Using the Bi-modality of Speech for Convolutive
Frequency Domain Blind Speech Separation

By Andrew Aubrey, Jonathan Lees, Yulia Hicks and Jonathon Chambers
Centre of Digital Signal Processing, Cardiff School of Engineering, Cardiff University, UK

Abstract
The problem of blind source separation for the case of convolutive mixtures of speech is considered. A novel algorithm is proposed that exploits the bi-modality of speech. This is achieved by incorporating joint audio-visual features into an existing BSS algorithm for the purpose of improving the convergence rate of the source separation algorithm. The increase in the rate of convergence when using a joint audio-visual model compared to using raw audio data (i.e. no model) is shown with simulations. The difference between using time varying (HMM) and stationary (GMM) statistical models to model the joint audio-visual features is also considered.

1. Introduction
The cocktail party problem [2] is one of separating a mixture of speech signals into their individual signals. Proposed solutions to the problem usually employ blind source separation (BSS) techniques based on one or more of the following: channel estimation, direction of arrival information or time-frequency information of the signals. Most BSS algorithms rely solely on audio information alone; however it is becoming increasingly apparent that the audio information alone is insufficient to separate the signals. Extra information regarding the sources is required. Speech is bi-modal; not only do we have audio information (captured using microphones) but using video cameras we can also capture facial information. In fact, it has been shown that the human auditory system uses visual and audio information to improve the intelligibility of a person’s speech in the presence of noise [6]. Visual information has also been utilised in speech enhancement and speech recognition systems [4]. There have been solutions proposed that use visual information for solving instantaneous mixtures of speech, an overview of these methods can be found in [1]. However, instantaneous mixtures do not represent the mixing that takes place in the cocktail party scenario; to represent this more accurately we must consider convolutive mixtures. Solutions using visual information to solve for convolutive mixtures of speech have been proposed by Wang et al. [7] and Rivet et al. [5]. In [7] the audio-visual information was used in a penalty function in a second order frequency domain BSS algorithm to estimate the separating matrix, and in [5] the audio-visual information was used to solve the permutation and scaling ambiguities.

In this paper, we propose a novel BSS algorithm where the audio-visual information is used to control the learning rate of the second order frequency domain algorithm outlined in [8].

2. Basics
BSS is the method of obtaining estimates of sources from a mixture of signals with no \textit{a-priori} information about the nature of the signals or the mixing environment. We can however assume that the sources are statistically independent and that at most one of them is Gaussian. The cocktail party scenario can be represented by the following equation:

\[ x(t) = \sum_{\tau=0}^{L} H(\tau) s(t-\tau) \]  

(1)

where \(x(t)\) are the mixtures detected at the microphones, \(s(t)\) are the original source signals, \(H\) is the mixing matrix and \(\tau\) is a delay. Also in this paper we do not consider additive noise. Instead of estimating \(H\) it is common to estimate a separating matrix \(W\), and applying this to the observed signals we obtain estimates \(\hat{s}\) of the source signals \(s\).
The size of the filters in convolutive mixtures can be in the order of 1000’s of samples, dependent upon the size of the room and the sampling frequency. For this reason time domain solutions are computationally expensive, and so frequency domain approaches are preferred as the convolutive mixing in the time domain is transferred to that of instantaneous mixing in the frequency domain. In the frequency domain we separate the signals at each frequency bin $\omega$ at discrete time $k$ using a backward discrete-time model:

$$\hat{S}(\omega, k) = W(\omega)X(\omega, k)$$  \hspace{1cm} (2)

However, it is worth noting that working in the frequency domain amplifies the permutation and scaling ambiguities associated with BSS.

The challenge in audio-visual BSS is to use the visual information without a significant increase in the computational complexity. Audio-visual BSS has been previously used for online estimation of the separation matrix [7] and also in post processing for solving the permutation and scaling ambiguities [5]. Other examples of audio-visual BSS can be found in [1]. All of the audio-visual methods have reported improvements over similar audio only methods.

3. Visual Features and Appearance Models

It is imperative that the chosen visual features have a high correlation with the audio information. For this reason, and because we are processing speech signals, we use the speaker’s lips as the visual feature as they are the most visible component of speech production. Previous work has used either lip shape or the internal lip width and height, as shown in Figure 1. In our work we use a combination of the lip shape and the texture content therein.

![a) Lip Height and Width](Fig. 1. Example of visual features extracted from the lips.)

To extract our required visual features we use an Active Appearance Model (AAM) [3] of the speaker’s lips. AAMs are a joint statistical model of shape and texture, where a single appearance parameter defines a corresponding shape and texture vector. First the lip shape is tracked through the video by placing landmarks (manually or automatically) on the outer edge of the lips in each frame (Fig.1b is an example of the landmarks for one frame, shown as dots on the connected line). We are then able to obtain a model of the shape variation ($z$) of the lips from the landmarks in each frame. The texture within each shape is obtained and a model of the texture variation ($g$) over the set of images is generated. We obtain the shape and texture models by performing PCA (Principal Component Analysis) on the shape and texture data separately, and this allows us to represent the shape and texture of any image in the set using the following models:

$$z(j) = z + P_s b_z(j), \quad \text{and} \quad g(j) = g + P_g b_g(j)$$  \hspace{1cm} (3)

where $z$ and $g$ are the mean shape and texture vectors. $P_s$ and $P_g$ are matrices formed from eigenvectors (obtained from the PCA operation). By varying the shape and texture parameters $b_z(j)$ and $b_g(j)$ we are able to approximate the shape and texture of any of the existing images. Next we concatenate the shape and texture vectors into a single vector set $\{b_{sg}(j)\}$, then the required appearance parameters $c(j)$ are obtained by performing PCA on the set of vectors $\{b_{sg}(j)\}$, and forming a matrix $P_c$ from a certain number of the resulting eigenvectors:

$$b_{sg}(j) = P_c c(j)$$  \hspace{1cm} (4)
Thus $\mathbf{c}(j)$ is a vector of appearance parameters describing shape and texture of a speaker’s mouth region at time $j$.

4. Incorporating Visual information into BSS

We wish to maximise the coherence between a set of visual features $v_i$ and a set of audio features $a_i$ to provide a criterion for controlling the learning rate of a second order frequency domain BSS algorithm. It should be noted that we do not use all of the appearance parameters $c$ obtained from (6); we use a dimensionally reduced vector $v$. For $N$ speakers we maximise the coherence in the following way:

$$J(W) = \arg \max_w \sum_{i=1}^{N} C(v_i, a_i)$$

(5)

To obtain the visual features we use an AAM of the speaker’s lips over a number of frames. For the corresponding audio feature we use MFCCs (Mel-cepstral coefficients) as they mimic the non-linear frequency resolution of the human ear. The audio and visual features are then concatenated to provide joint audio-visual feature vectors:

$$u = [v^T, a^T]$$

(6)

The probability distribution of $u$ is modelled by using either a Gaussian mixture model (GMM) or an Hidden Markov model (HMM). Training of the models is achieved using the same method as in [7]. In our experiments we compare the results of both models. Next we integrate the audio-visual information into the BSS algorithm. The frequency domain BSS algorithm we use is that proposed by Wang et al. [8]. In their approach a penalty function is incorporated into the cost function to constrain the gradient search. The penalty function has an associated learning rate $\mu_{IC}$. In [8] the learning rate is controlled by a function of the present penalty value. In the current work it is controlled by a function of the audio-visual coherence

$$\mu_{IC}(\omega) = \frac{\xi}{\zeta + f'(P_{uv})}$$

(7)

where $P_{uv}$ is the joint audio-visual probability (a measure of the coherence), $\xi$ and $\zeta$ are constants and $f'$ is a non-linear mapping.

The algorithm is performed using the following steps:

1. Estimate the source signals from the current estimate of $W(\tau)$ (for $\tau=0…k$) and calculate the audio features.
2. Concatenate the audio and visual features to form a new joint audio-visual feature.
3. Calculate the joint probability $P_{uv}$ using the model parameters from the GMM or HMM.
4. Calculate a new value for $\mu_{IC}$.
5. Update $W(\tau)$ (for $\tau=0…k$) until converged.

The algorithm is said to have converged when the change in value of $\mu_{IC}$ falls below a chosen threshold.

5. Experimental Comparison

The GMM and HMM were trained on the audio-visual features extracted from a video of a subject in an office environment with low level acoustic noise and artificial front on lighting. Visual data were captured at 25fps and the audio was sampled at 32kHz, 16-bit mono. The AAM of the lip region was described using 10 appearance parameters per frame and the speech features were extracted using Mel-cepstral analysis with a 20ms Hamming window, providing 12 MFCCs per frame. It was necessary to interpolate the appearance parameters (40ms) to retain one-to-one correspondence with the audio features (20ms). Only $2*2$ mixtures were considered, where the speech signal of the speaker present in the video was mixed with another speaker in a convolutive system with 9 taps. Several datasets were used for the simulations and Figure 2 shows the results from two of these datasets. It can clearly be seen that using visual features increases the convergence rate of the BSS algorithm. Moreover, when using the HMM to model the audio-visual features the algorithm converges quicker than the GMM model. This is assumed to be due to the
ability of the HMM to capture the co-articulation in speech. Finally, to ensure there is no loss in separation performance the signal to interference ratio (SIR) was calculated for each implementation (HMM, GMM and audio only) and found to be essentially identical. The real advantage afforded by our approach is that the increased learning rate would be very useful in a dynamic environment (e.g. where a speaker is moving).

![Graphs](image1.png)

**Fig. 2.** Comparison of learning rate for two datasets using (a) HMM, (b) GMM, (c) audio only to control the step size

6. Conclusion

While the results shown here are promising, audio-visual BSS is still a relatively new area and there are still many outstanding challenges, such as: extracting useful features from video, appropriate modelling of these features and incorporating the visual information into the BSS algorithm. Currently, visual lip feature extraction techniques are still in their infancy. Several methods are available but none are robust to both speaker identity and lip motion that is not speech related, and while modelling the audio-visual features is not always necessary it can improve robustness. Simulation results show an advantage of incorporating visual information into a BSS algorithm. However, for audio-visual BSS to be successful a technique is required that exploits the video data but is robust to wide range of speakers. This is the subject of our future work.

REFERENCES


