathrm{LW}} = \alpha_{\mathrm{LH}} = 1$. For simplicity, we only used 20 ($I = 20$) kernels to approximate the audio-visual coherence.

The algorithm was tested on convolutive mixtures synthesized on computer. We used the real room recordings from the binaural room impulse response database (i.e. AIR database) [35] for the mixing filters. The measurements were recorded with or without dummy head in a low-reverberant studio booth, an office room, a meeting room and a lecture room respectively, of which we chose the meeting room scenario[5]. For a 2×2 mixing system, we chose two positions for two source signals, and used the corresponding room impulse responses as mixing filters. As a result, $C_5^2 = 10$ combinations of mixing filters were used in the following performance evaluations. The two source signals, one was the previously mentioned 8-second truncated segment

[5]In the meeting room scenario, a dummy head was placed on a fixed position and the room impulse responses were captured for five different positions opposite of the head. Each room impulse has 10923 taps (approximately 700 ms) where sampling frequency is 16 kHz, and the reverberation time (T_{60}) is about 300 ms.

from the test audio, and another source signal [6] was continuous speech from the XM2VTS database [36]. Gaussian white noise was added to both mixtures at different signal to noise ratios (SNRs). Figure 6(a) shows the source signals used in our experiments (note that only 2 s are shown here).

In the frequency domain we use the global filters and signal to interference and noise ratio (SINR) as criteria to evaluate the performance of our bimodal BSS algorithm at different signal to noise ratios (SNRs). Suppose $s_1(n)$ is the target source, then in the 2×2 case:

$$\mathbf{G}(f) = \begin{bmatrix} \mathbf{G}_{11}(f) & \mathbf{G}_{12}(f) \\ \mathbf{G}_{21}(f) & \mathbf{G}_{22}(f) \end{bmatrix} = \mathbf{W}(f)\mathbf{H}(f), \tag{17}$$

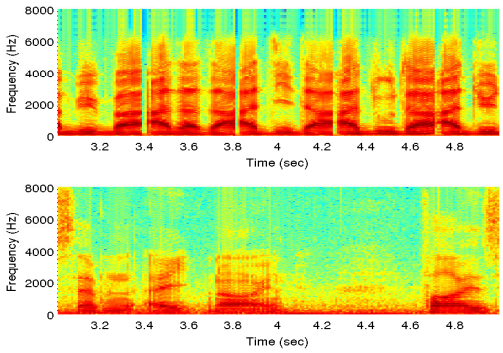
$$\text{SINR} = 10 \log \frac{P_{s_1}}{P_{\hat{s}_1 - s_1}} = 10 \log \frac{\sum_n \parallel \sum_{p=1}^{P} w_{1p} * h_{p1} * s_1(n) \parallel}{\sum_n \parallel \hat{s}_1(n) - \sum_{p=1}^{P} w_{1p} * h_{p1} * s_1(n) \parallel}. \tag{18}$$

5.2. Experimental Results

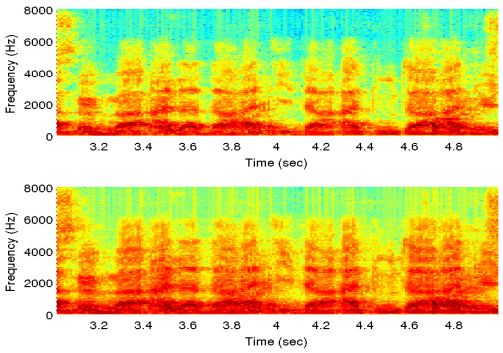
5.2.1. An example

In the first experiment, we show the effectiveness of our algorithm by comparing it with the FD-BSS method without any permutation alignment (denoted as "No Alignment"). For the FD-BSS, **the joint approximate diagolization method proposed in [30] is used for source separation at each frequency bin.** Note that the same ICA algorithm was used here for the contrast methods [30] and [33] in our comparisons in Section 5.2.2. **In addition, the scaling problem is addressed based on the minimal distortion principle [29] for all these methods before performing the permutation alignment.** We used the same parameter set-up as described

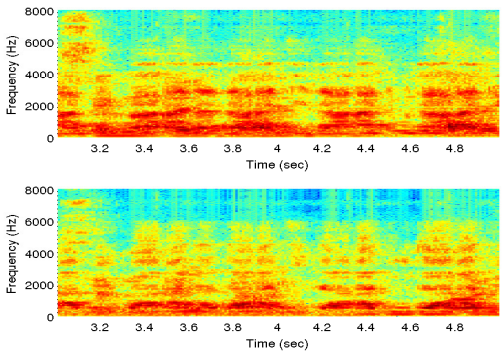
[6] The source signals can also be chosen from a same dataset as done in our work [15].



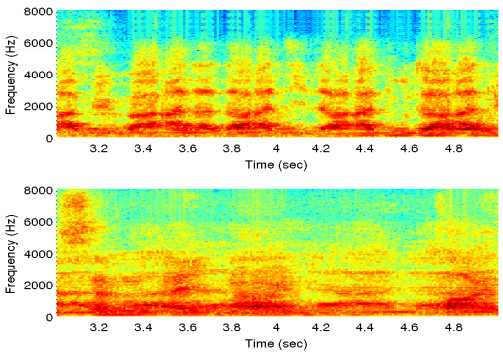
(a) Source signals



(b) Mixtures



(c) Estimates (no alignment)

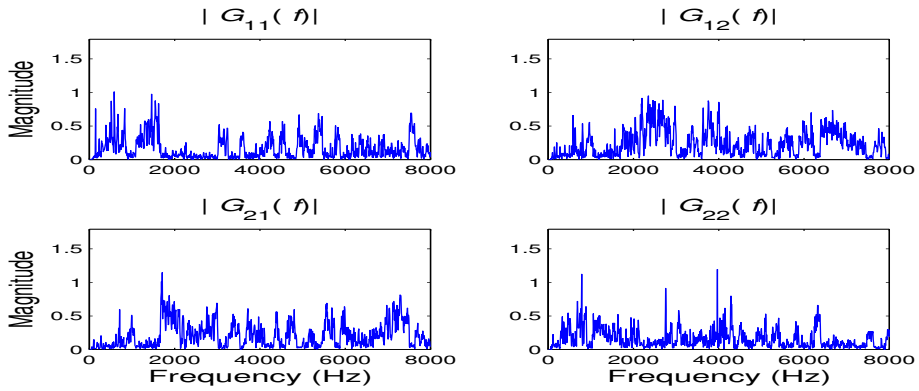


(d) Estimates (audio-visual)

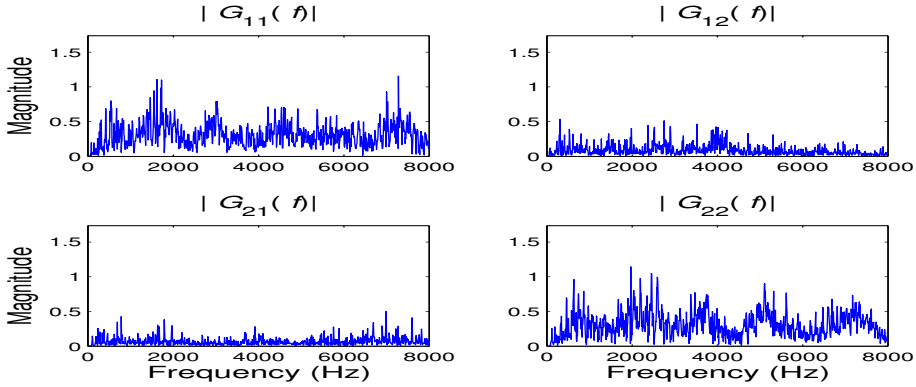
Figure 6: Spectrograms of the sources (a), the mixtures (b) and the estimated sources without permutation alignment (c) where SINR=0.48 dB and the estimated sources by the proposed algorithm (d) where SINR=7.91 dB. The upper sub-plot in (a) is the target signal.

in Section 5.1. The room impulse responses obtained from the AIR database are used as mixing filters, where the sources are placed respectively on the first and fourth position of the meeting room. Each filter has 10923 taps, and the reverberation time (T_{60}) of the meeting room is about 300 ms. We set $N = 2048$ and $I_{max} = 5$. No noise is added to the mixtures. The generated mixtures are shown in Figure 6(b). Figure 6(c) and 6(d) show the separated sources from the mixtures without and with our proposed algorithm respectively, where it can be clearly observed that the audio-visual approach has improved the quality of the separated speech.

Figure 7 shows typical global filters obtained from this experiment. Ideally the global filter should be an identity matrix for each frequency bin. From this figure, it can also be seen that the permutation problem is well addressed by the audio-visual approach, as the magnitudes of G_{12} and G_{21} have been reduced considerably. Alternatively, based on the diagonal property of the global filter matrix, we can also use the criterion $|G_{11}|\cdot|G_{22}|-|G_{12}|\cdot|G_{21}|$ to evaluate the consistency of the permutations across the frequency bins, as shown in Figure 8. **By comparing Figure 8(d) with Figure 8(a), it can be observed that our algorithm successfully corrected the permutation ambiguities at most frequency bins, as the values of $|G_{11}|\cdot|G_{22}|-|G_{12}|\cdot|G_{21}|$ are consistently positive across the frequency bins. In addition, our algorithm performs better than the two baseline methods shown in Figures 8(b) and 8(c). To observe the effect of noise on the permuation errors, we have also plotted the quantity $|G_{11}|\cdot|G_{22}|-|G_{12}|\cdot|G_{21}|$ for the case where 10 dB white Gaussian noise was added to the mixtures, as shown in Figure 9, both the**

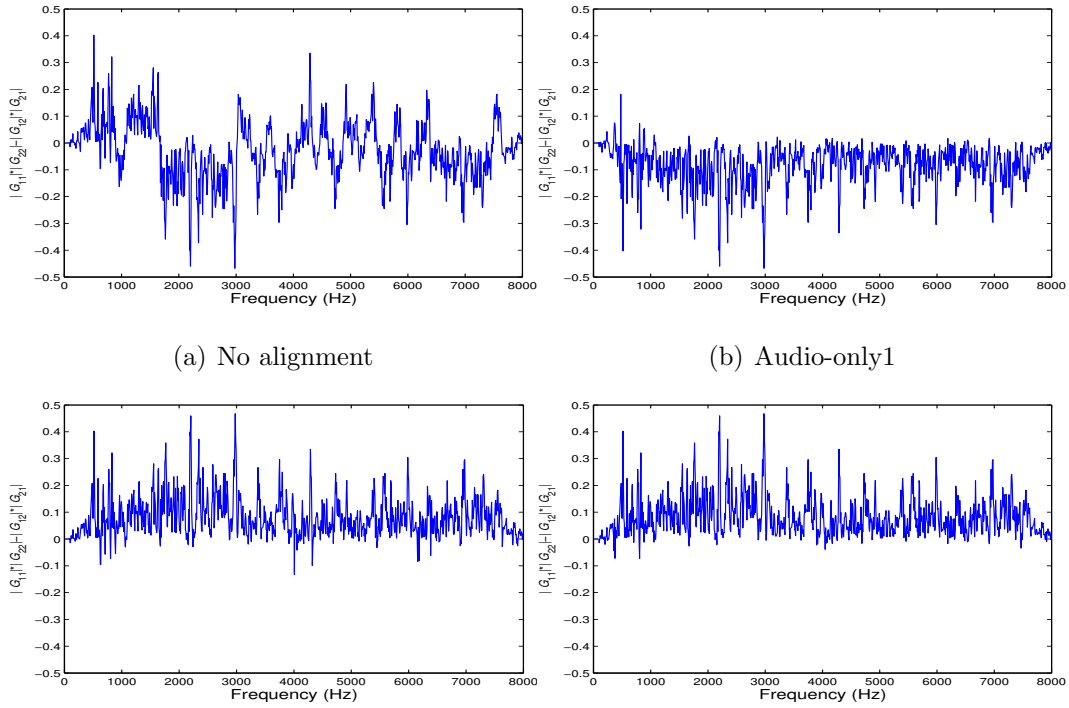


(a) No alignment



(b) Audio-visual

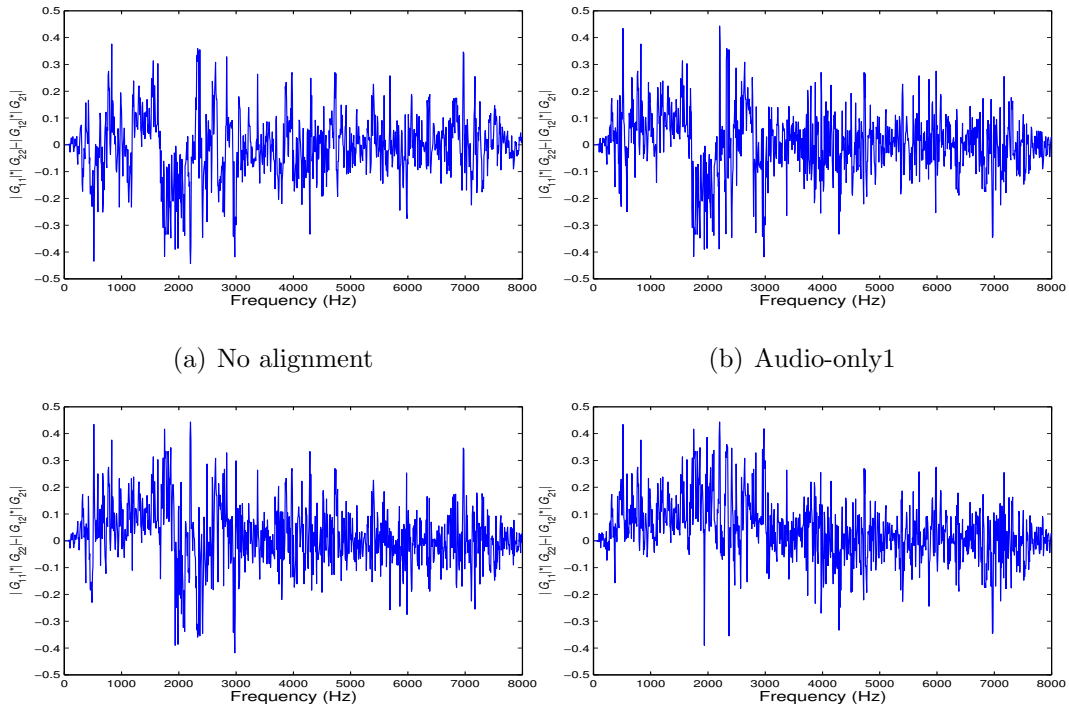
Figure 7: Global filters in the frequency domain obtained by the FD-BSS without permutation alignment (a) and the audio-visual alignment (b). The same mixtures and parameters as those in Figure 6 were used. The permutation ambiguities occur in many frequency bands in (a). Therefore, the spectrum of the recovered signal $y_1(t)$ contains distortions from the other source signal $s_2(t)$, which can also be observed in Figure 6(c). With audio-visual alignment, the permutation ambiguities have been significantly reduced, as shown in (b) **as well as Figure 6(d)**.



(a) No alignment (b) Audio-only1

(c) Audio-only2 (d) Audio-visual

Figure 8: **The plot of $|G_{11}| \cdot |G_{22}| - |G_{12}| \cdot |G_{21}|$ for the noise-free case, when no permutation alignment is conducted (a) and the permutations are aligned by Audio-only1 (b) in [30], Audio-only2 (c) in [33] and the audio-visual approach (d).** $|G_{11}| \cdot |G_{22}| - |G_{12}| \cdot |G_{21}|$ should be consistently larger or smaller than 0, if the permutations across the frequency bins are correctly aligned.



(a) No alignment

(b) Audio-only1

(c) Audio-only2

(d) Audio-visual

Figure 9: **The plot of $|G_{11}| \cdot |G_{22}| - |G_{12}| \cdot |G_{21}|$ for the case where 10 dB Gaussian white noise was added to the mixtures. The meanings of the sub-plots correspond similarly to those in the noise-free case shown in Figure 8.**

audio-only methods failed to group the spectral components accurately in the frequency band between approximately 1700 Hz and 3000 Hz, which lies in the range of frequencies that are essential to human intelligibility of speech. The audio-visual method provides more accurate alignment in this case. The results and comparisons in terms of SINR measurements are provided in the next section.

The magnitude spectra in Figures 7, 8, and 9 appear somewhat peaky, which implies that a certain amount of scale ambiguity remains despite having applied the MDP-based post-processing to the unmixing filters. However, our informal listening tests show that the effect of the peakedness of the frequency spectrum on the perceptual quality of the recovered speech signals is negligible. In other words, we didn't observe strong distortions of the signal as a result of the residual spectral peaks.

5.2.2. Comparisons

First, we compare the effect of different frequency partitions (i.e. the choice of I_{max}) on the performance of the proposed method for permutation alignment. To this end, we used the same experimental set-up as described in Sections 5.1 and 5.2.1, except the following two changes. First, we change the values of I_{max} by selecting it from [3 4 5 6]. Second, we perform 10 experiments in each of which the mixing filters were selected from the 10 sets of the room impulse responses described in Section 5.1. The average results over the 10 different sets of mixtures are shown in Table 1. From this table, it can be observed that the different number of frequency partitions does influence the separation performance. In general, $I_{max} = 4$ **or** 5**, which corresponds**

Table 1: The effect of the choice of I_{max} on the separation performance measured by SINR in dB

I_{max}	3	4	5	6
SINR(dB)	6.39	8.98	8.21	6.40

to 8 or 16 partitions respectively, gives reasonably good performance. More partitions however do not increase the system performance. We choose $I_{max} = 5$ for subsequent performance comparisons.

We then compare our algorithm with other de-permutation methods using only audio signals. We use the method in [30] (denoted as "Audio-only1") and the approach in [33] (denoted as "Audio-only2") as contrast algorithms for permutation alignment. Audio-only1 integrates information across different frequencies with the assumption that signal profiles in different bins undergo interrelated changes, even for distant frequency channels. Audio-only2 exploits the cross-frequency correlation between neighbouring frequency bands based on a hierarchical structure.

We evaluate the performance of our algorithm and the above baseline algorithms with respect to different signal to noise ratios (SNRs) and different FFT sizes. First, we fix the parameters as used in above sections, and only change the values of FFT size among [512 1024 2048 4096]. For each FFT size N, ten independent tests were run on the ten different sets of mixtures for each of the algorithms under comparison. No noise was added to these mixtures. The average SINR (dB) results are shown in Table 2. From this table, it can be observed that the choice of N has impacts on the separation performance. For example, a smaller N, such as $N = 512$, gives relatively lower performance for almost all the tested algorithms. Our proposed algo-

rithm offers competitive performance in all the test cases. $N = 4096$ appears to offer better performance for the majority of the algorithms. However, a higher computational cost is usually involved for a larger N when performing permutation alignment.

Table 2: SINR measurements for different FFT size (i.e. N).

N	512	1024	2048	4096
No alignment	2.19	1.85	3.53	3.87
Audio-only1	**4.23**	**6.36**	6.12	7.55
Audio-only2	2.38	2.76	**8.69**	**8.88**
Audio-visual	**4.15**	4.56	**8.21**	**8.98**

Then, we changed the SNR level from 5 dB to 30 dB while maintaining $N = 2048$, $I_{max} = 5$, and other parameters as set in the above two sections. For each SNR, again ten independent sets of tests were run for each of the algorithms under comparison. The average results measured by SINR are shown in Figure 10. It is shown from this figure that the proposed algorithm tends to offer better performance for a higher noise level. For example, when SNR = 10 dB, our algorithm provides 0.73 dB improvement over Audio-only1 and 1.75 dB over Audio-only2. For the noise-free case (not plotted on this figure), our algorithm gains 2.08 dB improvement over Audio-only1, but is 0.48 dB lower than Audio-only2. This suggests that the advantage of the audio-visual method is more prominent for noisy mixtures than the noise-free ones.

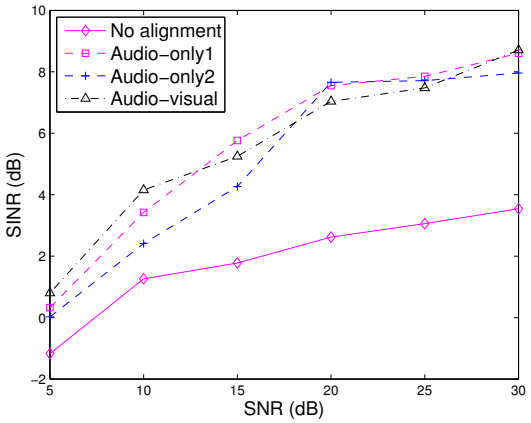


Figure 10: Average SINR computed over the independent tests for the 4 algorithms, where $N = 2048$ and $I_{max} = 5$.

6. Conclusions

We have presented a new audio-visual convolutive BSS system. In this system, we have used the MFCCs as audio features, which were combined with geometric visual features to form an audio-visual feature space. We have considered using some dynamic features in video for robust feature selection in audio-visual space. An adapted EM algorithm has been proposed by exploiting the different influences of the audio features for statistically modelling the audio-visual coherence. A new recursive sorting scheme based on the maximization of the audio-visual coherence has also been developed to solve the permutation problem.

Acknowledgement

This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) (Grant number EP/H012842/1) and the MOD Uni-

versity Defence Research Centre on Signal Processing (UDRC). **The authors would like to thank the anonymous reviewers and the guest editors for their insightful comments that considerably improved the quality of this paper.**

References

[1] D. A. Bulkin, J. M. Groh, Seeing sounds: visual and auditory interactions in the brain, IEEE Transactions on Neural Networks 16 (4) (2006) 415–419.

[2] J.-L. Schwartz, F. Berthommier, C. Savariaux, Seeing to hear better: evidence for early audio-visual interactions in speech identification, Cognition 93 (2004) B69–B78.

[3] D. Sodoyer, J.-L. Schwartz, L. Girin, J. Klinkisch, C. Jutten, Separation of audio-visual speech sources: a new approach exploiting the audio-visual coherence of speech stimuli, EURASIP Journal on Applied Signal Processing 2002 (11) (2002) 1165–1173.

[4] G. Monaci, P. Vandergheynst, F. T. Sommer, Learning bimodal structure in audio-visual data, IEEE Transactions on Neural Networks 20 (12) (2009) 1898–1910.

[5] A. L. Casanovas, G. Monaci, P. Vandergheynst, R. Gribonval, Blind audiovisual source separation based on sparse redundant representations, IEEE Transactions on Multimedia 12 (5) (2010) 358–371.

[6] S. G. Mallat, Z. Zhang, Matching pursuits with time-frequency dictionaries, IEEE Transactions on Signal Processing 41 (12) (1993) 3397–3415.

[7] D. Ormoneit, V. Tresp, Averaging, maximum penalized likelihood and bayesian estimation for improving gaussian mixture probability density estimates, IEEE Transactions on Neural Networks 9 (4) (1998) 639–650.

[8] G. Potamianos, C. Neti, G. Gravier, A. Garg, A. W. Senior, Recent advances in the automatic recognition of audio-visual speech, in: Proc. IEEE, 2003, pp. 1306–1326.

[9] M. Gurban, Multimodal speaker localization in a probabilistic framework, in: Proc. of EUSIPCO, 2006.

[10] L. Girin, J.-L. Schwartz, G. Feng, Audio-visual enhancement of speech in noise, The Journal of the Acoustical Society of America 109 (6) (2001) 3007–3020.

[11] Q. Liu, W. Wang, P. Jackson, Bimodal coherence based scale ambiguity cancellation for target speech extraction and enhancement, in: Proc. Interspeech, 2010, pp. 438–441.

[12] B. Rivet, L. Girin, C. Jutten, Mixing audiovisual speech processing and blind source separation for the extraction of speech signals from convolutive mixtures, IEEE Transactions on Audio, Speech and Language Processing 15 (1) (2007) 96–108.

[13] W. Wang, D. Cosker, Y. Hicks, S. Sanei, J. Chambers, Video assisted speech source separation, in: Proc. ICASSP, 2005, pp. 425–428.

[14] Q. Liu, W. Wang, P. Jackson, Audio-visual convolutive blind source separation, in: Proc. Sensor Signal Processing for Defence (SSPD), 2010.

[15] Q. Liu, W. Wang, P. Jackson, Use of bimodal coherence to resolve spectral indeterminacy in convolutive BSS, in: Proc. LVA/ICA, 2010, pp. 131–139.

[16] S. M. Naqvi, M. Yu, J. A. Chambers, A multimodal approach for blind source separation of moving sources, IEEE Journal Selected Topics in Signal Processing 4 (5) (2010) 895–910.

[17] P. Comon, Independent component analysis, a new concept?, Signal Processing 36 (3) (1994) 287–314.

[18] C. Jutten, J. Herault, Blind separation of sources, part i: An adaptive algorithm based on neuromimetic architecture, Signal Processing 24 (1) (1991) 1–10.

[19] J.-F. Cardoso, A. Souloumiac, Blind beamforming for non-gaussian signals, Radar and Signal Processing, IEE Proceedings F 140 (6) (1993) 362–370.

[20] A. J. Bell, T. J. Sejnowski, An information-maximization approach to blind separation and blind deconvolution, Neural Computation 7 (6) (1995) 1129–1159.

[21] A. Belouchrani, K. Abed-Meraim, J.-F. Cardoso, E. Moulines, A blind source separation technique using second-order statistics, IEEE Transactions on Signal Processing 45 (2) (1997) 434–444.

[22] A. Hyvärinen, E. Oja, A fast fixed-point algorithm for independent component analysis, Neural Comput. 9 (1997) 1483–1492.

[23] S. Amari, S. C. Douglas, A. Cichocki, H. H. Yang, Multichannel blind deconvolution and equalization using the natural gradient, in: Proc. IEEE Int. Workshop on Wireless Commun., 1997, pp. 101–104.

[24] J. Thomas, Y. Deville, S. Hosseini, Time-domain fast fixed-point algorithms for convolutive ICA, IEEE Signal Process. Lett. 13 (4) (2006) 228–231.

[25] T. Mei, J. Xi, F. Yin, A. Mertins, J. F. Chicharo, Blind source separation based on time-domain optimization of a frequency-domain independence criterion, IEEE Transactions on Audio, Speech, and Language Processing 14 (6) (2006) 2075–2085.

[26] J. Anemüller, B. Kollmeier, Amplitude modulation decorrelation for convolutive blind source separation, in: Proc. ICA, 2000, pp. 215–220.

[27] M. Z. Ikram, D. R. Morgan, A beamforming approach to permutation alignment for multichannel frequency-domain blind speech separation, in: Proc. ICASSP, 2002, pp. 881–884.

[28] F. Nesta, M. Omologo, P. Svaizer, A novel robust solution to the permutation problem based on a joint multiple TDOA estimation, in: Proc. IWAENC, 2008.

[29] K. Matsuoka, Minimal distortion principle for blind source separation, in: Proc. SICE, Vol. 4, 2002, pp. 2138–2143.

[30] D.-T. Pham, C. Servière, H. Boumaraf, Blind separation of speech mixtures based on nonstationarity, in: Proc. ISSPA, Vol. 2, 2003, pp. 73–76.

[31] H. Sawada, R. Mukai, S. Araki, S. Makino, A robust and precise method for solving the permutation problem of frequency-domain blind source separation, IEEE Transactions on Speech and Audio Processing 12 (5) (2004) 1063–6676.

[32] R. Mazur, A. Mertins, An approach for solving the permutation problem of convolutive blind source separation based on statistical signal models, IEEE Transactions on Audio, Speech and Language Processing 17 (1) (2009) 117–126.

[33] K. Rahbar, J. P. Peilly, A frequency domain method for blind source separation of convolutive audio mixtures, IEEE Transactions on Speech and Audio Processing 13 (5) (2005) 832–844.

[34] D. Sodoyer, B. Rivet, L. Girin, C. Savariaux, J.-L. Schwartz, C. Jutten, A study of lip movements during spontaneous dialog and its application to voice activity detectio, The Journal of the Acoustical Society of America 125 (2) (2009) 1184–1196.

[35] M. Jeub, M. Schafer, P. Vary, A binaural room impulse response database for the evaluation of dereverberation algorithms, in: 16th International Conference on Digital Signal Processing, 2009, data online: `http://www.ind.rwth-aachen.de/AIR`.

[36] XM2VTS, Website, `http://www.ee.surrey.ac.uk/CVSSP/xm2vtsdb/`.