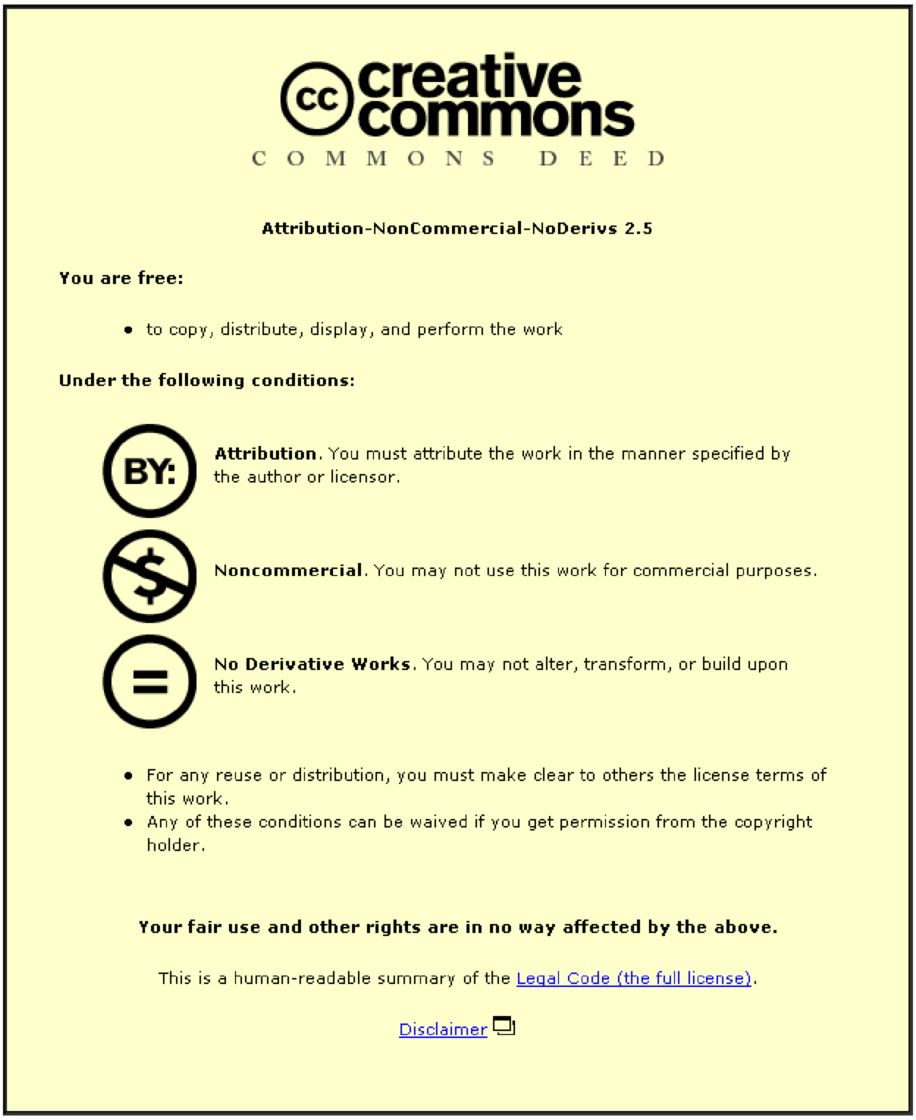


STUDY OF VIDEO ASSISTED BSS FOR CONVOLUTIVE MIXTURES

Andrew Aubrey, Yulia Hicks, Saeid Sanei and Jonathon Chambers

Centre of Digital Signal Processing, Cardiff School of Engineering,
Cardiff University, U.K.

ABSTRACT

In this paper we present an overview of recent research in the area of audio-visual blind source separation (BSS), together with new results of our work that highlight the advantage of including visual information into a BSS algorithm. In our work the visual information is combined with audio information to form joint audio-visual feature vectors. The audio-visual coherence is then modelled using statistical models. The outputs of these models are used within a frequency domain BSS algorithm to control the step size. Experimental results verify the improvement of the audio-visual method compared to audio only BSS. We also discuss visual feature extraction techniques, along with several recently published methods for audio-visual BSS, and conclude with suggestions for future research.

Index Terms— active appearance models, audio-visual processing, blind source separation, feature extraction.

1. INTRODUCTION

Blind source separation (BSS) is a method of separating the original source signals from a mixture of the sources with little or no information about the original source signals or the manner in which they were mixed. The most challenging BSS problem is separating convolutive mixtures of signals, which is more commonly known as "The Cocktail Party Problem". The cocktail party problem was first defined by Colin Cherry in 1953 [1] and is defined as: "*Suppose there is a room where several people are talking simultaneously. How does the listener recognize what one person is saying among the mixture of voices and background noise (e.g music)*?" The human auditory system is well adapted to such situations as it is able to use a wealth of information about the speaker and their surroundings, but from a signal processing perspective this problem has yet to be solved. Traditionally, BSS is performed using audio information alone, but the performance of these methods degrades in a noisy environment. Speech is a bi-modal signal, with both audio and visual aspects. It has been shown [2] that when we are able to see the speakers face, the intelligibility of that persons' voice in a noisy environment increases. The McGurk effect [3] also highlights the relationship between the audio and visual aspects of speech and how humans perceive speech. Petajan [4] also showed that audio-visual speech recognition performs better than audio only speech recognition.

The key challenges when using visual information to improve the performance of BSS are selecting the visual features that have a high correlation with the audio features and extracting that data. Encouraging results were obtained in early work by Girin et al. [5], where they used visual features to enhance speech embedded in noise by selecting filter parameters that were partially determined by visual information. Sodoyer et al. [6] extended this idea to combine audio-visual speech processing and blind source separation to form an early contribution to audio-visual source separation research. More recently Wang et al. [7] and Sodoyer et al. [8] have used visual information to help solve the convolutive case of BSS. However, audio-visual BSS is still in the early stages of research compared to audio only BSS; this paper gives an overview of the research area so far, and provides recently obtained results and suggestions for future research. The outline of this paper is as follows. Section 2 covers the basics of BSS, section 3 describes methods of extracting visual features, with an overview of active appearance models. In Section 4 we discuss previous approaches of audio-visual BSS. Sections 5 and 6 present a novel AVSS algorithm and results of simulations. We conclude in Section 7 by suggesting directions for future research.

2. BLIND SOURCE SEPARATION

As stated earlier, (BSS) is a method of extracting the source signals from a mixture of signals with no a-priori information about the nature of the signals or the mixing environment. Typically, BSS is performed using Independent Component Analysis (ICA)[9]. ICA is a statistical tool used to break down a set of random variables or signals into their independent components to reveal hidden factors of that data. However, ICA algorithms suffer from the permutation problem, i.e. indeterministic order of the estimated signals. To use ICA, we make an assumption that the sources are statistically independent.

This research was supported by funding from the EPSRC and QinetiQ as part of the Cardiff partnership project "SISTER". The authors also wish to thank Dr Darren Cosker and Dr Wenwu Wang for their advice.

1-4244-0535-1/06/$20.00/©2006 IEEE

The instantaneous form of BSS is defined as

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) + \mathbf{n}(t) \tag{1}$$

where **x**(t) represents the zero mean signal mixtures detected at the sensors at discrete time *t*, **s**(t) are the original source signals, **n**(t) is additive noise and **A** is the mixing matrix. In this work we assume **A** is not time varying and the number of sensor measurements is at least as large as the number of sources. However, in real room environments a signal is mixed with other source signals in the room together with reflections of both itself (which are caused by reverberations within the room) and other signals present. Thus the problem is now one of convolutive mixing:

$$\mathbf{x}(t) = \mathbf{A} * \mathbf{s}(t) + \mathbf{n}(t) \tag{2}$$

where the elements of the mixing matrix A are no longer scalar, they now represent a filter:

$$\mathbf{x}(t) = \sum_{\tau=0}^{k} \mathbf{A}(\tau)\mathbf{s}(t-\tau) \tag{3}$$

To recover the original sources from the mixtures we rewrite (3) to get:

$$\hat{\mathbf{s}}(t) = \sum_{\tau=0}^{k} \mathbf{W}(\tau)\mathbf{x}(t-\tau) \tag{4}$$

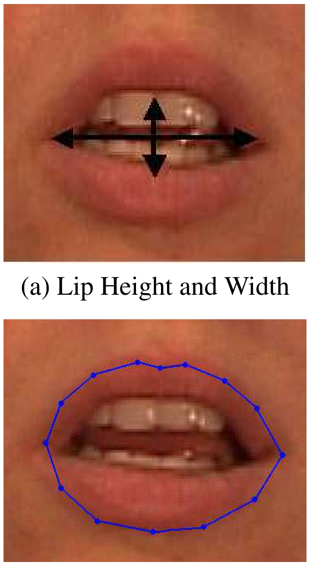
where $\hat{\mathbf{s}}$ represents the estimated source signals, $\mathbf{W}(\tau)$ (for τ=0...k) are the separation matrices and **x**(t) are the mixtures.

In convolutive mixing the length of the mixing filters can be of the order of 1000's of samples, dependent upon the size of the room and the sampling frequency. Thus time domain based approaches are computationally expensive, therefore frequency domain methods are preferred. In frequency domain approaches, the time domain convolutive mixture is turned into that of instantaneous complex mixing at each frequency bin. Unfortunately, working in the frequency domain amplifies the permutation problem meaning that we need to correctly separate each frequency bin.

The challenge in audio-visual BSS is to be able to use the visual information without significant increase in the computational complexity. There are a number of options when incorporating visual information within BSS: pre-processing the speech to remove unwanted noise, online to aid in finding the separating matrix or to post process the estimated speech sources to solve the permutation problem. So far all have been attempted, either on their own or in combination and all have reported improvements over similar audio only methods. Several such methods are discussed in section 4.

3. VISUAL FEATURE CHOICE AND EXTRACTION

A key challenge in audio-visual speech separation (AVSS) is to ensure that the chosen features have a high correlation with the audio information. As we are dealing with speech the most natural features to use would be those involved with the production of speech. The most visible components of speech production are the lips and they are the favoured feature in AVSS research. There are several lip features that have been used, such as lip height and width, lip shape and combined shape and texture information as shown in Fig 1.



(a) Lip Height and Width

(b) Lip Shape

Fig. 1. Example of visual features extracted from the lips.

3.1. Feature Extraction

Careful consideration needs to be given to extracting the lip features as a poor extraction method can lead to bad representation of the lip features. Visual features can be generally grouped into three main areas [10]: high-level lip contour based features, low level video pixel based and features that are a combination of both. High-level features consist of the inner or outer lip contour, which are then modelled in a statistical model, or alternatively geometric parameters such as lip height and width are used. Low-level features consist of using appropriate transforms, such as the discrete cosine transform, (DCT) on a region of interest (ROI), such as the speakers' mouth area and using the transformed pixel values as the features. There are many examples of methods where the low-level and high-level features are combined into one method to provide both shape and appearance features, such as the Active Appearance Model (AAM) [11]. To find more examples of feature extraction methods the reader is directed to the work by Potamianos et al. [10].

Sodoyer et al. [6],[8] extract the internal width and height of the lips using a chroma-key process and contour tracking on lips with blue makeup. Wang et al. [7] and Aubrey et al.

[12] use facial features found on the basis of an AAM. AAMs provide a statistical model of both shape and texture information of the lips, thus representing more information. Dansereau [13] captures visual information by combining motion, colour and edge information into a Markov random field (MRF) and extracting the three lip features; outer lip height, inner lip height and corner to corner mouth width. Rajaram et al. [14] implement a neural network based approach that uses a set of support vector machine classifiers to track the mouth and generate a binary sequence of visual features.

3.2. Basics of Active Appearance Models (AAMs)

Here, we present a brief overview of AAMs. A more detailed description of the method can be found in [11].

Cootes and Taylor [11] introduced active appearance models (AAMs) as a way of modelling selected features. An AAM is a joint statistical model of shape and grey-level parameters (texture), where a single appearance parameter defines a corresponding texture and shape vector. Our model is built in several stages. Firstly, the lip shape is tracked through the video by placing landmarks (manually or automatically) on the outer edge of the lips (Figure 1b is an example of this). Each landmark is represented with its cartesian coordinates (x_i, y_i), so for a single image, the vector $\mathbf{x}$ describing the lip shape is:

$$\mathbf{x} = (x_1, \ldots, x_N, y_1, \ldots, y_N) \tag{5}$$

For a given set of $\mathbf{j}$ images there are $\mathbf{x_1}, \ldots, \mathbf{x_j}$ such vectors. Next, we generate a statistical model of the shape variation from the landmarks. The mean shape is found and all images in the set are warped to the mean shape. We proceed by building a statistical model of the texture within the shape and apply principal component analysis (PCA) to shape and texture features separately to obtain:

$$\mathbf{x} = \bar{\mathbf{x}} + \mathbf{P_s b_s} \tag{6}$$

$$\mathbf{g} = \bar{\mathbf{g}} + \mathbf{P_g b_g} \tag{7}$$

where $\bar{\mathbf{x}}$ and $\bar{\mathbf{g}}$ are the mean normalised shape and grey-level vectors. $\mathbf{P_s}$ and $\mathbf{P_g}$ are matrices formed from eigenvectors, and $\mathbf{b_s}$, $\mathbf{b_g}$ are shape and grey-level parameters. By concatenating $\mathbf{b_s}$ and $\mathbf{b_g}$ and performing PCA, we obtain our required appearance parameters $\mathbf{c}$:

$$\mathbf{c} = \mathbf{P_c^T b} \tag{8}$$

where $\mathbf{P_c}$ is a set of matrices formed from eigenvectors describing to shape and texture and $\mathbf{b}$ are the combined parameters $\mathbf{b_s}$ and $\mathbf{b_g}$.

4. AUDIO-VISUAL ALGORITHMS

AVSS is still a new area of research, and so only a small number of papers have been published on using visual information to aid source separation. What follows is an overview of existing techniques that use visual information.

Early work by Sodoyer et al. [6] proposed a method that used a statistical model of the coherence of audio and visual speech features to estimate the separating matrix for the case of simple additive mixtures. A Gaussian mixture model (GMM) was built that provides the joint probability of a video vector (containing the visual features) and an audio vector containing spectral characteristics of the sound, which was then used to estimate the separating matrix. Rajaram et al. [14] developed a Bayesian framework for speech separation of a 2x2 linear mixture using a Kalman filter on the audio-visual observations, while Dansereau [13] used spectral matching of the audio-visual inputs to separate a mixture of speech for the case of a 2x2 first order decorrelation filter. These methods are concerned with simple mixtures of speech, but in a real environment the speech mixture is more complex. A convolutive mixing model best describes a realistic mixture.

Taking this into account, Rivet et al. [15] extended the audio-visual coherence idea in [6] for the case of convolutive speech mixtures. In [15] the speech signals were first separated using a time-frequency domain BSS algorithm. The results were then post processed using the audio-visual coherence model to solve the permutation and scaling ambiguities by cumulating the joint probability of audio-visual coherence for consecutive frames. Wang et al. [7] also exploited the audio-visual coherence of speech. In [7] the visual features were found on the basis of an AAM, and the speech features are extracted using Mel-frequency cepstral coefficients (MFCC's). A collection of joint audio-visual vectors is formed and modelled using a GMM. At each iteration the joint audio-visual probability between the estimated source and the video parameters is calculated using the above GMM and used as a penalty to estimate the separating matrix in a second order penalty function based BSS algorithm [16]. More recently, Aubrey et al. [12] presented a novel algorithm that uses an audio-visual model to control the learning rate of the penalty function BSS algorithm in [16]. In the experiments the audio-visual coherence was initially captured using a GMM, and then in later experiments the GMM was substituted for an Hidden Markov Model (HMM) to capture the coarticulation in speech (Section 5).

An alternative use of audio-visual information is presented in [8]. There the visual information was used for a voice activity detector (VAD). The VAD finds the silence periods in a speech signal by using the change in lip height and width at each frame. If the change over a number of frames is greater than a threshold value then it is said that the segment was speech, if not then it was assumed to be a silence period. However, this method assumes that when a person is not speaking, their lips remain stationary, which is not always the case.

5. AUDIO-VISUAL EXPERIMENTS

The basis of our approach is to maximise the coherence C 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 [16]. It should be noted that we do not use all of the appearance parameters c, we use a dimensionally reduced vector v.

For N speakers we maximise the coherence in the following way:

$$J(W) = \arg\max_{W} \sum_{i=1}^{N} C(a_i, v_i) \tag{9}$$

An AAM is used to obtain a model of the visual features of the speakers lips over a number of frames. To extract the corresponding audio features we use Mel-cepstral coefficients (MFCCs). MFCCs were chosen for their ability to mimic non-linear frequency resolution of the human ear. The audio and visual features are concatenated to provide joint audio-visual feature vectors:

$$\mathbf{u} = [\mathbf{v}^T, \mathbf{a}^T] \tag{10}$$

The probability distribution of $\mathbf{u}$ can be modelled using either a GMM or an HMM. Training of the model is achieved using the same method as in [7]. For our experiments we compare the results using both models.

Next we integrate the audio-visual information into a BSS algorithm. For our experiments we used a penalty function based frequency domain BSS algorithm [16]. In the original algorithm of [16] the learning rate μ_{Jc} 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.

$$\mu_{Jc}(\omega) = \frac{\xi}{\zeta + f'(P_{av})} \tag{11}$$

where P_{av} is the joint audio-visual probability (a measure of the coherence), ξ and ζ are constants and f' is a certain non-linear mapping [12]. It is necessary to calculate P_{av} using a different method when using a GMM or HMM to model the training data. For the case of a GMM we have:

$$p(\mathbf{u}_s) = \sum_{i=1}^{K} \mathbf{w}_i \frac{exp\{-1/2(\mathbf{u}_s - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1}(\mathbf{u}_s - \boldsymbol{\mu}_i)\}}{\sqrt{(2\pi)^K \mid \boldsymbol{\Sigma}_i \mid}} \tag{12}$$

where $\boldsymbol{\mu}_i$, $\boldsymbol{\Sigma}_i$, $\mathbf{w}_i$ and K are the mean vector, covariance matrix, kernel weights and the number of Gaussian kernels respectively. P_{av} is then found by summing the log of (12) and for the HMM the log probabilities were calculated using the method in [17].

The AVSS algorithm is performed using the following steps:

1. Estimate the source signals from the current estimate of $\mathbf{W}(\tau)$ (for τ = 0...k) and calculate the audio features.
2. Concatenate the audio feature with the visual features to form a new joint audio-visual feature.
3. Calculate the joint probability P_{av} using either the GMM or HMM model parameters.
4. Calculate a new value for (11).
5. Update $\mathbf{W}(\tau)$ (for τ = 0...k) until converged.

The algorithm is said to have converged when the change in value of μ_{Jc} falls below a chosen threshold.

6. RESULTS

The statistical models (GMM, 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. Video data were captured using a digital video camera at 25fps and the audio was captured using a directional microphone, and sampled at 32KHz, 16-bit mono. The lip region in the video was tracked using an AAM to provide a joint model of shape and texture information with 10 appearance parameters per frame. The speech features were extracted using Mel-cepstral analysis with a 20ms Hamming window, providing 12 MFCCs per frame. The appearance parameters (40ms) were then interpolated in order to retain one-to-one correspondence with the audio parameters (20ms). The number of Gaussian kernels for the GMM and the number of states for the HMM were set to 10. Finally, the audio-visual feature space had 22 dimensions, 10 video plus 12 audio and remained the same size during separation.

Only 2x2 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. Figure 2 shows the results of the simulations. It can be seen that the audio-visual model requires fewer iterations to converge (the end of the curve denotes convergence of the BSS algorithm), hence there is a lower overall complexity which is very likely to be useful in a non-stationary environment when the speaker is moving. Furthermore, the advantage of using an HMM compared to a GMM was also observed. This could be contributed to the fact that HMMs are better able to capture the co-articulation of speech. The quality of the reconstructed sources was judged subjectively by listening tests to be essentially identical for the 3 methods.

7. CONCLUSION

Using video to aid BSS is still a new area. The first steps in using video to control the search parameters in a BSS algorithm show promising results. Experimental results indicate that by combining audio and visual information we achieve a faster learning rate. The results were confirmed by using several different datasets to evaluate the method, although due to limited space we only present one example.

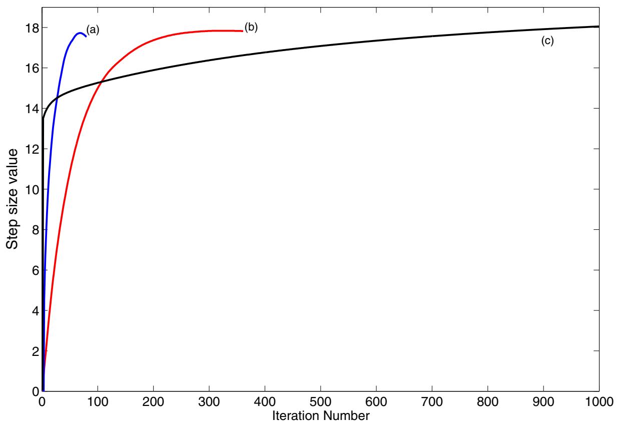


Fig. 2. Comparison of learning rate using (a)HMM, (b)GMM, (c)audio only to control the step size

Nonetheless there are still many outstanding challenges for audio-visual BSS:

- Finding the best method for extracting the chosen visual features.
- Appropriate modelling of the audio-visual features.
- Incorporating the visual information into the BSS algorithm.

Currently, visual lip feature extraction techniques are in their infancy. There are several methods available depending on what features you require but none are robust to both speaker identity and to lip motion that is not speech related (e.g. people often smile during a conversation and this can be wrongly classified as speech). Modelling the audio-visual features is not always necessary but can improve the robustness. For audio-visual blind source separation to be successful, what is required is a technique that exploits the video data but that is robust to a wide range of possibly moving speakers, and this is the subject of our future work.

8. REFERENCES

[1] C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *The Journal Of The Acoustical Society Of America*, vol. Vol 25, no. 5, pp. 975–979, September 1953.

[2] W. Sumby and I. Pollack, "Visual contribution to speech intelligibility in noise," *Journal Acoustical Society of America*, vol. 26, pp. 212–215, 1954.

[3] H. McGurk and J. McDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, pp. 746–48, December 1976.

[4] E.D. Petajan, *Automatic Lipreading to Enhance Speech Recognition*, Ph.D. thesis, University of Illinois, 1984.

[5] L. Girin, J.L. Schwartz, and G. Feng, "Audio-visual enhancement of speech in noise," *Journal of the Acoustical Society of America*, vol. Vol 109, no. 06, pp. 30007–3020, 2001.

[6] D. Sodoyer, L. Girin, C. Jutten, and J.L. Schwartz, "Developing an audio-visual speech source separation algorithm," *Speech Communication*, vol. 44(1-4), pp. 113–125, 2004.

[7] W. Wang, D. Cosker, Y. Hicks, S. Sanei, and J. Chambers, "Video assisted speech source separation," *ICASSP*, Philadelphia, 2005.

[8] D. Sodoyer, B. Rivet, L. Girin, J.L. Schwartz, and C. Jutten, "An analysis of visual speech information applied to voice activity detection," *ICASSP*, Toulouse, 2006.

[9] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, John Wiley and Sons, 2001.

[10] G. Potamianos, N. Chalapathy, J. Luettin, and I. Matthews, "Chapter 10: Audio-visual automatic speech recognition: An overview," *In Audio-Visual Speech Processing, MIT Press*, September, 2005.

[11] T.F. Cootes and C.J. Taylor, "Statistical models of apperance for computer vision," *Available from T.F.Cootes webpage, http://www.isbe.man.ac.uk/ bim/.*

[12] A. Aubrey, J. Lees, Y. Hicks, and J. Chambers, "Using the bi-modality of speech for convolutive frequency domain blind source separation," *IMA 7th International Conference on Mathematics in Signal Processing (Accepted for publication)*, December, 2006.

[13] R.M. Dansereau, "Co-channel audiovisual speech separation using spectral matching constraints," *ICASSP*, Montreal, Canada, 2004.

[14] S. Rajaram, A.V. Nefian, and T.S. Huang, "Bayesian separation of audio-visual speech sources," *ICASSP*, Montreal, Canada, 2004.

[15] B. Rivet, L. Girin, and C Jutten, "Solving the indeterminations of blind source separation of convolutive speech mixtures," *ICASSP*, Philadelphia, 2005.

[16] W. Wang, S. Sanei, and J. Chambers, "Penalty function based joint diagonalization approach for convolutive blind source separation of nonstationary sources," *IEEE Transactions on Signal Processing*, vol. 53, no. 05, pp. 1654–69, May 2005.

[17] L.R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. Vol 77, no. Issue 2, pp. pp. 257–286, 1989.