BAYESIAN SEPARATION OF AUDIO-VISUAL SPEECH SOURCES

Shyamsundar Rajaram[1], *Ara V. Nefian*[2] *and Thomas S. Huang*[1]

[1]University of Illinois at Urbana Champaign
Image Formation and Processing Group, Urbana, IL

[2] Intel Corporation
Architecture Research Labs, Santa Clara, CA

ABSTRACT

In this paper we investigate the use of audio and visual rather than only audio features for the task of speech separation in acoustically noisy environments. The success of existing independent component analysis (ICA) systems for the separation of a large variety of signals, including speech, is often limited by the ability of this technique to handle noise. In this paper, we introduce a Bayesian model for the mixing process that describes both the bimodality and the time dependency of speech sources. Our experimental results show that the online demixing process presented here outperforms both the ICA and the audio-only Bayesian model at all levels of noise.

1. INTRODUCTION

Robust separation of speech sources [1, 2] has a wide range of applications from automatic speech recognition to robotics or indexing. However, the performance of most of these systems decays rapidly in noisy environments. Recently, the use of visual features in audio-visual speech separation (AVSS) systems [3, 4] has shown encouraging results as it provides a set of unmixed observations independent from the acoustic noise. The key challenges in AVSS systems are finding a set of visual features that are highly correlated with the acoustic features and an appropriate model for the mixing of the audio-visual sources. The AVSS system presented in this paper and illustrated in Figure 1 combines a set of visual features extracted from the mouth region, with the audio samples obtained from a microphone array and uses a novel statistical model for the mixing process. The outline of this paper is as follows. In Section 2 we describe the extraction of the visual features. In Section 3 we describe the audio-only source separation models and in Section 4 we introduce a new mixing model for the audio-visual sources and compare it with previous models. In Section 5 we present the experimental results obtained using these mixing models. In Section 6 we discuss the conclusions of this work and suggests directions for future research.

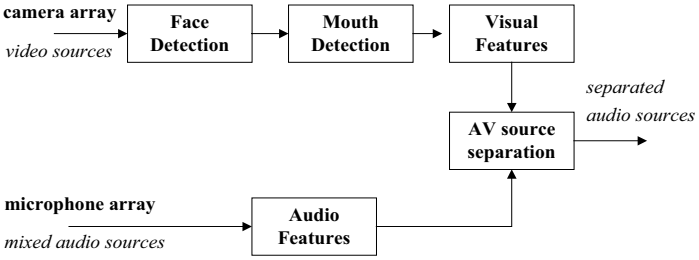

Figure 1: The overall audio-visual speech separation system.

2. THE VISUAL FEATURE EXTRACTION

One key challenge in AVSS is to obtain a set of visual features that can describe the acoustic data with high accuracy. While in lip reading the visual features can be enough for trained individuals or machines to recognize speech, a perfect reconstruction of the audio data from video is impossible even for humans. This is because most of the relevant audio information is contained in the frequency domain which cannot be estimated from video sequences. Fisher et. al [3] described a mutual information maximization criterion for audio and visual feature selection that projects both the audio and visual features in a low-dimensional space. The method does not allow the audio sequences to be perfectly reconstructed from their projections in the lower dimensional space, although encouraging separation results were obtained.

In our system, the extraction of the visual features starts with a neural network-based face detection system followed by the detection and tracking of the mouth region using a set of support vector machine classifiers (Figure 1). We described this method in detail in [5]. Next, we calculate the number of pixels within the mouth region for which the

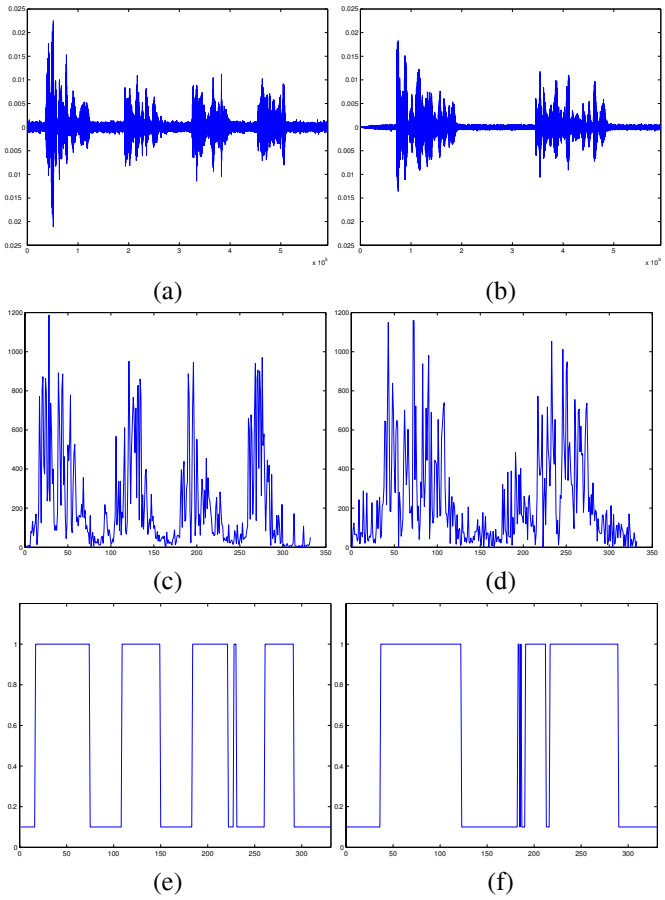

Figure 2: Audio source signals (a,b), the corresponding number of pixels within the mouth region for which the absolute difference at consecutive frames increases a fixed threshold (c,d), and the visual observations for the above source signals (e,f).

absolute difference at consecutive frames increases a fixed threshold (Figure 2c and d). These sequence are next low pass filtered with a filter of size 9, and thresholded to generate a binary sequence of visual features. As shown in Figure 2 the visual features described here can discriminate well between silence and speech periods. The resulting visual observations are upsampled to match the frequency of the audio samples and used together with the audio samples in the de-mixing process as explained in the next sections.

3. AUDIO-ONLY SOURCE SEPARATION

Let N be the number of sources, T be the number of audio samples, and M the number of microphones. Then, $\mathbf{S} = \{\mathbf{s}_t|\mathbf{s}_t = [s_{1t}, \ldots s_{Nt}]^T\}$ and $\mathbf{X} = \{\mathbf{x}_t|\mathbf{x}_t = [x_{1t}, \ldots x_{Nt}]^T\}$, are the sequence of the unmixed sources and mixed audio sequences from the microphones, respectively. The ICA method [2] uses a natural gradient technique [6] to maximize the mutual information criteria between the sources [1]. The method assumes that the sources are statistically independent (Equation 1) and characterized by a non-Gaussian density function described in Equation 2. The mixing process assumes that the mixed signals are a linear combination of the original source, with no additional noise (Equation 3). In addition, ICA requires that the number of microphones equals the number of sources.

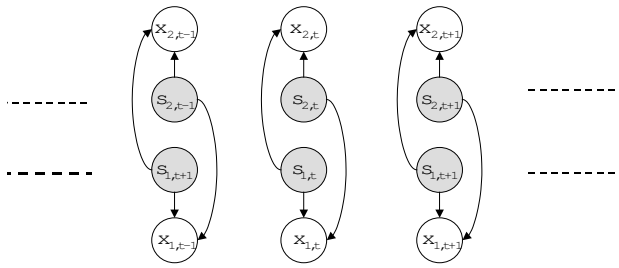

Figure 3: The source mixing model for ICA.

$$P(s_{1t}, \ldots, s_{Nt}) = \prod_{i=1}^{N} P(s_{it}) \quad (1)$$

$$P(s_{it}) = \frac{1}{\exp(s_{it}) + \exp(-s_{it})} \quad (2)$$

$$P(\mathbf{x}_t|\mathbf{s}_t, \mathbf{A}) = \delta(\mathbf{x}_t - \mathbf{A}\mathbf{s}_t) \quad (3)$$

Figure 3 illustrates the ICA mixing model over time as a Bayesian network where the transparent nodes represent the microphone observations, the grayed nodes represent the hidden unmixed sources, and the edges represent the conditional dependencies between observation and source nodes.

3.1. Audio-only Bayesian source separation

The Bayesian mixing model for source separation [7, 8, 9] relaxes the ICA assumptions by handling noise and different number of sources and microphones. The source signals are modelled using a Gaussian density function with zero mean and variance $\mathbf{C}_s$ (Equation 4) while the observations are described by a Gaussian density function with variance $\mathbf{C}_x$ (Equation 6). The mixing model used in this paper for audio-only Bayesian source separation (ABSS) also considers the first order temporal dependencies for each of the sources (Figure 4). The temporal dependencies of s_{it} from its previous sample s_{it-1} are described by a Gaussian density function with variance $\mathbf{C}_{ss}$ and mean bs_{it-1}, where b is a constant (Equation 5).

$$P(s_{i0}) \sim N(0, \mathbf{C}_s) \quad (4)$$

$$P(s_{it}|s_{it-1}) \sim N(, bs_{it-1}, \mathbf{C}_{ss}) \quad (5)$$

$$P(x_{it}|\mathbf{A}, s_{1t}, \ldots, \mathbf{s}_{Nt}) \sim N(\sum_j a_{ij}s_{jt}, \mathbf{C}_x) \quad (6)$$

It can be seen that the above equations describe a Kalman filter with the additional independence assumptions for the sources. With these constraints, the parameters of the mixing model ($\mathbf{A}, b, \mathbf{C}_s, \mathbf{C}_{ss}$ and $\mathbf{C}_x$) are learned online using the maximum likelihood estimation method in [10]. Next, the unmixed signals are obtained using the constrained Kalman filter (Equations 4-6)with the learned parameters.

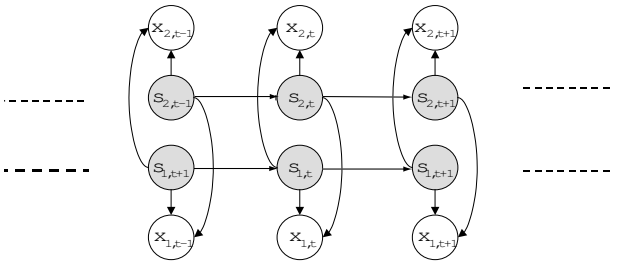

Figure 4: The source mixing model for ABSS.

4. AUDIO-VISUAL BAYESIAN SOURCE SEPARATION

The use of visual features in speech source separation is motivated by the audio and visual nature of speech and by the orthogonality of visual speech to the acoustic noise. However, finding robust statistics that describe the nonlinearity between audio and visual features remains the key challenge in audio visual source separation. In [3], the authors derived a space of maximum mutual information between audio and visual features and obtained the audio-visual features from the projections on this space. Sodoyer et al. [4] used a Bayesian framework to model the joint distribution of the audio and visual features using a set 32 Gaussian density functions trained off line from a large number of examples.

In our audio-visual Bayesian source separation (AVBSS) approach we extend the mixing model described in the previous section with a set of nodes that describe the audio-visual observations (Figure 5). The audio-visual observations $\mathbf{z}_{it} = [w_{it}x_{1t}, \ldots, w_{it}x_{Mt}]^T, i = 1 \ldots, N$, $t = 1, \ldots, T$ are obtained as the multiplications between the mixed audio observations $x_{jt}, j = 1 \ldots, M$ and the visual features w_{it} described in Section 2 and illustrated in Figure 2 e and f. This choice of the audio-visual observations improves the acoustic silence detection, by allowing a sharp reduction of the audio signal when no visual speech is observed. Formally, the audio-visual speech mixing process is given by the Equations 4- 6, and

$$P(\mathbf{z}_{it}|\mathbf{s}_t) \quad \sim \quad N(\mathbf{V}_i\mathbf{s}_t, \mathbf{C}_z) \tag{7}$$

where $\mathbf{V}_i$ is a $M \times N$ matrix, and $\mathbf{C}_z$ is the covariance matrix of the audio-visual observations. As with the audio only mixing model, the audio-visual Bayesian mixing model can be seen as a Kalman filter with the source independence constraints (Equation 1). In learning the model parameters, the whitening of the audio observations provides an initial estimate of the $\mathbf{A}$. The model parameters $\mathbf{A}, \mathbf{V}_i, b_i, \mathbf{C}_s, \mathbf{C}_{ss}, \mathbf{C}_z$) are learned on-line using the maximum likelihood estimation method in [10]. Finally, the sources are estimated using a constrained Kalman filter and the learned parameters.

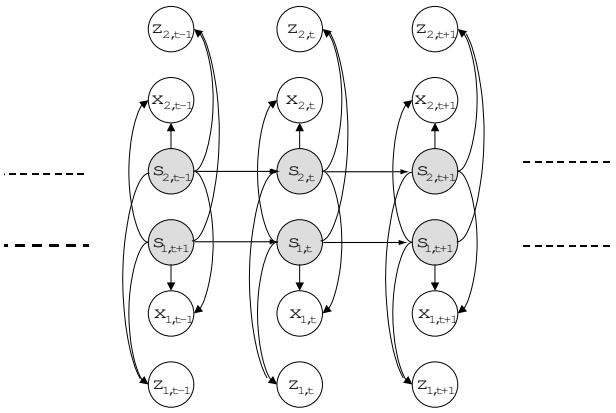

Figure 5: The source mixing model for AVBSS.

5. EXPERIMENTAL RESULTS

The audio and audio-visual source separation methods described in the previous sections were tested using two audio-visual sequence uttered by different people ("...I would like to know the weather in Beijing today....Is it going to rain in the afternoon after 4:00pm? ... What is the forecast for tomorrow in San Francisco, California... ?" and "...May I speak to Catherine please?...This is a friend from work...What time is she returning?...Can I leave a message?..."). The sequences were captured in an office environment with a low level of acoustic noise. To simulate noisy acoustic conditions and a scenario with two microphones the original sequences were mixed with a 2×2 matrix $A = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}$ and white noise was added to the mixed signals at different signal to noise ratios (SNR). The online source separation using the ABSS and AVBSS models were implemented using the Bayes Net toolbox [11]. Figure 6 shows the input mixed signals at SNR = =-4db and the output signals obtained using AVBSS for the source signals showed in Figure 2(a,b). Figure 7 illustrates the signal to noise and interference ratio (SNIR) of the input signals, ICA, ABSS and AVBSS for different levels of the input SNR. It can be seen that at all levels of SNR, the AVBSS outperforms ABSS and ICA confirming the intuition that features of visual speech improve the separation of the input mixed signals.

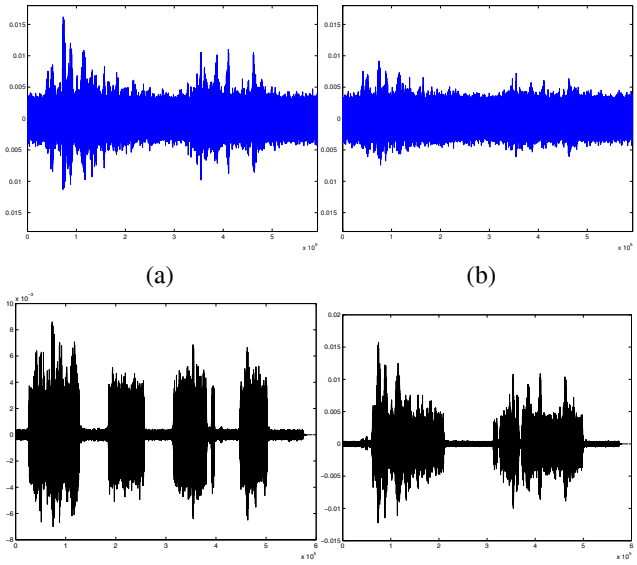

Figure 6: An example of the audio input signals at an average SNR=-4db (a,b) and the separated signals using AVBSS (c,d).

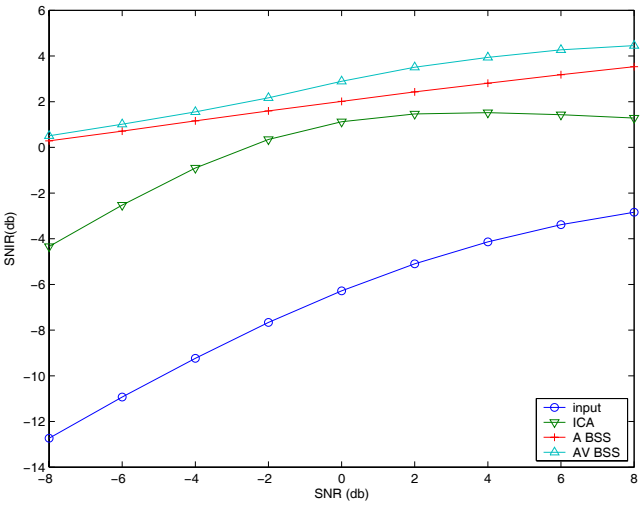

Figure 7: The signal-to-noise and interference ratio for the input signals, ICA, audio-only and audio-visual Bayesian source separation at different levels of signal-to-noise ratio of the input signals.

6. CONCLUSIONS

In this paper we proposed a Bayesian framework for speech source separation in noisy environments that uses both audio and visual observations. The model for the source mixing process presented in this paper describes the visual observations and the first order temporal dependencies between audio samples. The parameters of the mixing process are learned online using a maximum likelihood technique. The sources are estimated through a constrained Kalman filtering scheme of the visual observations and mixed audio signals. The audio visual source separation technique presented in this paper outperforms the audio only Bayesian approach and the ICA method for a wide range of input SNR levels. The success of the current approach relies on the accurate modelling of the source mixing process and the selection of visual observations that can discriminate robustly between periods of active speech and silence. The performance of the proposed approach comes however with a very high computational complexity cost compared to ICA methods. A real-time implementation of the proposed algorithm remains an important issue to be addressed in future work.

We are currently investigating visual features that better correlate with the acoustic speech sources and nonlinear techniques to handle the audio-visual dependencies. Future research includes audio-visual source separation for scenarios where the number of microphones is smaller than the number of sources and the extension of the current framework to audio-visual blind deconvolution.

7. REFERENCES

[1] A. J. Bell and T. J. Sejnowski, "An information maximization approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, November 1995.

[2] Kevin H. Knuth, "A Bayesian approach to source separation," in *Independent Component Analysis*, 1999, pp. 283–288.

[3] John Fisher III, Trevor Darrell, William T. Freeman, and Paul Viola, "Learning joint statistical models for audio-visual fusion and segregation," in *Advances in Neural Information Processing Systems*, 2000.

[4] D. Sodoyer, J-L Schwartz, L. Girin, J. Klinkisch, and 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*, pp. 1165–1173, 2002.

[5] L. Liang, X. Liu, X. Pi, Y. Zhao, and A. V. Nefian, "Speaker independent audio-visual continuous speech recognition," in *International Conference on Multimedia and Expo*, 2002, vol. 2, pp. 25–28.

[6] S. Amari, A. Cichocki, and H. H. Yang, "A new learning algorithm for blind signal separation," in *Avnaces in Neural Information (Cambridge,MA: MIT Press)*, 1996, pp. 757–763.

[7] Daniel B. Rowe, "A Bayesian approach to blind source separation," 2002.

[8] E. Wan and A. Nelson, "Neural dual extended kalman filtering: applications in speech enhancement and monaural blind signal separation," 1997.

[9] Ding Liu, Xiaoyan Liu, Fucai Qian, and Han Liu, "Blind source separatio based on dual adaptive control," in *4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003), Nara, Japan*, April 2003, pp. 445–450.

[10] V. Digalakis, J. R. Rohlicek, and M. Ostendorf, "Ml estimation of a stochastic linear system with the em algorithm and its applications to speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 1, no. 4, October 2002.

[11] ," Bayes Net Toolbox for Matlab, `http://www.cs.berkeley.edu/ murphyk/Bayes/bnt.html`.