VIDEO ASSISTED SPEECH SOURCE SEPARATION

Wenwu Wang†, Darren Cosker‡, Yulia Hicks†, Saeid Sanei†, and Jonathon Chambers†

† Cardiff School of Engineering, Cardiff University, CF24 0YF, U.K.
E-mails: [wangw2, hicksya, saneis, and chambersj]@cf.ac.uk
‡ Cardiff School of Computer Science, Cardiff University, CF24 3AA, U.K.
Email: D.P.Cosker@cf.ac.uk

ABSTRACT

In this paper we investigate the problem of integrating the complementary audio and visual modalities for speech separation. Rather than using independence criteria suggested in most blind source separation (BSS) systems, we use the visual feature from a video signal as additional information to optimize the unmixing matrix. We achieve this by using a statistical model characterizing the nonlinear coherence between audio and visual features as a separation criterion for both instantaneous and convolutive mixtures. We acquire the model by applying the Bayesian framework to the fused feature observations based on a training corporus. We point out several key exisiting challenges to the success of the system. Experimental results verify the proposed approach, which outperforms the audio only separation system in a noisy environment, and also provides a solution to the permutation problem.

1. INTRODUCTION

In the past decade, BSS has attracted tremendous research interests in the signal processing community. The success of BSS algorithms has made solutions to many problems possible, including the *cocktail party* problem. For this problem, it is classically addressed within the framework of convolutive BSS or its more effective implementation in a transform domain, such as frequency domain BSS [1]. However, the performance of the algorithms based on BSS highly depends on the acoustic condition. Many, if not most, of them tend to degrade considerably in a noisy environment.

Interestingly, rather than using only audile organs, humans are able to infer the meaning of spoken sentences by reading the movement of mouth and facial muscles. Indeed, human speech is inherently bimodal: *audio* and *visual* [2], in both production and perception. At a cocktail party, the visual modality, such as lipreading, helps the people to separate speech from background noise and multiple competing speakers to some extent [3]. There have been attempts in exploiting the visual information for speech separation [4], [5]. However, there are several open issues demanding research effort, such as addressing the convolutive mixtures and noisy mixtures. This paper, therefore, takes into account the intrinsic coherence between audition and vision in speech separation. The details of the approach combining audio and visual modalities will be presented in Section 2. Experimental results are given in Section 3 and the paper is concluded in Section 4.

This work was supported by EPSRC of the U.K in the form of the project GR/R69228/02 "Blind Signal Processing for Multichannel Speech Enhancement".

2. THE APPROACH

2.1. Problem Formulation and System Structure

Let's consider an acoustic application. A set of mixtures $\mathbf{x}(n) \in \mathbb{R}^M$, observed at M microphones with each picking up a weighted components of N source speeches $\mathbf{s}(n) \in \mathbb{R}^N$, can be modelled as

$$\mathbf{x}(n) = \mathbf{H} * \mathbf{s}(n) + \mathbf{e}(n), \tag{1}$$

where $\mathbf{e}(n) \in \mathbb{R}^M$ denotes the possible additive noise, $*$ represents the mixing operation (multiplication or convolution) and n is discrete time index, being omitted hereafter for notation simplicity if not specified. An initial assumption of BSS is having the sources independent so that we can apply a separation matrix $\mathbf{W} \in \mathbb{R}^{N \times M}$ to conduct the unmixing operation

$$\mathbf{y}(n) = \mathbf{W} * \mathbf{x}(n), \tag{2}$$

where $\mathbf{y}(n)$ are the separated signals, assuming to be the estimates of sources, i.e. $\mathbf{y} = \hat{\mathbf{s}}$.

In this study, we present a new separation system which imitates the humans in speech perception by exploiting the coherence between audio and visual modalities. Fig. 1 shows its functional block diagram. In this system, the audio and visual modalities are integrated at the feature level (or early stage) in which both stimuli are synchronized and merged, by concatenating or averaging the vocal tract functions, for joint learning and separation. We elaborate on the functionality of Fig. 1 in the later sections.

2.2. Audio-visual Coherence for Separation

To begin with, let us consider the separation problem of instantaneous mixtures. Suppose we have obtained, by applying unmixing operation $\mathbf{W}$, a separated signal y_i, which is equivalent to s_i, with the only difference in amplitude

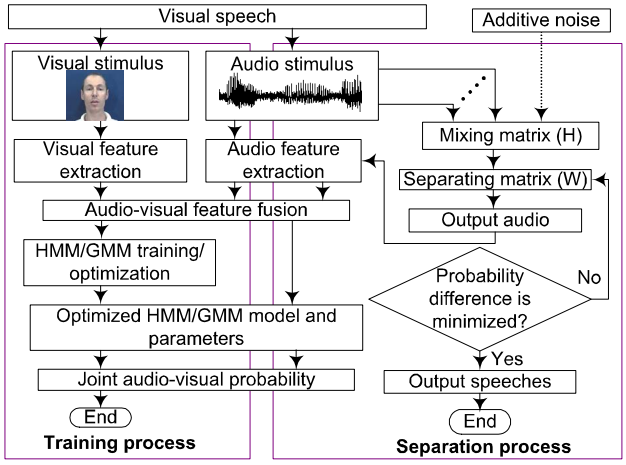

Fig. 1. Audio-visual speech separation system

for the case of having no permutation ambiguity. Ideally, $y_i = s_i$. For this case, y_i will have the same coherence as s_i with the associated visual sequence v_i, i.e. the maximum coherence between the audio and visual modalities. Therefore, maximizing the coherence $\mathcal{C}$ between y_i and v_i provides a criterion for separating the mixtures,

$$\mathcal{J}(\mathbf{W}) = \arg\max_{\mathbf{W}} \sum_{i=1}^{N} \mathcal{C}(y_i, v_i). \quad (3)$$

One crucial problem remaining to answer is to find out a suitable statistical model characterizing such coherence. From the speech production point of view, sounds are produced from the invisible vibration of the vocal cords and soft palate along with visible moments of lips, teeth and tongues. Therefore, we can consider the spectral information for the audio modality and mouth feature for visual modality. To represent their correlation, we use the feature extraction and statistical modeling techniques as elaborated hereafter.

2.3. Feature Extraction and Fusion

To extract the audio feature, we resort to the spectral information by applying the well-established filter bank analysis approach to short time-windowed segments of audio speech, using mel-scaled filters. Mel-frequency cepstral coefficients (MFCCs) $\mathbf{a}_{MFCC}$ are thereby computed by taking discrete cosine transform (DCT) of the log of the mel-scale filter-bank magnitudes. This approach is suggested due its ability of mimicing human ear's non-linear frequency resolution and its robustness in the presence of source degradation. Applying PCA to $\mathbf{a}_{MFCC}$, we obtain $\mathbf{a}_s = [a_{s1}, \ldots, a_{sq}]^T \in \mathbb{R}^q$, where subscript s denotes the *source*.

There are a variety of techniques for extracting visual features from a video signal. In this study, we use the active appearance model (AAM) proposed in [7]. The dimension of the acquired feature is further reduced with the PCA to generate the final visual feature, denoted as $\mathbf{v}_s = [v_{s1}, \ldots, v_{sp}]^T \in \mathbb{R}^p$. Note that, $\mathbf{a}_s$ should be calculated synchronously with $\mathbf{v}_s$ in implementation. With these time-synchronous and dimension-reduced feature vectors $\mathbf{a}_s$ and $\mathbf{v}_s$ on hand, we can generate a joint audio-visual observation by concatenation, i.e. $\mathbf{u}_s = [\mathbf{v}_s^T \ \mathbf{a}_s^T]^T \in \mathbb{R}^{p+q}$. This implies that the feature fusion is conducted by mapping on the feature level between the parameters.

2.4. Statistical Modeling and Training

Both the widely used hidden Markov model (HMM) and the Gaussian mixture model (GMM) can be applied to model the probability distribution of $\mathbf{u}_s$. The probability density of $\mathbf{u}_s$ is modeled by Gaussian mixture density, given by

$$p(\mathbf{u}_s) = \sum_{i=1}^{K} \boldsymbol{\sigma}_i \frac{\exp\{-\frac{1}{2}(\mathbf{u}_s - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1}(\mathbf{u}_s - \boldsymbol{\mu}_i)\}}{\sqrt{(2\pi)^K |\boldsymbol{\Sigma}_i|}}, \quad (4)$$

where $\boldsymbol{\mu}_i$, $\boldsymbol{\Sigma}_i$, $\boldsymbol{\sigma}_i$ and K are the mean vector, covariance matrix, weights and the number of Gaussian kernels respectively. Therefore, the parameter space for GMM is denoted as $\boldsymbol{\lambda}_{GMM} = (\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \sigma_i)$, $i = 1, \ldots, K$. The expectation-maximization (EM) algorithm [8] is used to obtain *maximum likelihood* estimates of $\boldsymbol{\lambda}$. This optimal $\boldsymbol{\lambda}$ is used for computing the joint audio-visual probability of a training video, and will also be used in the separation algorithm.

Based on the above analysis, the training process we used for acquiring the joint probability is conducted as following steps:

Step 1: Extract the visual feature $\mathbf{v}_s$ from video sequence by applying the AAM approach followed by PCA;

Step 2: Extract the audio feature $\mathbf{a}_s$ from audio sequence, i.e. MFCC followed by PCA, and form the audio-visual feature $\mathbf{u}_s = [\mathbf{v}_s^T \ \mathbf{a}_s^T]^T$;

Step 3: Train the HMM/GMM model based on a training corporus of $\mathbf{u}_s$ by using EM algorithm to optimize $\boldsymbol{\lambda}$;

Step 4: Calculate the joint probability $p(\mathbf{u}_s)$ using (4).

2.5. Audio-Visual Source Separation

After obtaining the joint statistical model as described above, $\mathcal{C}(y_i, v_i)$ in (3) takes the following expression,

$$\mathcal{J}(\mathbf{W}) = \arg\max_{\mathbf{W}} \sum_{i=1}^{N} p(\mathbf{u}_{y_i}), \quad (5)$$

where $\mathbf{u}_{y_i} = [\mathbf{v}_{s_i}^T \ \mathbf{a}_{y_i}^T]^T$ denotes the concatenated audio-visual feature of the i-th separated speech signal. Note that, visual feature $\mathbf{v}_s^T$ extracted from the source video from the training process is used to compose the new joint audio-visual observation. According to the analysis in section (2.2), it is clear that maximization of (5) leads to the optimal separation matrix $\mathbf{W}_{opt}$. Recalling the training process, we summarize our audio-visual speech source separation algorithm for instantaneous mixtures in the following steps:

Step 1: Estimate source signals $\mathbf{y} = \mathbf{W}\mathbf{x}$ from a $\mathbf{W}$ and calculate the audio feature $\mathbf{a}_y = [a_{y1}, \ldots, a_{yq}]^T$ from $\mathbf{y}$;

Step 2: Concatenate $\mathbf{a}_y$ with $\mathbf{v}_s$ to form a new joint audio-visual feature $\mathbf{u}_y = [\mathbf{v}_s^T \ \mathbf{a}_y^T]^T$;

Step 3: Calculate the joint probability $p(\mathbf{u}_y)$ using the GMM model, whose parameters $\boldsymbol{\lambda} = (\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \sigma_i)$ are obtained

from the training process; if $\mathcal{J}(\mathbf{W})$ in (5) is a maxima then stop, otherwise go back to *step 1* and repeat until it is maximized.

To see how the algorithm works, let us examine a two-input-two-output system. Substituting (1) into (2) in entry-form, we have $y_i = \sum_{j=1}^{2}(w_{i1}h_{1j} + w_{i2}h_{2j})s_j$, $i = 1, 2$. After separation, y_i should ideally correspond to s_j for $i = j$, with $w_{i1}h_{1j} + w_{i2}h_{2j} = 0$ for $i \neq j$. This leads to $w_{i2}/w_{i1} = -h_{1j}/h_{2j}$ for $i \neq j$, and w_{i2} can be derived by fixing w_{i1}. In implementation, the maximization of $p(\mathbf{u}_{y_i})$ in (5) gives the optimal w_{i2} after certain trials. In a strict sense, it is not a "blind" algorithm and does not assume the independence of sources. A potential way of exploiting both *independence* and *coherence* is to apply a penalty function based framework [6], detailed for the following convolutive case.

2.6. Audio-visual Constraint on Convolutive Mixtures

In a realistic environment, the output of the j-th microphone is modeled as a weighted sum of convolutions of the source signals, i.e., $x_j(n) = \sum_{i=1}^{N}\sum_{p=0}^{P-1} h_{jip}s_i(n-p)+e_j(n)$,where h_{jip} is the P-point impulse response from source i to microphone j. In this model, a large number of coefficients have to be estimated in the time domain. Moreover, there is no direct procedure available for introducing $p(\mathbf{u}_{y_i})$ to implement the algorithm described in section 2.5, due to the time delays being involved in the unmixing filters. For this case, we use a computationally efficient implementation in the frequency domain. Specifically, we use the following criterion,

$$\mathcal{J}(\mathbf{W}(\omega)) = \arg\min_{\mathbf{W}} \sum_{\omega=1}^{T}\sum_{l=1}^{L} \mathcal{F}(\mathbf{W})(\omega, l) \quad (6)$$

where $\mathcal{F}(\mathbf{W}) = \|\mathbf{R}_Y(\omega,l) - diag[\mathbf{R}_Y(\omega,l)]\|_F^2$, $diag(\cdot)$ is an operator which zeros the off-diagonal elements of a matrix, and $||\cdot||_F^2$ is the squared Frobenius norm, L is the number of time-blocks, $\mathbf{R}_Y(\omega,l)$ is the cross-power spectrum matrices of separated signals $\mathbf{y}$. We address the joint audio-visual model as a constraint, which allows to consider a trade-off between the *independence* and *coherence*, and employ a penalty function framework to incoporate the constraint [6], that is $\mathcal{U}(\mathbf{W}) = \left|1/\sum_{i=1}^{N} p(\mathbf{u}_{y_i})\right|^2$. For the sake of computational simplicity, this penalty function is calculated in the time domain from the *posterior* $\mathbf{y}$. The main steps of the algorithm are summarized:

Step 1: Compute a $\mathbf{W}(\omega)$ from the frequency domain BSS algorithm [6], and compute the time domain $\mathbf{W}$ by applying IFFT and estimate $\mathbf{y}$;

Step 2: Estimate $p(\mathbf{u}_{y_i})$ by following the same procedure as in section 2.5;

Step 3: Adjust the gradient $\nabla\mathcal{J}(\mathbf{W}(\omega))$ with the penalty $-1/(\sum_{i=1}^{N} p(\mathbf{u}_{y_i}))^4$, go back to *step 1* and repeat until $\mathcal{J}(\mathbf{W})$ is minimized.

3. EXPERIMENTAL RESULTS

The statistical model $p(\mathbf{u}_s)$ was trained based on two audio-visual sequences, uttered by two subjects. One of the subjects was recorded reciting part of a children's fairy-tale, while the other subject was recorded separately uttering a series of single words. Video data was obtained using a standard Digital Video (DV) camera at 25 fps, while audio was sampled at 32KHz, 16bit mono. The sequences were captured in an office environment with a low level of acoustic noise and front-on artificial lighting. Then, the mouth regions in the video were tracked using an AAM [7]. This provided a set of robust visual features, which encode both shape and texture information in a compact form.

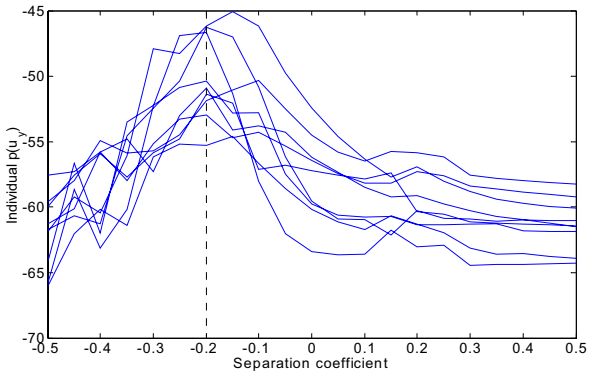

Fig. 2. The probability distribution of the separated signal y_1, i.e. $\log p(\mathbf{u}_{y_1})$, changes with separation coefficients using only one frame.

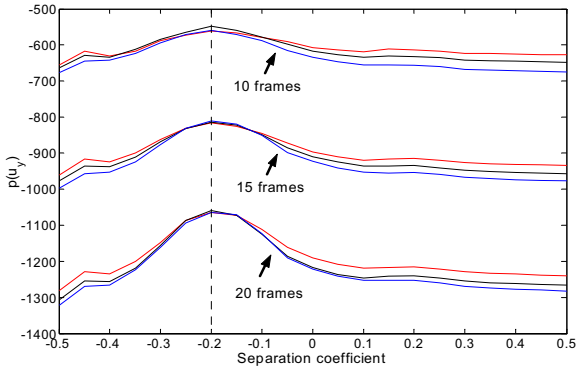

Fig. 3. The probability distribution of the separated signal y_1, i.e. $\log p(\mathbf{u}_{y_1})$, varies with different separation coefficients using multiple audio-visual frames. (Note that, from blue line to red line, there are 2 frames shifts between them.)

Both speech signals were processed using Mel-Cepstral analysis with 20ms Hamming windows, yielding 12 MFCCs per frame. Cepstral Mean Normalisation (CMN) was then performed on each set of coefficients to reduce any ambient background noise effects. A PCA model was constructed for each speaker using their respective MFCCs, which were projected through the models to obtain $\mathbf{a}_s$ with 12 dimensions (100% of the total energy). Finally, in order to retain

SNR(dB)		-6	0	6	12	clean
I	AV	6.2	7.5	9.5	12.6	14.1
I	A	1.3	4.1	7.3	10.1	12.2
C	AV	4.5	5.2	8.3	10.2	11.2
C	A	2.4	4.7	7.8	9.9	11.1

Table 1. SIR comparison between audio-visual (AV) and audio (A) only source separation for both instantaneous (I) and convolutive (C) mixtures.

one-to-one correspondences between the audio (20ms) and video (40ms) signals, linear interpolation was then applied to the visual appearance parameters to obtain $\mathbf{v}_s$ with 10 dimensions (82.2% of the total energy). Therefore, the dimension of the audio-visual space was 22 (10 video + 12 audio). These dimensions remain unchanged in the separation process. Both the training and testing data set contained 1411 audio-visual feature vectors. The number of Gaussian kernels modeling the training data set was 10.

For simplicity, only 2×2 systems were considered. The instantaneous mixtures were obtained by mixing two audio signals extracted from the training process with matrix [3, 0.8; 2, 4]. The separation algorithm described in section 2.5 was applied. Fig. 2 shows the joint audio-visual probability of y_1 (9 individual frames), which gives the optimal solution of w_{12} around -0.2 (assuming $w_{11} = 1$). However, the optimal probability was not accurate for some individual audio-visual frames. Hence, multiple frames were examined in Fig. 3, which clearly show much more robust solution to w_{12}. To simulate noisy acoustic environment, we added white noise to the mixed signals at different signal to noise ratio (SNR), see the result in Table 1. The GMM model was used for the experiments. For robust estimation, 20 audio-visual frames were used in this experiment. The signal to interference ratio (SIR) was used for performance evaluation, $10\log(|\mathbb{M}_{ii}|^2 \left\langle |s_i|^2 \right\rangle / \sum_{i\neq j} |\mathbb{M}_{ij}|^2 \left\langle |s_j|^2 \right\rangle)$, in which s_i and s_j are the i-th and j-th source signals, $\mathbb{M}_{ii}$ and $\mathbb{M}_{ij}$ are respectively the direct and cross channels of a multi-path channel $\mathbb{M}$.

For convolutive mixtures, we use the following SIR for performance evaluation [6],

$$SIR_i(\omega) = 10\log \frac{|\mathbb{M}_{ii}(\omega)|^2 \left\langle |s_i(\omega)|^2 \right\rangle}{\sum_{i\neq j} |\mathbb{M}_{ij}(\omega)|^2 \left\langle |s_j(\omega)|^2 \right\rangle}, \quad (7)$$

where $SIR_i(\omega)$ represents the SIR improvement at the i-th channel, and $s(\omega)$ are source signals in the frequency domain. The overall $SIR(\omega)$ is $(1/N)\sum_{i=1}^{N} SIR_i(\omega)$. The algorithm in section 2.6 was applied to an artificially convolutive system with 9 taps. Fig. 4 shows the result of the audio-visual constraint for the permutation problem, where we saw considerable SIR improvement along the frequency axis. However, this gained performance will not be considered in noisy mixtures (see Table 1), where the permutation problem was addressed by using a hybrid approach

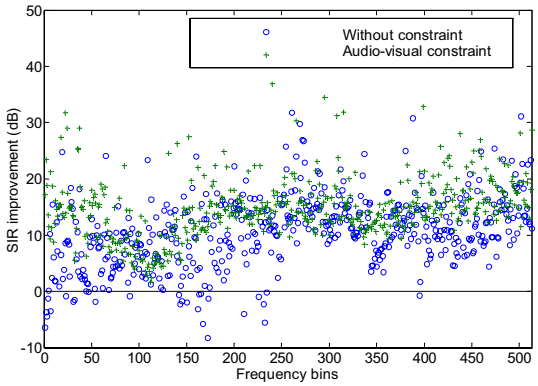

Fig. 4. Audio-visual constraint reduced the permutation effect along frequency axis.

proposed in [9]. Therefore, the SIR difference between AV and V for clean mixtures are not significant. However, the SIR improvement is considerably increased for the noisy mixtures.

4. CONCLUSION

The audio-visual source separation problems for both instantaneous and convolutive mixtures have been discussed. Experimental results indicate that using audio and visual modalities gives more precise separation than using audio only modality, especially for noisy mixtures. The bimodality of visual speech is also useful for addressing the frequency domain permutation problem as a result of the incorporated correlation between audio and visual features. However, there are several challenging problems demanding further research. For example, what is the best feature for representation of visual modality and on which level should the modalities be combined? How to effectively build robust model to represent the nonlinear dependency between audio-visual modalities?

5. REFERENCES

[1] L. Parra and C. Spence, "Convolutive blind source separation of non-stationary sources," *IEEE Trans. on SAP* , pp. 320–327, May 2000.

[2] D. G. Stork and M. E. Hennecke, Eds. *Speechreading by Humans and Machines: Models Systems and Applications*, Springer-Verlag, 1996.

[3] Q. Summerfield, "Lipreading and audiovisual speech perception," *Trans. R. Soc. Lond.*, pp. 71-78, 1992.

[4] D. Sodoyer, J.L. Schwartz, L. Girin, J. Klinkisch, and C. Jutten, "Separation of audio-visual speech sources," *EURASIP Journal on Applied Signal Processing*, pp. 1165-1173, Nov., 2002.

[5] S. Rajaram, A. V. Nefian, and T. S. Huang, "Bayesian separation of audio-visual speech separation," *Proc. ICASSP*, May, 2004.

[6] W. Wang, J. A. Chambers, and S. Sanei, "Penalty function based joint diagonalization approach for convolutive constrained BSS of nonstationary signals," *Proc. EUSIPCO*, Sept., 2004.

[7] T. F. Cootes and C. J. Taylor, "Statistical models of appearance for medical image analysis and computer vision," *Proc. SPIE Medical Imaging*, 2001.

[8] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Royal Statistical Society*, vol. 39, no. 1, pp. 1-38, 1977.

[9] W. Wang, J. A. Chambers, and S. Sanei, "A novel hybrid approach to the permutation problem of frequency domain blind source separation," *Proc. ICA*, Granada, Spain, Sept. 22-24, 2004.