MULTIMODAL SPEAKER IDENTIFICATION USING CANONICAL CORRELATION ANALYSIS

M. E. Sargın, E. Erzin, Y. Yemez and A. M. Tekalp

Multimedia, Vision and Graphics Laboratory
College of Engineering, Koç University, Sarıyer, Istanbul, 34450, Turkey
{msargin,eerzin,yyemez,mtekalp}@ku.edu.tr

ABSTRACT

In this work, we explore the use of canonical correlation analysis to improve the performance of multimodal recognition systems that involve multiple correlated modalities. More specifically, we consider the audiovisual speaker identification problem, where speech and lip texture (or intensity) modalities are fused in an open-set identification framework. Our motivation is based on the following observation. The late integration strategy, which is also referred to as decision or opinion fusion, is effective especially in case the contributing modalities are uncorrelated and thus the resulting partial decisions are statistically independent. Early integration 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 texture also consist of some components that are mutually independent. Thus we first perform a cross-correlation analysis on the audio and lip modalities so as to extract the correlated part of the information, and then employ an optimal combination of early and late integration techniques to fuse the extracted features. The results of the experiments testing the performance of the proposed system are also provided.

1. INTRODUCTION

Audio is probably the most natural modality to recognize speech content and a valuable source to identify a speaker [1]. Video also contains important biometric information, which includes face/lip texture and lip motion information that is correlated with the audio. Audio-only speaker/speech recognition systems are far from being perfect especially under noisy conditions. Performance problems are also observed in video-only speaker recognition systems, where poor picture quality, changes in pose and lighting conditions, and varying facial expressions may have detrimental effects [2]. Hence, robust solutions for both speaker and speech recognition should employ multiple modalities, such as audio, lip texture, and lip motion in a unified scheme.

The design of a multimodal recognition system requires addressing three basic issues: i) Which modalities to fuse, ii) How to represent each modality with a discriminative and low-dimensional set of features, and iii) How to fuse existing modalities. The second issue, representative feature selection, also includes modeling of classifiers through which each class is represented with a statistical model or a representative feature set. As for the last issue, that is the fusion problem, different strategies are possible: In the so-called "early integration", modalities are fused at data or feature level, whereas in "late integration" decisions or scores resulting from each unimodal recognition are combined to give the final conclusion.

Audio and lip information have been used for speaker identification in only few works such as [3, 4, 5]. These works mainly focus on fusion of decisions coming from different feature sets. In the speaker recognition literature, audio is generally modeled by mel frequency cepstral coefficients (MFCC). Unlike audio, several feature sets can be used to represent the lip modality such as shape, motion and texture. In texture-based approaches, pure or DCT-domain lip image intensities are used as features [6, 4]. The dimension reduction techniques, such as principle component analysis, linear discriminant analysis or DCT, are applied independently to lip or speech regardless of the mutual information between them. However one should also consider the correlation between speech and lip features since it is a widely accepted fact that these features are coupled and highly correlated with each other. Despite the fact that multimodal speech and speaker recognition systems implicitly use the audio-visual correlation, in the literature there is relatively little work for explicit audio-visual correlation analysis. In [7], the mutual information between a speech spectrogram and an intensity based image is examined to detect image segments where these two modalities are highly correlated. Our previous work [8] addresses the use of canonical correlation analysis to find the relationship between audio and visual modalities so as to apply it to synchronization of audio-visual features.

In this study, we explore the use of canonical correlation analysis to improve the performance of multimodal recognition systems that involve multiple modalities that are correlated with each other. More specifically, we consider the audio-visual speaker identification problem, where speech and lip texture (or intensity) modalities are fused in an open-set identification framework. Our motivation is based on the following observation. The late integration strategy, which is also referred to as decision or opinion fusion, is effective especially in case the contributing modalities are uncorrelated and thus the resulting partial decisions are statistically independent. Early integration 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 texture also consist of some components that are mutually independent. Thus we first perform a cross-correlation analysis on the audio and lip modalities so as to extract the correlated part of the information, and then employ an optimal combination of early and late integration techniques to fuse the extracted features.

In Section 2, we describe the probabilistic framework that we use for the open-set speaker identification problem. Section 3 presents the extraction of the initial audio and lip features. The proposed multimodal fusion scheme with canonical correlation analysis is presented in Section 4. Experimental results are discussed in Section 5

This work has been supported by TUBITAK under the project EEEAG-101E026 and by the European FP6 Network of Excellence SIMILAR (http://www.similar.cc).

and finally concluding remarks are given in Section 6.

2. OPEN-SET SPEAKER IDENTIFICATION

Speaker recognition task can be formulated as either verification or identification problem. The latter can further be classified as open-set or closed-set identification. In the closed-set identification problem, a reject scenario is not defined and an unknown observation is classified as belonging to one of the R registered pattern classes. In the open-set problem, the objective is, given the observation from an unknown pattern, to find whether it belongs to a pattern class registered in the database or not; the system identifies the pattern if there is a match and rejects otherwise. Hence, the problem can be thought of as an $R+1$ class identification problem, including also a reject class. Open-set identification has a variety of applications such as the authorized access control for computer and communication systems, where a registered user can log onto the system with her/his personalized profile and access rights. In this paper, we formulate the speaker recognition problem in an open-set identification framework, which is a more challenging and realistic way of addressing the problem as compared to closed-set speaker identification and verification. Note that verification is a special case of the general open-set identification problem.

In the open-set identification problem, an imposter class λ_{R+1} is introduced as the $R+1$'th class. Since it is difficult to accurately model the imposter class, λ_{R+1}, we employ the following solution which includes a reject strategy through the definition of the likelihood ratio:

$$\rho(\lambda_r) = \log \frac{P(\boldsymbol{f}|\lambda_r)}{P(\boldsymbol{f}|\lambda_{R+1})} \tag{1}$$

Then, the decision strategy of the open-set identification can be implemented in two steps. First, determine

$$\lambda_* = \underset{\lambda_1,\ldots,\lambda_R}{\arg\max}\, \rho(\lambda_r), \tag{2}$$

and then

$$\begin{array}{ll} \text{if } \ \rho(\lambda_*) \geq \tau & \text{accept} \\ \text{otherwise} & \text{reject} \end{array} \tag{3}$$

where τ is the optimal threshold which is usually determined experimentally to achieve the desired false accept or false reject rate.

Computation of class-conditional probabilities needs a prior modeling step, through which a probability density function of feature vectors is estimated for each class $r = 1, 2, \ldots, R$ by using available training data. A common and effective approach to model the impostor class is to use a universal background model, which is estimated by using all available training data regardless of which class they belong to.

3. MODALITIES AND FEATURES

In this paper, audio and lip texture are considered as different but partially correlated modalities. The mel-frequency cepstral coefficients are used as features for the audio modality. The audio feature vector $\boldsymbol{f}_A$ is formed as a collection of 13 MFCC coefficients along with the first and second derivatives, total of 39 coefficients. The features for the lip texture modality, $\boldsymbol{f}_L$, are 50 base-band zig-zag scanned 2-D DCT coefficients of the luminance component within a rectangular box about the lip region. Since audio features are extracted at a rate of 100 fps and lip texture features have a frame rate of 15 fps, visual features are interpolated to 100 fps rate using bi-cubic interpolation to synchronize the audio and visual features.

A preprocessing step is required to locate the lip region and eliminate the global motion of the head between the frames so that the extracted motion features within the lip region provides us with the pure movement of the speaking act. To this effect, each face frame is aligned with the first frame of the sequence using a 2D parametric motion estimator. For every two consecutive face images, global head motion parameters are calculated using hierarchical Gaussian image pyramids and 12-parameter quadratic motion model [9]. The face images are successively warped according to these calculated parameters [10]. In the resulting aligned image sequence, the location of the lip region remains almost unchanged except for local movements. Thus, by only hand-labeling the mid-point of the lip region on the first frame, we automatically extract a region of interest around this point so as to obtain a sequence of lip frames of size 128×80.

4. MULTIMODAL FUSION USING CANONICAL CORRELATION ANALYSIS

In this section we consider the early and late integration of the extracted audio and lip texture features. In the early integration the audio and lip texture features are mapped to an audio-lip correlated space using canonical correlation analysis (CCA). Subsequently, possible late integration techniques that combine audio, lip texture and correlated audio-lip features are exploited.

4.1. Early Integration using CCA

The early integration, or equivalently data fusion, is implemented with concatenation of the correlated audio and lip texture features that are obtained using CCA. The canonical correlation analysis provides a way of measuring the linear relationship between two multidimensional variables [11]. It finds two basis spaces, one for each multidimensional variable, that are optimal with respect to correlations and, at the same time, it finds the corresponding correlations. In other words, it finds two basis spaces in which the cross-correlation matrix between the variables is diagonal and the correlations on the diagonal are maximized. The dimension of each resulting space is equal to or less than the smallest of the dimensions of the two variables. An important property of canonical correlations is that they are invariant with respect to affine transformations of the variables. This is the most important difference between the CCA and the ordinary correlation analysis which highly depends on the basis in which the variables are described.

4.1.1. CCA Problem

Let two multidimensional biometric signals be represented with $\mathbf{x}$ and $\mathbf{y}$. Further let the projection matrices be $\mathbf{w}_x$ and $\mathbf{w}_y$ such that the correlations between the projections of $\mathbf{x}$ and $\mathbf{y}$ onto range space of $\mathbf{w}_x$, $R(\mathbf{w}_x)$ and $R(\mathbf{w}_y)$ are mutually maximized. The problem becomes estimating the projection matrices $\mathbf{w}_x$ and $\mathbf{w}_y$ while maximizing the correlation coefficient, r, of the projected signals $\mathbf{w}_x^T\mathbf{x}$ and $\mathbf{w}_y^T\mathbf{y}$,

$$r \quad = \quad \frac{\mathbf{w}_x^T \mathbf{C}_{\mathbf{xy}} \mathbf{w}_y}{\sqrt{\mathbf{w}_x^T \mathbf{C}_{\mathbf{xx}} \mathbf{w}_x \mathbf{w}_y^T \mathbf{C}_{\mathbf{yy}} \mathbf{w}_y}} \tag{4}$$

such that

$$\rho_{i,j} = \left\{ \begin{array}{l} E[x_i, x_j] = \mathbf{w}_x^T \mathbf{C}_{\mathbf{xx}} \mathbf{w}_x = 0 \\ E[y_i, y_j] = \mathbf{w}_y^T \mathbf{C}_{\mathbf{yy}} \mathbf{w}_y = 0 \quad for \quad i \neq j \\ E[x_i, y_j] = \mathbf{w}_x^T \mathbf{C}_{\mathbf{xy}} \mathbf{w}_y = 0 \end{array} \right. \tag{5}$$

4.1.2. CCA Solution

Let $\mathbf{x}$ and $\mathbf{y}$ be zero mean random variables. The joint covariance matrix is defined as:

$$\mathbf{C} = \begin{bmatrix} \mathbf{C}_{xx} & \mathbf{C}_{xy} \\ \mathbf{C}_{yx} & \mathbf{C}_{yy} \end{bmatrix} = E\left[\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} \begin{bmatrix} \mathbf{x}^T & \mathbf{y}^T \end{bmatrix} \right] \quad (6)$$

where, $\mathbf{C}_{xx}$ and $\mathbf{C}_{yy}$ are the within set covariance matrices, and $\mathbf{C}_{xy} = \mathbf{C}_{yx}^T$ is the between set correlation matrix. The canonical correlations between $\mathbf{x}$ and $\mathbf{y}$ can be found by solving the eigenvalue equations,

$$\begin{aligned} \mathbf{C}_{xx}^{-1}\mathbf{C}_{xy}\mathbf{C}_{yy}^{-1}\mathbf{C}_{yx}\mathbf{w}_x &= \rho^2\mathbf{w}_x \\ \mathbf{C}_{yy}^{-1}\mathbf{C}_{yx}\mathbf{C}_{xx}^{-1}\mathbf{C}_{xy}\mathbf{w}_y &= \rho^2\mathbf{w}_y \end{aligned} \quad (7)$$

where the eigenvalues ρ^2 are the squared canonical correlations and the eigenvectors $\mathbf{w}_x$ and $\mathbf{w}_y$ are the normalized canonical correlation basis vectors. Only one of the eigenvalue equations needs to be solved since the solutions are related by

$$\begin{aligned} \mathbf{C}_{xy}\mathbf{w}_y &= \rho\lambda_x\mathbf{C}_{xx}\mathbf{w}_x \\ \mathbf{C}_{yx}\mathbf{w}_x &= \rho\lambda_y\mathbf{C}_{yy}\mathbf{w}_y \end{aligned} \quad . \quad (8)$$

where,

$$\lambda_x = \lambda_y^{-1} = \sqrt{\frac{\mathbf{w}_y^T\mathbf{C}_{yy}\mathbf{w}_y}{\mathbf{w}_x^T\mathbf{C}_{xx}\mathbf{w}_x}}. \quad (9)$$

4.1.3. Audio-Visual Feature Extraction Using CCA

Let speech and lip texture features be represented with $\boldsymbol{f}_A$ and $\boldsymbol{f}_L$, respectively. One can apply CCA to find two new feature sets $\boldsymbol{f}'_A = \mathbf{w}_A^T\boldsymbol{f}_A$ and $\boldsymbol{f}'_L = \mathbf{w}_L^T\boldsymbol{f}_L$ such that the between class correlation coefficient matrix of $\boldsymbol{f}'_A$ and $\boldsymbol{f}'_L$ is diagonal with maximized diagonal terms. However, maximized diagonal terms do not necessarily mean that all the diagonal terms exhibit strong correlations. Hence one can pick the maximally correlated components that are above a certain correlation threshold t_h. Let us denote the projection vector that corresponds to the diagonal terms larger than the threshold t_h by $\tilde{\mathbf{w}}_A$. Then, the corresponding projections of $\boldsymbol{f}_A$ and $\boldsymbol{f}_L$ are given as,

$$\tilde{\boldsymbol{f}}_A = \tilde{\mathbf{w}}_A^T\boldsymbol{f}_A \quad (10)$$

$$\tilde{\boldsymbol{f}}_L = \tilde{\mathbf{w}}_L^T\boldsymbol{f}_L \quad (11)$$

Here, $\tilde{\boldsymbol{f}}_A$ and $\tilde{\boldsymbol{f}}_L$ are the correlated components that are embedded in $\boldsymbol{f}_A$ and $\boldsymbol{f}_L$. Furthermore, the early integration, or equivalently data fusion,is performed with concatenation of correlated audio-visual components. The resulting combined audio-visual feature is defined as,

$$\tilde{\boldsymbol{f}}_{AL} = [\tilde{\boldsymbol{f}}_A\tilde{\boldsymbol{f}}_L] \quad (12)$$

4.2. Late Integration

In the Bayesian framework, late integration can be performed using product rule assuming statistically independent modalities. Various methods have been proposed in the literature as an alternative to the product rule such as max rule, min rule and reliability-based weighted summation. In fact, the most generic way of computing joint ratios (or scores) can be expressed as a weighted summation:

$$\rho(\lambda_r) = \sum_{n=1}^{N} \omega_n\rho_n(\lambda_r) \quad \text{for } r = 1, 2, ..., R, \quad (13)$$

where $\rho_n(\lambda_r)$ is the log-likelihood of the class-conditional probability, $\log P(\boldsymbol{f}_n|\lambda_r)$, for the n-th modality $\boldsymbol{f}_n$ with class λ_r, and ω_n denotes the weighting coefficient for modality n, such that $\sum_n \omega_n = 1$. Then, the fusion problem becomes finding the optimal weight coefficients. Note that when $\omega_n = \frac{1}{N} \ \ \forall n$, (13) is equivalent to the product rule. Since the ω_n values can be regarded as the reliability values of the classifiers, we referred to this combination method as RWS (Reliability Weighted Summation) rule in [4]. The statistics and the numerical range of these likelihood scores mostly vary from one classifier to another, and thus using sigmoid and variance normalization as described in [4], they can be normalized into $(0, 1)$ interval before the fusion process. The RWS rule is employed for the decision fusion of audio, lip texture and correlated audio-lip features, using the reliability value estimation, which is described in [4].

5. EXPERIMENTAL RESULTS

Hidden Markov Models (HMM) are known to be effective structures to model the temporal behavior of the speech signal, and thus they are widely used both in audio-based speaker identification and speech recognition applications. In this work, we employ text dependent HMM structures for the open-set speaker identification system. Our database consists of audio and video signals belonging to individuals of a certain population. Thus in our system the temporal characterization of the lip-texture and audio-lip modalities are also performed using HMMs. We use word-level continuous-density HMM structures. Each speaker or utterance in the database is modeled using a separate HMM that is trained over some repetitions of the modality streams of the corresponding speaker. In the recognition process, given a test feature set, each HMM structure associated with a speaker or an utterance produces a likelihood. A world HMM model is also trained over the whole training data of the population. The log-ratio of the speaker likelihoods and the world class likelihood results in a stream of log-likelihood ratios that are used in the speaker identification process. The system identifies the person if there is a match and rejects otherwise. The performance of the speaker verification systems is often measured using the equal error rate (EER) figure. The EER is calculated as the operating point where false accept rate (FAR) equals false reject rate (FRR).

The experiments have been conducted using the MVGL-AVD audio-visual database $\mathcal{D}_n$ [4]. The database $\mathcal{D}_n$ includes 50 subjects, where each subject utters ten repetitions of her/his name as the secret phrase. A set of impostor data is also available for each subject in the population uttering five different names from the population. The $\mathcal{D}_n$ database is partitioned into two sets namely $\{\mathcal{D}_{n_A}$ and $\mathcal{D}_{\bar{n}_A}\}$, where $\mathcal{D}_{n_A}$ and $\mathcal{D}_{\bar{n}_A}$ are mutually exclusive sets each having five repetitions from each subject in the database. The subsets $\mathcal{D}_{n_A}$ and $\mathcal{D}_{\bar{n}_A}$ are used for training and testing, respectively. Since there are 50 subjects and five repetitions for each true and imposter client tests, the resulting total number of trials for both the true accepts and true rejects is 250.

Fig. 1 plots the maximized diagonal terms of the between class correlation coefficient matrix after the CCA analysis of audio and lip-texture features. As observed from Fig. 1, the maximum correlation coefficient is around 0.7, and 15 correlation coefficients out of 39 are higher than 0.1 threshold. Table 1 presents the EER performance of the early integration of correlated audio-lip combined features for varying correlation coefficient threshold t_h. Note that, when all the 39 transformed coefficients are used, the EER performance is 6.6%. The EER performance is observed to have a minimum around 4.8% for threshold values from 0.1 to 0.4. The optimal

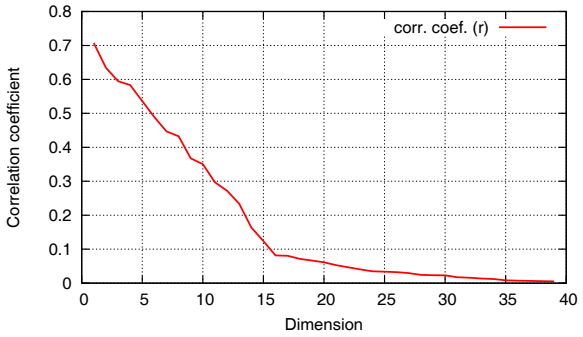

Fig. 1. Sorted correlation coefficient plot for audio and lip texture CCA analysis.

threshold is found to be 0.4 that minimize the EER performance and the feature dimension. Table 2 presents the EER performances for

Table 1. Speaker identification results for data fusion with CCA: Equal error rates at varying correlation coefficient threshold values (t_h) with the corresponding projection dimension (Dim).

	EER (%) at (t_h, Dim)						
t_h	0.0	0.1	0.2	0.3	0.4	0.5	0.6
Dim	39	15	13	11	8	6	3
$\tilde{\boldsymbol{f}}_{AL}$	6.6	4.8	5.2	5.0	4.8	7.2	10.1

various combinations of early and late integration techniques with using the optimal threshold 0.4. We observe that the use of correlated audio-lip features improves the performance over the early fusion of audio and lip texture features obtained without correlation analysis. Here, note that the early fusion of audio and lip texture features with and without correlation analysis are denoted with $\tilde{\boldsymbol{f}}_{AL}$ and $\boldsymbol{f}_{AL}$ with corresponding feature dimensions $8+8=16$ and $39+50=89$ respectively. It is worth to mention that RWS decision fusion do not introduce the data sparsity since it is late integration.

6. CONCLUSION

In this paper, we have presented a multimodal fusion system with canonical correlation analysis to improve the performance of audio-visual speaker identification systems that use audio and lip modalities. The most important finding of this work is that the performance of early integration (or data fusion) for audio-visual speaker identification can be increased by concatenating the inter-correlated parts of the audio and lip feature vectors. Since each modality carries also additional independent information, e.g. the texture of the lip region, about the identity of a speaker, the overall recognition system employs an optimal combination of the early and late integration techniques combining lip, audio and correlated audio-lip feature vectors. The future work will address the robustness of the proposed scheme against noise, where audio alone feature set do not yield such low EER, and will explore the use of CCA for extraction and fusion of mutually uncorrelated components.

Table 2. Speaker identification results for decision fusion: Equal error rates for different modalities and multimodal fusion scenarios ($+$ represents RWS decision fusion).

Modality	EER (%)
$\boldsymbol{f}_A$	1.0
$\boldsymbol{f}_L$	6.0
$\boldsymbol{f}_{AL}$	7.2
$\tilde{\boldsymbol{f}}_{AL}$	4.8
$\boldsymbol{f}_A + \boldsymbol{f}_{AL}$	0.6
$\boldsymbol{f}_A + \tilde{\boldsymbol{f}}_{AL}$	0.4
$\boldsymbol{f}_A + \boldsymbol{f}_L + \boldsymbol{f}_{AL}$	1.2
$\boldsymbol{f}_A + \boldsymbol{f}_L + \tilde{\boldsymbol{f}}_{AL}$	1.0

7. REFERENCES

[1] J.P. Campbell, "Speaker recognition: A tutorial," *Proceedings of the IEEE*, vol. 85, no. 9, pp. 1437–1462, September 1997.

[2] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, vol. 3, no. 1, pp. 586–591, September 1991.

[3] R.W. Frischholz and U. Dieckmann, "BioID: A multimodal biometric identification system," *Journal of IEEE Computer*, vol. 33, no. 2, pp. 64–68, February 2000.

[4] E. Erzin, Y. Yemez, and A.M. Tekalp, "Multimodal speaker identification using an adaptive classifier cascade based on modality reliability," *IEEE Transactions on Multimedia*, vol. 7, no. 5, pp. 840–852, October 2005.

[5] T. Wark and S. Sridharan, "Adaptive fusion of speech and lip information for robust speaker identification," *Digital Signal Processing*, vol. 11, no. 3, pp. 169–186, July 2001.

[6] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A.W. Senior, "Recent Advances in the Automatic Recognition of Audio-Visual Speech," *Proc. of the IEEE*, vol. 91, no. 9, September 2003.

[7] J. W. Fisher III and T. Darrell, "Speaker association with signal-level audiovisual fusion," *IEEE Transactions on Multimedia*, vol. 1, pp. 406–413, June 2004.

[8] M. E. Sargin, E. Erzin, Y. Yemez, and A. M. Tekalp, "Lip feature extraction based on audio-visual correlation," *Proc. of the European Signal Processing Conference 2005 (EUSIPCO'05)*, 2005.

[9] J.-M. Odobez and P. Bouthemy, "Robust Multiresolution Estimation of Parametric Motion Models," *Journal of Visual Communication and Image Representation*, vol. 6, no. 4, pp. 348–365, December 1995.

[10] H.E. Cetingul, Y. Yemez, E. Erzin, and A.M. Tekalp, "Discriminative Lip-Motion Features for Biometric Speaker Identification," *Proc. of the Int. Conf. on Image Processing 2004 (ICIP 2004)*, pp. 2023–2026, October 2004.

[11] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, pp. 321–377, 1936.