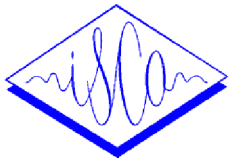

Auditory-Visual Speech Processing
2007 (AVSP2007)
Hilvarenbeek, The Netherlands
August 31 - September 3, 2007

Audiovisual Speaker Identity Verification Based on Cross Modal Fusion

Girija Chetty, Michael Wagner

School of Information Sciences and Engineering
University of Canberra, Australia

girija.chetty@canberra.edu.au , michael.wagner@canberra.edu.au

Abstract

In this paper, we propose the fusion of audio and explicit correlation features for speaker identity verification applications. Experiments performed with the GMM based speaker models with hybrid fusion technique involving late fusion of explicit cross-modal fusion features, with eigen lip and audio MFCC features allow a considerable improvement in EER performance An evaluation of the system performance with different gender specific datasets from controlled VidTIMIT data base and opportunistic UCBN database shows, that is possible to achieve an EER of less than 2% with correlated component hybrid fusion, and improvement of around 22 % over uncorrelated component fusion.

Index Terms: audiovisual, speaker identity verification, liveness checking, cross modal association

1. Introduction

The performance of a speaker verification system can be enhanced by including visual information from the lip region, as it would be more difficult for an impostor to imitate both audio and dynamical visual information simultaneously [1] [2] and [3]. Some of the recent findings in psychophysical analysis of visual speech by Kuratate, Munhall et.al [4], and Shinji Maeda [5] suggest that a speaking face is a kinematic-acoustic system in motion, and the shape, the texture, and the acoustic features during speech production are correlated in a complex way, with a single neuromotor source controlling the vocal tract behavior, and being responsible for both the acoustic and the visible attributes of speech production. Hence, the speaker models built with explicit audio-lip correlation features can allow better modeling of intrinsic temporal correlations between acoustic-labial articulators and vocal tract dynamics during speech production, enhancing the performance of speaker identity verification systems. Further, it would also allow liveness checks to be performed as it would be extremely difficult for an impostor to manufacture the complex intrinsic temporal correlations and make fraudulent replay attacks on speaker verification system [8]. The motivation to investigate explicit cross-modal features was based on the following two observations:

The first observation is in relation to any video event, for example a speaking face video, where the content usually consists of the synchronized audio and the visual elements. A series of psychological experiments on the cross-modality influence [5] have proved the importance of synergistic integration of the multiple modalities in the human perception system. A typical example of this kind is the well-known McGurk effect [5]. Cognitive psychologists in [4] and [5] suggest that the type of multi-sensory interactions occurring in the McGurk effect involves both the early and late stages of processing. It is likely that a human brain uses a hybrid form of fusion that depends on the availability and quality of different sensory cues. Yet, in audio visual speech and speaker verification systems, the analysis is usually performed separately on different modalities, first on the visual signal, then the audio signal and the results are brought together using different fusion methods. However, in this process of separation of modalities, we lose valuable information about the whole event and/or object we are trying to analyze and detect. As there is an inherent association between the two modalities, the analysis should take advantage of the synchronized appearance of the relationship between the audio and the visual signal.

The second observation relates to different types of fusion techniques used for joint processing of audio visual speech signals. The late-fusion strategy, which is also referred to as the decision or the score fusion, is effective especially in case the contributing modalities are uncorrelated and thus the resulting partial decisions are statistically independent. Feature level fusion techniques, on the other hand, can be favored only if a couple of modalities are highly correlated. However, coupled modalities such as audio and lip information during speech production can also consist of some components that are mutually independent. By extracting the explicit correlation features from the correlated audio and lip components, and then by employing an optimal combination of feature-level and late fusion techniques, the hybrid fusion of correlated and the mutually independent components can result in significant improvement in the performance of Speaker Identity Verification (SIV) systems. This is due to the ability of the speaker models built with hybrid fusion vectors in representing the intrinsic cross-modal association that exists between the closely coupled audio and lip modalities in speaking faces.

In this paper, we propose explicit correlation features based on two different cross modal association techniques namely cross modal factor analysis (CFA) to model the intrinsic temporal correlations. We first perform a cross correlation analysis on the audio and lip modalities to extract the correlated part of the information, and then employ a hybrid fusion approach based on the optimal combination of feature-level and late fusion techniques to fuse the correlated and the mutually independent components. The explicit correlation features proposed are based on the novel CFA technique, extract the multimodal content by identifying and measuring the intrinsic associations between different correlated modalities. The organization of this paper is as follows: In the next section, we present the details of proposed cross modal factor analysis (CFA) technique for extracting explicit correlation features. Section 3 describes the hybrid fusion scheme based on the late fusion of the explicit correlation features with audio and lip features for modeling the

correlated and uncorrelated components. The details of the experimental setup are described in Section 4. The performance evaluation for different datasets is discussed in Section 5, and the Section 6 summarizes the conclusions on the proposed cross modal analysis technique and the hybrid fusion technique.

2. Explicit Cross Modal Features

In this section we describe the details of extracting explicit correlation features based on cross modal factor analysis (CFA), a cross modal association technique, which allows the modeling of the correlated components in audio and lip modalities during speaking act. In CFA space, the problem can be defined as finding the optimal transformations under linear correlation model that best represents or identifies the coupled patterns between the features of the two different subsets. We adopt the following optimization criterion to obtain the optimal transformations:
Given two mean centered matrices X and Y, which compose of row-by-row coupled samples from two subsets of features, we want orthogonal transformation matrices A and B that can minimize the expression:

$$\|XA - YB\|_F^2 \tag{1}$$

Where $A^T A = I$ and $B^T B = I$

$\|M\|_F$ denotes the Frobenius norm of the matrix M and can be expressed as:

$$\|M\|_F = \left(\sum_i \sum_j |m_{ij}|^2 \right)^{1/2} \tag{2}$$

In other words, A and B define two orthogonal transformation spaces where coupled data in X and Y can be projected as close to each other as possible.
Since we have:

$$\begin{aligned} \|XA - YB\|_F^2 &= trace\left((XA - YB).(XA - YB)^T\right) \\ &= trace\left(XAA^T X^T + YBB^T Y^T - XAB^T Y^T - YBA^T X^T\right) \\ &= trace\left(XX^T\right) + trace(YY^T) - 2.trace(XAB^T Y^T) \end{aligned} \tag{3}$$

where trace of a matrix is defined to be the sum of the diagonal elements. We can easily see from above that matrices A and B which maximize trace (XAB^TY^T) will minimize Eqn. 3. It can be shown that such matrices are given by:

$$\begin{cases} A = S_{xy} \\ B = D_{xy} \end{cases}$$

where

$$X^T Y = S_{xy}.V_{xy}.D_{xy} \tag{4}$$

With the optimal transformation matrices A and B, we can calculate the transformed version of X and Y as follows:

$$\begin{cases} \tilde{X} = X.A \\ \tilde{Y} = Y.B \end{cases} \tag{5}$$

Corresponding vectors in $\tilde{X}$ and $\tilde{Y}$ are thus optimized to represent the coupled relationships between the two feature subsets without being affected by distribution patterns within each subset. Traditional Pearson correlation or mutual information calculation can then be performed on the first *k* corresponding vectors in $\tilde{X}$ and $\tilde{Y}$. In addition to feature dimension reduction, feature selection capability is another feature of CFA. The weights in A and B automatically reflect the significance of individual features. We show in Fig. 1 the absolute values of the first seven vectors of matrix A obtained from the training of 300 frames of visual faces and their associated speech features. For better visualization each vector has been reshaped according to the corresponding visual location. It is obvious that matrix A is able to highlight those facial areas corresponding most to the speech. This clearly demonstrates the great feature selection capability of CFA, which makes it a promising tool for different multimedia applications including audio-visual speaker identity verification.

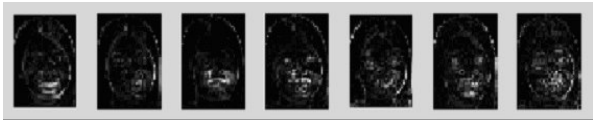

Fig. 1. Absolute values of the first seven vectors of matrix A reshaped according to the corresponding visual location

3. Hybrid Audio-Visual Fusion

In this section, we describe proposed hybrid fusion scheme for combining the audio-lip correlated CFA features extracted in Section 2 with mutually independent audio and lip region features. Using Eqn.5, an off-line training can be performed to calculate the optimized transformation matrices for the low-level audio and visual features. Video clips with speaking faces and synchronized speech from VidTIMIT and UCBN were used as the ground truth data for the training. The algorithm for audio-visual correlated component extraction is described now.

3.1. Feature Fusion of Correlated Components

Let f_A and f_L represent the audio MFCC and lip-region eigen lip features respectively. A and B represent the CFA transformation matrices. One can apply CFA to find two new feature sets $f_A^{'} = A^T f_A$ and $f_L^{'} = B^T f_L$ such that the between-class cross modal association coefficient matrix of $f_A^{'}$ and $f_L^{'}$ is diagonal with maximised diagonal terms. However, maximised diagonal terms do not necessarily mean that all the diagonal terms exhibit strong cross-modal association. Hence, one can pick the maximally correlated components that are above a certain correlation threshold θ. Let us denote the projection vector that corresponds to the diagonal terms larger than the threshold θ by $\tilde{w}_A$ and $\tilde{w}_L$. Then the corresponding projections of f_A and f_L are given as:

$$\tilde{f}_A = \tilde{w}_A^T . f_A \quad (6)$$

$$\tilde{f}_L = \tilde{w}_L^T . f_L \quad (7)$$

Here $\tilde{f}_A$ and $\tilde{f}_L$ are the correlated components that are embedded in f_A and f_L. By performing feature fusion of correlated audio and lip components, we obtain the CFA optimized feature fused audio-lip feature vector:

$$\tilde{f}_{AL} = \begin{bmatrix} \tilde{f}_A & \tilde{f}_L \end{bmatrix} \quad (8)$$

3.2. Late Fusion of Mutually Independent Components

In the Bayesian framework, late fusion can be performed using the product rule assuming statistically independent modalities. Various methods have been proposed in the literature [2] [3] and [6] as an alternative to the product rule such as the max rule, the min rule and the reliability-based weighted summation rule. We can compute joint or scores as a weighted summation:

$$\rho(\lambda_r) = \sum_{n=1}^{N} w_n \log P(f_n | \lambda_r) \; for \; r = 1,2,\ldots,R \quad (9)$$

Where $\rho_n(\lambda_r)$ is the logarithm of the class-conditional probability $P(f_n|\lambda_r)$ for the n^{th} modality, with class λ_r, and w_n denotes the weighting coefficient for modality n, such that $\Sigma_n w_n = 1$. Note that when $w_n = \frac{1}{N} \quad \forall n$, Eqn. 9 is equivalent to the product rule. Since the w_n values can be regarded as the reliability values of the classifiers, this combination method is also referred to as RWS (Reliability Weighted Summation) rule [4]. The statistical and the numerical range of these likelihood scores vary from one classifier to another. Thus using sigmoid and variance normalization as described in [4], the likelihood scores can be normalized to be within the (0, 1) interval before the fusion process.

The hybrid audio visual fusion vector is finally obtained by late fusion of feature fused correlated components ($\tilde{f}_{AL}$) with uncorrelated and mutually independent eigen lip features, and audio features with weights selected using RWS rule. An overview of the fusion method described is given in Figure 2.

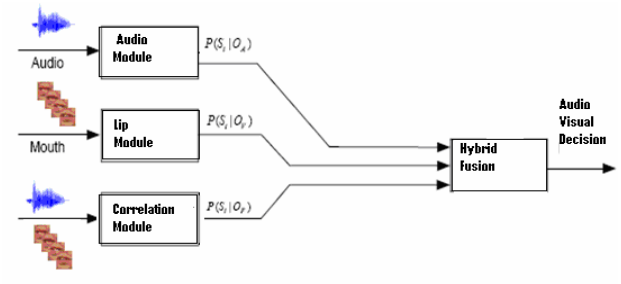

Fig. 2: System Overview of Hybrid Fusion Method

4. Experimental Setup

The audio visual data from two different data corpora, VidTIMIT and UCBN was used for evaluating the performance of explicit cross modal features and hybrid fusion approach. The VidTIMIT multimodal person authentication database [6], consists of video and corresponding audio recordings of 43 people (19 female and 24 male). The mean duration of each sentence is around 4 seconds, or approximately 100 video frames. A broadcast quality digital video camera in a noisy office environment was used to record the data. The video of each person is stored as a sequence of JPEG images with a resolution of 512×384 pixels with corresponding audio provided as a 16-bit 32-kHz mono PCM file.

The second type of data used is the UCBN database, a free to air broadcast news database. The broadcast news is a continuous source of video sequences, which can be easily obtained or recorded, and has optimal illumination, colour, and sound recording conditions. However, some of the attributes of broadcast news database such as near-frontal images, smaller facial regions, multiple faces and complex backgrounds present more realistic Speaker Identity Verification applications scenarios. The database consists of 20 - 40 second video clips for anchor persons and newsreaders with frontal/near-frontal shots of 10 different faces (5 female and 5 male). Each video sample is 25 frames per second MPEG2 encoded stream with a resolution of 720 × 576 pixels, with corresponding 16 bit, 48 kHz PCM audio. Figure 3 shows some sample images from the VidTIMIT database (first two rows) and UCBN database (last two rows).

Fig 3: Sample faces from VidTIMIT and UCBN databases

The impostor population was generated by leave one out scheme. The gender-specific universal background models (UBMs) were developed using the training data from two sessions, Session 1 and Session 2, of the VidTIMIT corpus, and for testing Session 3 was used. Due to the type of data available (test session sentences differ from training session sentences), only text independent speaker verification experiments could be performed with VidTIMIT. This gave

1536 (2*8*24*4) seconds of training data for the male UBM and 576(2*8*19*4) seconds of training data for the female UBM. The GMM topology with 10 Gaussian mixtures was used for all the experiments. The number of Gaussian mixtures were determined empirically to give the best performance. For the UCBN database, similar gender-specific universal background models (UBMs) were obtained using training data from the text independent subsets 3 & 4. Ten sessions of the male and female speaking face data from these subsets were used for training and 5 sessions for testing.

For all the experiments, the global threshold was set using test data. For the male only subset for the VidTIMIT database, there were 48 client trials (24 male speakers x 2 test utterances in Session 3) and 1104 impostor trials (24 male speakers x 2 test utterances in Session 3 x 23 impostors/client), and for the female VidTIMIT subset, there were 38 client trials (19 male speakers x 2 test utterances in Session 3) and 684 impostor trials. For the male only subset for UCBN database, there were 25 client trials (5 male speakers x 5 test utterances in each subset) and 100 impostor trials, and for the female UCBN subset, there were similar number of the client and impostor trials as in the male subset as we used 5 male and 5 female speakers from different subsets. The results for VidTIMIT male subset and UCBN female subset only are discussed in this paper.

5. Results and Discussion

Different sets of experiments were conducted to evaluate the performance of the explicit cross modal features and hybrid fusion features in terms of DET curves and equal error rates (EER). The EER performances in Table 2 and also the DET curves in Fig. 3, depict an evaluation of the hybrid fusion of correlated audio-lip features based on the cross modal factor analysis (CFA) with mutually independent audio-lip features for the SIV scenario.

As can be seen in Table 2 and DET curves in Fig. 3, the performance of ordinary features fusion of audio lip features , can be improved by cross modal analysis. For the feature fusion of the correlated components , the EER improves from 7.2 % to 4.7 % for CFA analysis for VidTIMIT male subset. Since each modality also carries mutually independent information, e.g. the texture of the lip region, possibly containing the information about the identity of a speaker, the overall performance can be enhanced with hybrid fusion, with an optimal combination of the feature-level and the late fusion techniques combining lip, audio and correlated audio-lip feature vectors.

As observed in Table 2 and DET curves in Figure 3, for the VidTIMIT male subset, the hybrid fusion involving late fusion of audio features with feature-level fusion of correlated audio–lip features based on CFA analysis + , yields a best EER of 0.68 %. Similar performance can be observed for different combinations of correlated component and independent component fusion for UCBN female dataset. For both data sets, around 22% improvement in EER is achieved with correlated component hybrid fusion (+ +) as compared to uncorrelated component hybrid fusion (+ +).It can also be noted that all the hybrid fusion modes (last four rows in Table 2) resulted in synergistic fusion, with the EER performance better than baseline audio only and visual only EERs of 4.88% and 6.2% for VidTIMIT male subset and 5.7 % and 7.64 % for the UCBN female subset.

Table 1: EERs at varying correlation coefficient threshold values (θ) with the corresponding projection dimension (dim) for CFA features

	EER(%) at (θ , dim)						
θ	0.0	0.1	0.2	0.3	0.4	0.5	0.6
dim	40	15	12	10	8	6	4
$\tilde{f}_{AL}$	6.8	4.7	5.3	5.0	4.7	7.4	10.3

Table 8.2: EER performance with late fusion of correlated ( ) components with mutually independent (&) components: (+ represents RWS rule for late fusion, - represents feature level fusion)

Modality	VidTIMIT (male) EER (%)	UCBN (female) EER (%)
f_{mfcc}	4.88	5. 7
f_{eigLip}	6.2	7.64
$f_{mfcc-eigLip}$	7.2	8.9
$\tilde{f}_{mfcc-eigLip}$	4.7	5.81
$f_{mfcc} + f_{mfcc-eigLip}$	0.97	1.01
$f_{mfcc} + \tilde{f}_{mfcc-eigLip}$	0.68	0.75
$f_{mfcc} + f_{eigLip} + f_{mfcc-eigLip}$	1.85	2.01
$f_{mfcc} + f_{eigLip} + \tilde{f}_{mfcc-eigLip}$	1.46	1.53

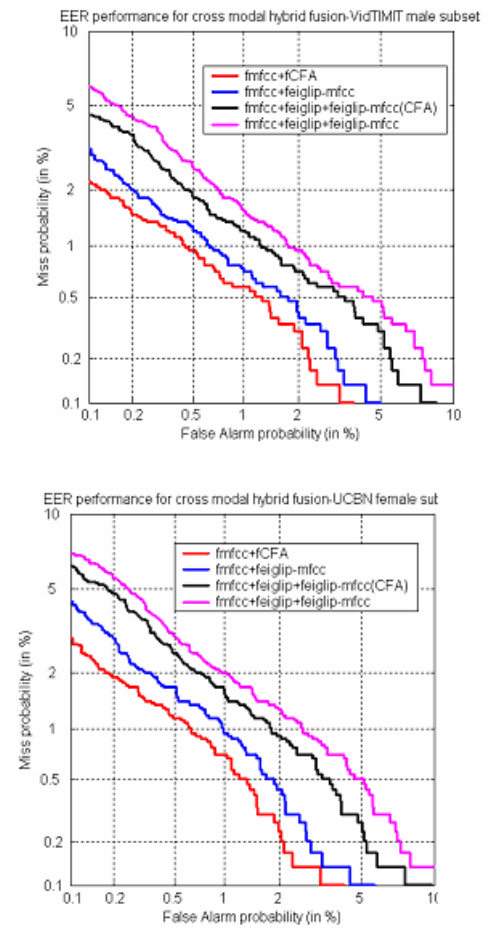

Fig. 4: DET curves for hybrid fusion of CFA correlated audio-lip features and mutually independent audio-lip features

6. Conclusions

In this paper, the performance evaluation of a novel hybrid fusion approach involving the correlated and the independent audio and lip modalities is proposed. The proposed cross modal factor analysis technique allows the extraction of the optimal correlated audio-lip features. An EER of less than 2 % was achieved for hybrid fusion with correlated features, and an EER improvement of around 22% is achieved with correlated component hybrid fusion when compared to uncorrelated component fusion. The EER performance for UCBN female subset for all fusion experiments was quite close to VidTIMIT male subset, even with poor quality of the visual data in UCBN dataset, with low resolution, small facial images, and presence of mostly irrelevant background information in the image sequences. Nevertheless, the performance for proposed technique with UCBN dataset from an opportunistic database depicts a more realistic speaker identity verification scenario.

7. References

[1] A. Ross and A. Jain, "Information fusion in biometrics," Pattern Recognition Letters, vol. 24, pp. 2115-2125, 2003/9 2003.

[2] R. Brunelli and D. Falavigna, "Person Identification Using Multiple Cues," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 17, pp. 955-966, Oct. 1995.

[3] C. C. Chibelushi, F. Deravi, and J. S. D. Mason, "A Review of Speech-Based Bimodal Recognition," IEEE Transactions on Multimedia, vol. 4, pp. 23-35, Mar 2002.

[4] Kuratate, T., Munhall, K.G., Rubin, P.E., Vatikiotis-Bateson, E., & Yehia, H. (1999). Audio-visual synthesis of talking faces from speech production correlates. Proceedings of EuroSpeech'99, ESCA.

[5] Shinji Maeda, " A face model derived from a guided PCA of motion capture data and McGurk effects" Proceedings of the ATR symposium on Cross-modal Processing of Faces and Voices, pp. 63-64, Jan. 2005.

[6] Sanderson, C. and K.K. Paliwal , "Fast features for face authentication under illumination direction changes", Pattern Recognition Letters 24, 2409-2419, 2003.

[7] Wojtek Krzanowski, Principles of multivariate analysis, Oxford University Press, Oxford, 1988.

[8] Chetty, G. and Wagner, M., "Liveness Detection Using Cross Modal Correlations in Face-Voice Person Authentication", in Proc. INTERSPEECH 2005, 9th European Conference on Speech Communication and Technology, 4th -7th October 2005, Lisbon, Portugal,. pp. 2181-2184.