Audio Visual Speaker Verification Based on Hybrid Fusion of Cross Modal Features

Girija Chetty and 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 hybrid fusion of audio and explicit correlation features for speaker identity verification applications. Experiments were performed with the GMM based speaker models with a hybrid fusion technique involving late fusion of explicit cross-modal fusion features, with implicit eigen lip and audio MFCC features. An evaluation of the system performance with different gender specific datasets from controlled VidTIMIT data base and opportunistic UCBN database shows a significant performance improvement.

Keywords: Audio-visual, speaker identity verification, liveness checking, cross modal correlations.

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] and [4]. Some of the recent findings in psychophysical analysis of visual speech by Kuratate, Munhall et.al [2], and Shinji Maeda [3] 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.

In this paper, we propose several such explicit correlation features to model the intrinsic temporal correlations that exist in visual speech. We first perform a cross correlation analysis on the audio and lip modalities to extract the correlated part of the information, and then employ an hybrid fusion approach based on the optimal combination of feature-level and late fusion techniques to fuse the correlated and the

A. Ghosh, R.K. De, and S.K. Pal (Eds.): PReMI 2007, LNCS 4815, pp. 469 – 478, 2007.

mutually independent components. Further, an automatic weight selection technique which automatically adapts the fusion weights to the audio noise conditions is proposed. We propose three different types of cross-modal association approaches under the linear correlation model: the Latent Semantic Analysis (LSA), the Cross-modal Factor Analysis (CFA), and the Canonical Correlation Analysis (CCA).

Experiments performed with the hybrid fusion based on the optimal combination of feature-level fused explicit cross-modal features (LSA, CFA and CCA), and the late fusion of lip features (eigenlip) features and audio MFCC features allow a considerable improvement in EER performance for both speaker identity verification scenarios. To improve the EER performance for noisy audio SIV scenarios, the late fusion weights were determined automatically, by performing a mapping between an audio reliability estimate and the modality weightings. It was found that the hybrid fusion with automatic weight adaptation improves the EER performance for different SIV scenarios including both the clean and the noisy audio conditions. This paper is organised as follows.

In the next section, the proposed explicit cross modal features (LSA, CFA and CCA), are described. Section 3 describes scheme for hybrid fusion of the explicit correlation features with audio and lip features for modeling the correlated and uncorrelated components. The automatic weight adaptation scheme is described in the Section 4. The details of the experimental results for the proposed approach are described in Section 5. Section 6 summarizes the findings on cross-modal features and describes details of the next stage of investigations.

2 Cross-Modal Association

In this section we describe the details of extracting explicit correlation features based on cross modal association (CMA) techniques which allow the modelling of the correlated components in audio and lip modalities.

2.1 Latent Semantic Analysis

Latent semantic analysis (LSA) is used as a powerful tool in text information retrieval to discover underlying semantic relationship between different textual units (.e.g. keywords and paragraphs) [5]. It is possible to detect the semantic correlation between visual faces and its associated speech based on LSA technique. The method consists of three major steps: the construction of a joint multimodal feature space, the normalization, the singular value decomposition (SVD), and the semantic association measurement.

Given n visual features and m audio features at each of the t video frames, the joint feature space can be expressed as:

$$X = [V_1, \ldots, V_i, \ldots, V_n, A_1, \ldots, A_i, \ldots A_m], \quad \text{where} \tag{1}$$

$$V_i = (v_i(1), v_i(2), \ldots, v_i(t))^T, \quad \text{and} \tag{2}$$

$$A_i = (a_i(1), a_i(2), \ldots, a_i(t))^T \tag{3}$$

Various visual and audio features can have quite different variations. Normalization of each feature in the joint space according to its maximum elements (or certain other statistical measurements) is thus needed and can be expressed as:

$$\hat{X}_{ij} = \frac{X_{ij}}{\max(\ abs\ (X_{ij}))} \quad \forall j \tag{4}$$

After normalization all elements in normalized matrix $\hat{X}$ have values between –1 and 1. SVD can then be performed as follows:

$$\hat{X} = S \,.\, V \,.\, D^{T} \tag{5}$$

where S and D are matrices composing of left and right singular vectors and V is diagonal matrix of singular values in descending order. Keeping only the first and most important *k* singular vectors in *S* and *D*, we can derive an optimal approximation of $\hat{X}$ with reduced feature dimensions, where semantic (correlation) information between visual and audio features is mostly preserved.

2.2 Cross-Modal Factor Analysis

LSA does not distinguish features from different modalities in the joint space. The optimal solution based on overall distribution which LSA models, may not best represent semantic relationships between features of different modalities, since distribution patterns among features from the same modality will also greatly impact LSA's results. A solution to the above problem is to treat the features from different modalities as two separate subsets and focus only on the semantic patterns between these two subsets. Under the linear correlation model, the problem now is to find the optimal transformations that can best represent (or identify) 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:

$$\| XA - YB \|_F^2 \text{ , where} \tag{6}$$

$$A^T A = I \text{ and } B^T B = I$$

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

$$\| M \|_F = \left(\sum_i \sum_j |m_{ij}|^2 \right)^{1/2} \tag{7}$$

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:

$$\begin{aligned} \|XA - YB\|_F^2 &= trace\left((XA - YB).(XA - YB)^T\right) \\ &= trace\left(XAA^T X^T + YBB^T Y^T - XAB^T Y^T - YBA^T X^T\right) \\ &= trace\left(XX^T\right) + trace\left(YY^T\right) - 2.trace\left(XAB^T Y^T\right) \end{aligned} \tag{8}$$

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:

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

$$X^T Y = S_{xy}.V_{xy}.D_{xy} \tag{9}$$

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

$$\begin{cases} \tilde{X} = X.A \\ \tilde{Y} = Y.B \end{cases} \tag{10}$$

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.

2.3 Canonical Correlation Analysis

Following the development of the previous section, we can adopt a different optimization criterion: Instead of minimizing the projected distance, we attempt to find transformation matrices A and B that maximize the correlation between X.A and Y.B. Given two mean centered matrices X and Y as defined in the previous section, we seek matrices A and B such that

$$correlation(XA, XB) = correlation(\tilde{X}, \tilde{Y}) = diag(\sigma_1, \cdots \sigma_i, \cdots, \sigma_l) \tag{11}$$

where $\tilde{X} = Y.B$, and $1 \geq \sigma_1 \geq, \cdots, \sigma_i, \cdots, \geq \sigma_l \geq 0$. σ_i represents the largest possible correlation between the i^{th} translated features in $\tilde{X}$ and $\tilde{Y}$. The CCA analysis is described in further detail in [6].

3 Hybrid Audio-Visual Fusion

In this section, we describe proposed hybrid fusion scheme for combining the audio-lip cross-correlation features extracted in Section 2, with the mutually independent audio and lip region features. 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 eigenlip features respectively. A and B represent the cross modal transformation matrices. One can apply LSA, CCA or 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:

$$\tilde{f}_A = \tilde{w}_A^T \cdot f_A \tag{12}$$

$$\tilde{f}_A = \tilde{w}_L^T \cdot f_A \tag{13}$$

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:

$$\tilde{f}_{AL} = \begin{bmatrix} \tilde{f}_A & \tilde{f}_L \end{bmatrix} \tag{14}$$

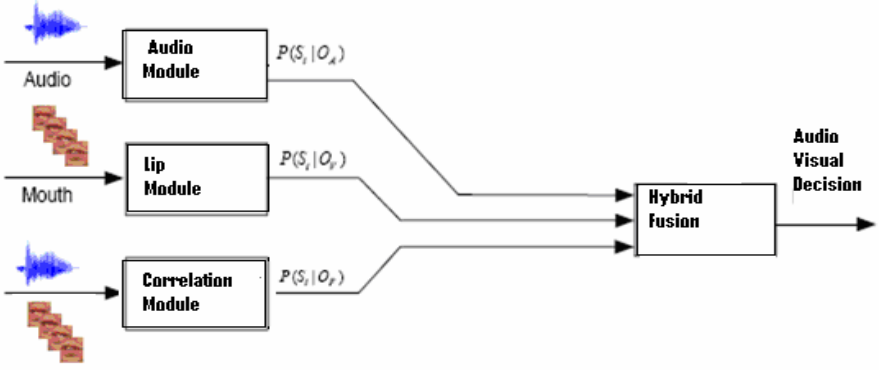

Fig. 1. System Overview of Hybrid Fusion Method

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:

$$\rho(\lambda_r) = \sum_{n=1}^{N} w_n \log P(f_n|\lambda_r) \; for \quad r = 1,2,\ldots,R \tag{15}$$

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. 15 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 eigenlip features, and audio features with weights selected using RWS rule. An overview of the fusion method described is given in Figure 1.

4 Automatic Weight Adaptation

For the RWS rule, the fusion weights are chosen empirically, whereas for the automatic weight adaptation, a mapping needs to be developed between an audio reliability estimate and the modality weightings. The late fusion scores can be fused via addition or multiplication as shown in Eqn. 16 and 17. Both methods were investigated and it was found that the results achieved for both were similar (based on empirically chosen weights). However, additive fusion has been shown to be more robust to classifier errors [4], and should perform better when the fusion weights are automatically, rather than empirically determined. Hence the results for additive fusion only, are presented in this chapter. Prior to late fusion, all scores were normalized to fall into the range of [0, 1], using *min-max* normalisation.

$$\begin{aligned} P(S_i|x_A,x_Y) &= \alpha P(S_i|x_A) + \beta P(S_i|x_Y), \quad \text{(a)} \\ P(S_i|x_A,x_Y) &= (P(S_i|x_A))^{\alpha} \times (P(S_i|x_Y))^{\beta}, \quad \text{(b)} \end{aligned} \tag{16}$$

where

$$\alpha = \begin{cases} 0, & c \le -1, \\ 1+c, & -1 < c < 0, \\ 1, & c \ge 0, \end{cases}$$
$$\beta = \begin{cases} 1, & c \le 0, \\ 1-c, & 0 < c < 1, \\ 0, & c \ge 1, \end{cases} \tag{17}$$

where x_A and x_V refer to the audio test utterance and visual test sequence/image respectively.

To carry out automatic fusion, that adapts to varying acoustic SNR conditions, a single parameter c, the *fusion parameter*, is used to define the weightings; the audio weight α and the visual weight β, i.e., both α and β dependent on c. Fig. 2 and Eqn. 17 show how the fusion weights, α and β, depend on the fusion parameter c. Higher values of c (>0) place more emphasis on the audio module whereas lower values (<0) place more emphasis on the visual module. For $c \geq 1$, $\alpha = 1$ and $\beta = 0$, hence the audio-visual fused decision is based entirely on the audio likelihood score, whereas, for $c \leq -1$, $\alpha = 0$ and $\beta = 1$, the decision is based entirely on the visual score. So in order to account for varying acoustic conditions, only c has to be adapted.

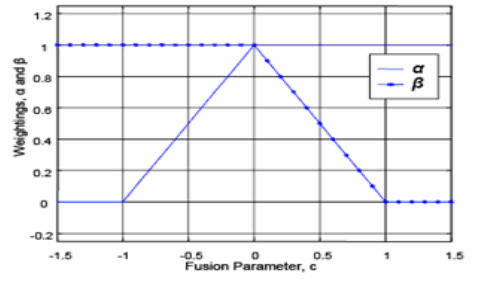

Fig. 2. The module weightings versus the fusion parameter "*c*"

The reliability measure was the audio likelihood score $\rho_n(\lambda_r)$. As the audio SNR decreases, this reliability measure decreases in absolute, and becomes closer to threshold for client likelihoods. Under clean test conditions, this reliability measure increases in absolute value because the client model yields a more distinct score. So, a mapping between ρ and c can automatically vary α and β and hence place more/less emphasis on the audio scores. To determine the mapping function c(ρ), the values of c which provided for optimum fusion, copt, were found by exhaustive search for the N tests at each SNR levels. This was done by varying c from –1 to +1, in steps of 0.01, in order to find out which c value yielded the best performance. The corresponding average reliability measures were calculated, ρmean, across the N test utterances at each SNR level.

$$c(\rho) = c_{oz} + \frac{h}{1 + \exp[d.(\rho + \rho_{oz})]}, \tag{18}$$

The described method can be employed to combine any two modules. It can also be adapted to include a third module. We assume here that only the audio signal is degraded when testing, and that the video signal is of fixed quality. The third module we use here is an audio-lip correlation module, which involves a cross modal

transformation of feature fused audio-lip features based on CCA, CFA or LSA cross modal analysis described in Sections 2 and 3.

5 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 [4], 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.

Fig. 3. VidTIMIT and UCBN databases

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. 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). Figure 3 shows some sample images from the VidTIMIT database (first two rows) and UCBN database (last two rows).

6 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). As can be seen in % EER table in Table 1, the performance of ordinary features fusion of audio lip features $f_{mfcc-eigLip}$, can be improved by cross modal analysis. For the feature fusion of the correlated components $\tilde{f}_{mfcc-eigLip}$, 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

Table 1. EER (%) performance with late fusion of correlated ($\tilde{f}_{mfcc-eigLip}$) components with mutually independent (f_{eigLip} & f_{mfcc}) components: (+ represents RWS rule for late fusion, - represents feature level fusion)

Dataset	**VidTIMIT male subset**			**UCBN female subset**		
Cross Modal Features	CFA EER	CCA EER	LSA EER	CFA EER	CCA EER	LSA EER
f_{mfcc}	4.88	4.88	4.88	5. 7	5. 7	5. 7
f_{eigLip}	6.2	6.2	6.2	7.64	7.64	7.64
$f_{mfcc-eigLip}$	7.2	7.87	12.47	8.9	9.54	17.36
$\tilde{f}_{mfcc-eigLip}$	4.7	5.18	8.09	5.81	6.28	9.74
$f_{mfcc} + f_{mfcc-eigLip}$	1.03	1.03	1.03	1.12	1.12	1.12
$f_{mfcc} + \tilde{f}_{mfcc-eigLip}$	0.68	0.86	1.29	0.79	1.17	1.34
$f_{mfcc} + f_{eigLip} + f_{mfcc-eigLip}$	1.26	1.26	1.26	1.46	1.46	1.46
$f_{mfcc} + f_{eigLip} + \tilde{f}_{mfcc-eigLip}$	1.06	1.85	2.22	1.23	2.31	2.46

combination of the feature-level and the late fusion techniques combining lip, audio and correlated audio-lip feature vectors.

Also, 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 $f_{mfcc} + \tilde{f}_{mfcc-eigLip}$, 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 ($f_{mfcc} + f_{eigLip} + \tilde{f}_{mfcc-eigLip}$) as compared to uncorrelated component hybrid fusion ($f_{mfcc} + f_{eigLip} + f_{mfcc-eigLip}$).It can also be noted that all the hybrid fusion modes (last four rows in Table 1) 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.

7 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.

References

[1] Brunelli, R., Falavigna, D.: Person Identification Using Multiple Cues. IEEE Transactions on Pattern Analysis and Machine Intelligence 17, 955–966 (1995)

[2] Kuratate, T., Munhall, K.G., Rubin, P.E., Vatikiotis-Bateson, E., Yehia, H.: Audio-visual synthesis of talking faces from speech production correlates. In: Proc. EuroSpeech 1999, ESCA (1999)

[3] Maeda, S.: A face model derived from a guided PCA of motion capture data and McGurk effects. In: Proceedings of the ATR symposium on Cross-modal Processing of Faces and Voices, pp. 63–64 (January 2005)

[4] Sanderson, C., Paliwal, K.K.: Fast features for face authentication under illumination direction changes. Pattern Recognition Letters 24, 2409–2419 (2003)

[5] Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. Journal of the American Society for Information Science 41(6), 391–407 (1990)

[6] Borga, M., Knutsson, H.: Finding Efficient Nonlinear Visual Operators using Canonical Correlation Analysis. In: Proc. of SSAB 2000, Halmstad, pp. 13–16