Audio-Visual Synchronization and Fusion using Canonical Correlation Analysis

M. E. Sargın, *Student Member, IEEE,* Y. Yemez, E. Erzin, *Member, IEEE,* and A. M. Tekalp, *Fellow, IEEE*

Abstract— **It is well-known that early integration (also called data fusion) is effective when the modalities are correlated, and late integration (also called decision or opinion fusion) is optimal when modalities are uncorrelated. In this paper, we propose a new multimodal fusion strategy for open-set speaker identification using a combination of early and late integration following canonical correlation analysis (CCA) of speech and lip texture features. We also propose a method for high precision synchronization of the speech and lip features using CCA prior to the proposed fusion. Experimental results show that i) the proposed fusion strategy yields the best equal error rates (EER), which are used to quantify the performance of the fusion strategy for open-set speaker identification, and ii) precise synchronization prior to fusion improves the EER; hence, the best EER is obtained when the proposed synchronization scheme is employed together with the proposed fusion strategy. We note that the proposed fusion strategy outperforms others because the features used in the late integration are truly uncorrelated, since they are output of the CCA analysis.**

I. Introduction

Speech and lip texture/movement are physiologically coupled modalities; hence, they are highly correlated. However, depending on the features employed for representation, they may also contain some uncorrelated components. Two fusion strategies are commonly employed in the literature: The late integration strategy [1], which is also referred to as decision or opinion fusion, is optimal in case the contributing modalities are uncorrelated, and thus the resulting partial decisions are statistically independent. On the other hand, early integration, which is also referred to as data fusion, combines modalities at the data or feature level and may be effective if the modalities are highly correlated. However, dimensionality is an important problem and in practice, decision fusion can outperform data fusion even if the modalities are tightly coupled. Neither of these two alternatives actually offers an optimal solution alone, especially when the modalities contain a mixture of correlated and uncorrelated components.

Lip information has extensively been employed in the state-of-the-art audio-visual speech recognition applications [2], since it is natural to expect that speech content can be revealed through lip reading. Lip movement patterns also contain information about the identity of the speaker. Yet, audio and lip information have been used for speaker identification and/or verification in relatively few works such as [3], [4], [5], [6], [7], [8]. These works are mainly based on decision fusion, where audio is generally modeled by mel frequency cepstral coefficients (MFCC). Several feature sets can be used for the lip modality such as shape, motion and texture. In texture-based approaches, pure or DCT-domain lip image intensities are commonly used as features [5], [9]. Dimension reduction techniques, such as principle component analysis (PCA), linear discriminant analysis (LDA) or Discrete Cosine Transform (DCT), are independently applied to the lip and speech modalities regardless of the mutual information between them.

There is relatively little work available on explicit analysis of audio-visual correlations. In [10], the speaker association problem is addressed via an information theoretic method, which aims to maximize the mutual information between the projections of audiovisual measurements so as to detect the parts of video, that are highly correlated with the speech signal. In [11], the information fusion problem is addressed in the context of handwritten character recognition. The correlated projections of multiple features, which are assumed to be maximally informative, are first extracted by using canonical

M. E. Sargın, Y. Yemez, E. Erzin and A. M. Tekalp are with the Koç University. (e-mail: msargin@ku.edu.tr; yyemez@ku.edu.tr; eerzin@ku.edu.tr; mtekalp@ku.edu.tr).

This work has been supported by the European FP6 Network of Excellence SIMILAR (http://www.similar.cc).

correlation analysis (CCA), and then concatenated. However their fusion scheme is not optimal, because the uncorrelated components, which may also be informative, are not taken into account; moreover, the combined features are all derived from a single modality. In [12], CCA is used for speaker adaptation to improve speech recognition performance.

Audiovisual correlation analysis has also been used in the literature to address the problem of temporal asynchrony between audio-visual features, such as in [13] that uses product HMMs (Hidden Markov Models) and as in [14] that uses CCA on audio and face video. In the case of lip movement and speech, asynchrony may occur not only due to imperfections of the acquisition setup, but also due to a natural delay between the acoustic and facial components of the speaking act.

We address the open-set speaker identification framework to demonstrate audio-visual synchronization and fusion using CCA. The speaker identification problem can be formulated as either an open-set or a closed-set identification problem. In the closed-set problem, a reject scenario is not defined and an unknown speaker is classified as one of the N registered people. In the open-set problem, the objective is, given the data of an unknown person, to find whether the person is registered in the database or not; the system identifies the person if there is a match and rejects otherwise. Hence, the problem can be thought of as an $N+1$ class problem, including a reject class. Verification problem can be considered as a special case of the open-set identification problem with $N=1$. 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 propose using canonical correlation analysis (CCA) to improve the performance of multimodal recognition systems that involve modalities having a mixture of correlated and uncorrelated components. More specifically, the multimodal recognition system is addressed within an open-set speaker identification framework. Audio and lip modalities are represented by mel-frequency cepstral and intensity-based DCT coefficients, respectively. There are two important contributions: First, we propose a simple CCA-based technique for synchronization of audio and lip modalities to optimize the performance of the data fusion process.[1] Second, we propose a multimodal fusion strategy based on canonical correlation analysis that first extracts the correlated components of audio and lip features, and then employs an optimal combination of early and late integration schemes to fuse the extracted features. The paper is organized as follows: In Section II, we review basics of the open-set audio-visual speaker identification problem. We address the audiovisual synchronization problem in Section III, and propose a CCA-based synchronization method. The proposed multimodal fusion scheme with canonical correlation analysis is presented in Section IV. Experimental results are discussed in Section V and concluding remarks are given in Section VI. Finally in Appendix, we provide a brief review of the canonical correlation analysis problem, where we also clarify the terminology and notation used throughout the paper.

II. The Audio-Visual Speaker Identification Problem

This section provides an overview of the open-set audio-visual speaker identification problem, since we present the proposed fusion strategy in the context of this application.

In open-set speaker identification, the objective is to find whether the given input audio and video features belong to one of the R subjects registered in the database or not; the system identifies the speaker if there is a match, rejects otherwise. Hence, the problem can be formulated as an $R+1$ class identification problem, where there are R subjects and a reject class. For the open-set identification problem, we employ a maximum likelihood solution through the likelihood ratio test as described in [5]. The likelihood ratio is defined as

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

where $\mathbf{f}$ is the observation from an unknown speaker, λ_r is the r-th registered speaker class, and λ_{R+1} is the impostor (reject) class. The conditional probability for the reject class, $P(\mathbf{f}|\lambda_{R+1})$, is approximated by using all available training data across all subjects. Then, the decision strategy can be implemented in two steps. First, determine

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

[1] A preliminary version of this method was presented in [15].

and then

$$\text{if}\ \ \rho(\lambda_*) \underset{\text{reject}}{\overset{\text{accept}}{\gtrless}} \tau \tag{3}$$

where λ_* denotes the speaker class with the maximum likelihood ratio and τ is the optimal threshold which is usually determined experimentally to achieve the desired false accept or false reject rate.

A. Computation of Class Conditional Probabilities

Computation of class-conditional probabilities needs a prior modeling step. Hidden Markov Models (HMM) are known to be effective structures to model the temporal behavior of the speech signal and hence they are widely used in audio-based speaker identification and speech recognition applications. In this study, we address a text-dependent open-set speaker identification application and our database consists of audio and video signals belonging to individuals of a certain population. We use word-level continuous-density HMM structures with 6 left-to-right states and single mixture for temporal characterization of both lip-texture and audio modalities. Each speaker in the database is modeled using a separate HMM, which is trained over some repetitions of the feature streams observed from the corresponding speaker and modality. An HMM model for the impostor class, λ_{R+1}, is also trained over the whole training data of the population. In the recognition process, given a test feature stream, each HMM structure associated with a speaker produces a likelihood ratio. The likelihood ratio test as defined in (3) identifies the person if there is a match and rejects otherwise.

The performance of speaker verification/identification systems is often measured using the equal error rate (EER). The EER is calculated as the operating point, where the false accept rate (FAR) equals the false reject rate (FRR). In the open-set identification problem, the false accept and false reject rates can be defined as,

$$\text{FAR} = 100 \times \frac{F_a}{N_a + N_r} \quad \text{and} \quad \text{FRR} = 100 \times \frac{F_r}{N_a}, \tag{4}$$

where F_a and F_r are the number of false accepts and rejects, and N_a and N_r are the total number of trials for the true and impostor clients in the testing, respectively.

B. Audio-Visual Feature Extraction

We use the mel-frequency cepstral coefficients (MFCC) as features for the audio modality, which are known to be robust and effective features and thus commonly employed in speaker recognition systems. The audio stream is processed over 10 msec frames centered on 25 msec Hamming window for 16 kHz sampled audio signal. The audio feature vector for each 10 msec frame is formed as a collection of 13 MFCC coefficients together with the first and second derivatives, for a total of 39 coefficients. We denote the audio feature vector by $\mathbf{f}_A$ and its dimension by N_A.

Each video stream has gray-level frames of size 720×576 pixels containing the frontal view of a speaker's head at a rate of 15 fps. For the visual features, a preprocessing step is employed to locate the lip region in each frame, and to eliminate global motion of the head between the frames so that the extracted motion features within the lip region provides us with only the motion 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 pyramids and 12-parameter quadratic motion model [16]. The face images are successively warped according to these calculated parameters [17] to align the lip regions. Since the viewing parameters of the camera are identical for all speakers, the lip covers a region of almost the same size in all video frames. Hence by hand-labeling the mid-point of the lip region in the first frame, we automatically extract a sequence of lip frames of size 128×80 from each video stream. The lip texture features, denoted by $\mathbf{f}_L$ of dimension N_L, are the first 50 zig-zag scanned 2D DCT coefficients of the luminance component within this rectangular lip region. These features implicitly represent lip movements with texture. The texture information itself might sometimes carry additional useful information for discrimination; but in some other cases it may also degrade the recognition performance since it is sensitive to acquisition conditions.

III. Audio-Visual Feature Synchronization using CCA

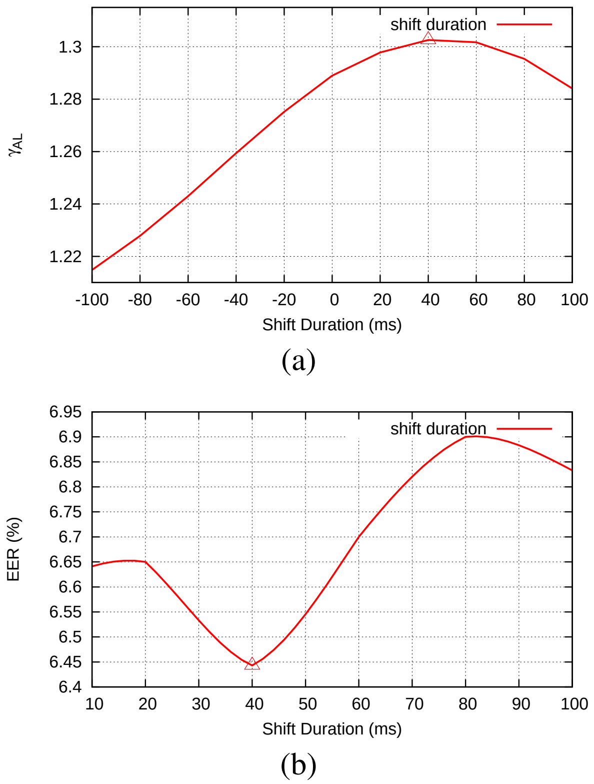
Early integration techniques require the features extracted from different modalities to be exactly at the same rate and in synchrony. In our case, the audio features are extracted at a rate of 100 audio fps, whereas the lip features have only a frame rate of 15 video fps. Thus prior to early integration, the lip features are interpolated using cubic splines to match the audio frame rate. Let us denote the audio and lip features of the k-th 10ms frame by $\mathbf{f}_A^k$ and $\mathbf{f}_L^k$, respectively. The audio and visual features need to be precisely synchronized in the interpolated frame scale before the data fusion, so that the correlations between them can better be exploited. We propose using the canonical correlation analysis (CCA) to achieve synchronization (see Appendix for a brief review of the canonical correlation analysis). The problem then becomes, given a set of realizations of $\mathbf{f}_A^{k+s}$ and $\mathbf{f}_L^k$, finding the delay s^* between audio and lip features, that maximizes the mutual information.

The CCA requires the covariance matrix of the concatenated audio and lip feature vector to be estimated using the whole set of realizations. The canonical correlations γ_i, $i = 1, 2, ..., N$, where N is the minimum of the audio and lip feature dimensions, which is 39 in our case, can then be computed from the estimated joint covariance matrix as described in Appendix. Based upon these canonical correlations, we define an overall audio-visual correlation measure $\gamma_{AL}(s)$ between audio-visual features $\mathbf{f}_A^{k+s}$ and $\mathbf{f}_L^k$ as,

$$\gamma_{AL}(s) = \sum_{i=1}^{N} \gamma_i^2 \tag{5}$$

which is a function of the delay variable s. The CCA is applied to the audio-visual features with varying values of s, and for each s the value of the correlation measure $\gamma_{AL}(s)$ is computed.

Figure 1(a) displays the behavior of $\gamma_{AL}(s)$ with varying s. As observed from the figure, the correlation measure, $\gamma_{AL}(s)$, is maximized for $s = 4$. This indicates that there is a 40 ms asynchrony between the features $\mathbf{f}_A$ and $\mathbf{f}_L$. Hence, for the rest of the paper, the lip features are shifted by 4 frames prior to their fusion with the audio features. This inference is also supported with the speaker identification results that we obtained using early integration of audiovisual features. The equal error rates obtained for varying shift durations are plotted in Figure 1(b), where we observe that the optimal shift s^* found by our CCA-based synchronization method yields the best EER performance.



(a)

(b)

Fig. 1. CCA-based synchronization results: (a) Correlation measure γ_{AL} and (b) speaker identification equal error rates, for varying values of shift duration s.

IV. Multimodal Fusion Using CCA

In this section, we propose a combination of early and late integration of the synchronized audio and lip texture features. For the early integration, the audio and lip features are first transformed using the CCA. New strategies for early integration of the correlated CCA components are proposed in Section IV-A, whereas the best combination of early and late integration schemes for the overall multi-modal fusion strategy is presented in Section IV-B.

A. *Integration of Correlated CCA Components*

Let the N-dimensional CCA-transformed audio and lip features be represented with $\mathbf{f}'_A$ and $\mathbf{f}'_L$, respectively, where N is chosen as the minimum of the audio feature dimension N_A and the lip feature dimension N_L. The between-set covariance matrix of $\mathbf{f}'_A$ and $\mathbf{f}'_L$ is a diagonal matrix with N diagonal terms, each of which corresponds to a squared canonical correlation (see Appendix). However, each of these diagonal terms does not necessarily exhibit a strong correlation. Hence, one can pick the highly correlated components from the transformed vectors, discarding those with small canonical correlations. Fig. 2 plots the canonical correlations of the audio-visual features, obtained by applying CCA to our database. As observed from Fig. 2, the maximum correlation coefficient is around 0.65, and 18 correlation coefficients out of 39 are higher than 0.05 threshold.

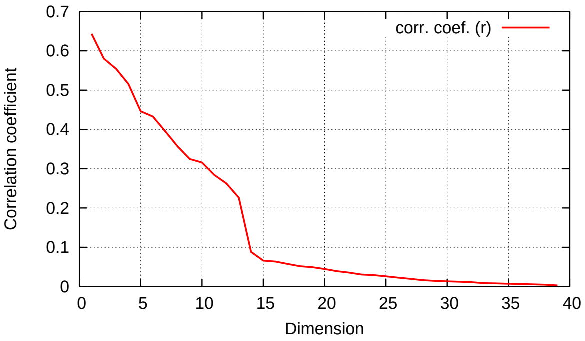


Fig. 2. Canonical correlations resulting from audio-lip CCA analysis (sorted in decreasing order).

We define the highly correlated components as the projections of the original features onto the CCA basis vectors along which the canonical correlations are above a certain threshold T_h. Let us denote the two transformations corresponding to these canonical basis vectors by $\tilde{\mathbf{H}}_A$ and $\tilde{\mathbf{H}}_L$, respectively for the audio and lip modalities. Then, the correlated projections, $\tilde{\mathbf{f}}_A$ and $\tilde{\mathbf{f}}_L$, each with dimension M, are given by

$$\begin{aligned} \tilde{\mathbf{f}}_A &= \tilde{\mathbf{H}}_A \mathbf{f}_A \\ \tilde{\mathbf{f}}_L &= \tilde{\mathbf{H}}_L \mathbf{f}_L \end{aligned} \tag{6}$$

Here, $\tilde{\mathbf{f}}_A$ and $\tilde{\mathbf{f}}_L$ can be regarded as the correlated components embedded in $\mathbf{f}_A$ and $\mathbf{f}_L$.

1) Early Integration by Concatenation: The early integration can simply be performed by concatenation of these correlated M dimensional projection vectors. The resulting combined audio-visual feature vector is thus given by

$$\tilde{\mathbf{f}}_{AL} = [\ \tilde{\mathbf{f}}_A^T \ \ \tilde{\mathbf{f}}_L^T\ ]_{2M \times 1}^T \tag{7}$$

2) Integration by Combining Weak Classifiers: An alternative integration strategy can be developed by decomposing the correlated CCA components, $\tilde{\mathbf{f}}_A$ and $\tilde{\mathbf{f}}_L$, into pairs of components, which are statistically independent from each other, but pairwise highly correlated. Recall from Appendix that the M pairs of canonical components, $(\tilde{f}_{Ai}, \tilde{f}_{Li})$, that are statistically independent from each other, can be computed via the projections

$$\begin{aligned} \tilde{f}_{Ai} &= \tilde{\mathbf{h}}_{Ai}^T \mathbf{f}_A \\ \tilde{f}_{Li} &= \tilde{\mathbf{h}}_{Li}^T \mathbf{f}_L \end{aligned} \tag{8}$$

where $\tilde{\mathbf{h}}_{Ai}$ and $\tilde{\mathbf{h}}_{Li}$ are the corresponding CCA basis vectors on which the projections are highly correlated so that $\gamma_i > T_h$.

In the new integration scheme, we employ M different HMM-based classifiers as defined in Section II, one for each pair of correlated speech-lip canonical components. Each canonical pair, that is, a two-dimensional concatenated vector, becomes input to the associated *weak* classifier. The decisions of these M weak classifiers are then combined using a late integration technique, as depicted in Fig. 3. The late integration computes a combined log-likelihood ratio $\rho_{AL}(\lambda_r)$ using Bayesian decision fusion (or the so-called product rule),

$$\rho_{AL}(\lambda_r) = \sum_{i=1}^{M} \rho_{ALi}(\lambda_r) \tag{9}$$

where $\rho_{ALi}(\lambda_r)$ is the likelihood ratio of the feature $\tilde{\mathbf{f}}_{ALi}$ for the r-th registered speaker class λ_r as defined in (1). The use of a weak classifier combination avoids the dimensionality problem of feature concatenation, and thus eases the task of feature modeling. Moreover, the late integration technique that combines the canonical pairs is optimal since these pairs of feature components are statistically independent.

B. The Proposed Multimodal Fusion Scheme

The two options presented in Section IV-A for integration of the correlated CCA components do not take into account the mutually independent information embedded in the features that might also convey discriminative information.

The solution that we propose to exploit the mutually independent information is to employ a final step of late integration that incorporates the original audio and lip feature vectors, $\mathbf{f}_A$ and $\mathbf{f}_L$, as depicted in Figure 3. The experiments that we have conducted show that the uncorrelated components of the intensity-based lip feature vector can be noisy and do not carry useful additional discriminative information about a speaker's identity. Hence, our optimal configuration discards the original lip feature vector and incorporates only the audio features into the fusion scheme.

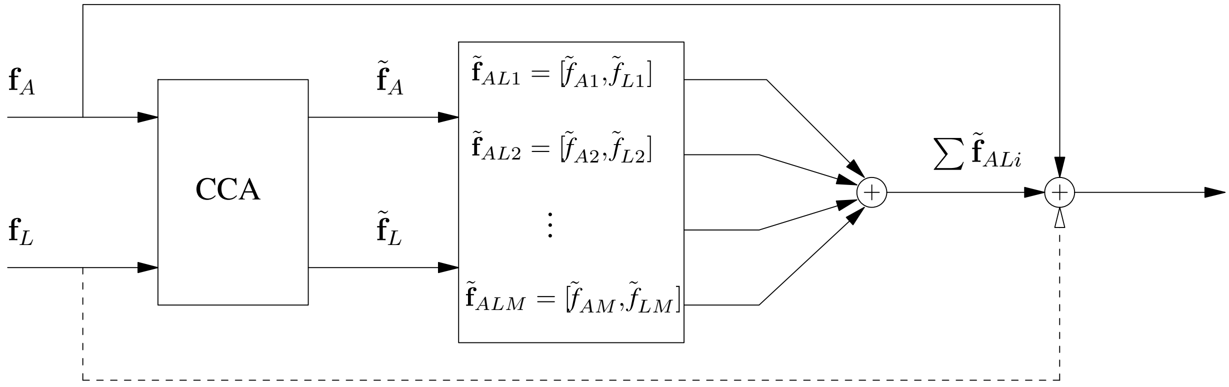

$\mathbf{f}_A$
$\mathbf{f}_L$
CCA
$\tilde{\mathbf{f}}_A$
$\tilde{\mathbf{f}}_L$
$\tilde{\mathbf{f}}_{AL1} = [\tilde{f}_{A1}, \tilde{f}_{L1}]$
$\tilde{\mathbf{f}}_{AL2} = [\tilde{f}_{A2}, \tilde{f}_{L2}]$
$\vdots$
$\tilde{\mathbf{f}}_{ALM} = [\tilde{f}_{AM}, \tilde{f}_{LM}]$
$\sum \tilde{\mathbf{f}}_{ALi}$
$+$
$+$

Fig. 3. The proposed fusion scheme, where $\sum$ denotes Bayesian decision fusion. The optimal configuration discards the late integration of the lip feature vector and incorporates only the audio features.

V. Experimental Results

The proposed multimodal speaker identification system has been tested on the MVGL-AVD^2 audio-visual database [5]. The database 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 with each subject in the population, uttering five different names from the population. Sample face images from the MVGL-AVD database are given in Fig. 4. Experimental results are extracted with two-fold cross-validation. In each experimental trial, the ten repetitions of secret phrase recordings for each subject are randomly divided into two sets, training and testing, of five repetitions. The training data is used in the CCA analysis of the audio-visual features and in the training of HMM structures for each subject. The testing data and the impostor recordings are used in the performance evaluations. The equal error rate (EER) figures are calculated over four independent experimental trials, where in each trial we have 250 true accept and 250 imposter recordings.

The EER results for various fusion strategies using CCA are presented in Table I for several values of the correlation threshold T_h, where M denotes the number of correlated components above the threshold. In Table I, $\tilde{\mathbf{f}}_{AL}$ and $\sum \tilde{\mathbf{f}}_{ALi}$ respectively denote integration by concatenation and integration by combining weak classifiers as described in Section IV-A, whereas $+$ stands for Bayesian decision fusion (also called product rule) [5]. The minimum equal error rates in each row are indicated in bold.

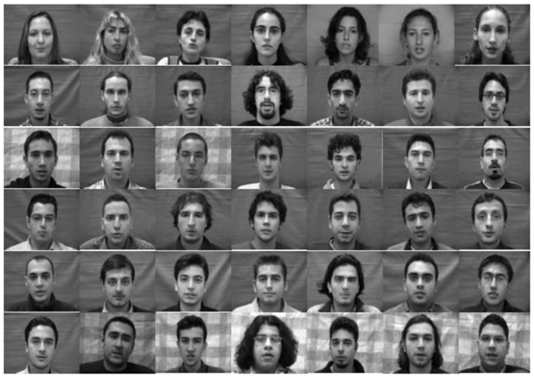


Fig. 4. Sample face images from the MVGL-AVD database.

We observe that for early integration by concatenation, as the threshold T_h decreases, that is, as the transformed vector dimension M increases, the EER for the concatenated audio-lip feature, $\tilde{\mathbf{f}}_{AL}$, first decreases and then increases, achieving an optimal 3.8% EER value at the threshold $T_h = 0.25$. On the other hand, the EER obtained using a combination of weak classifiers, $\sum \tilde{\mathbf{f}}_{ALi}$, first decreases with decreasing threshold and then saturates at 3.8% EER. Hence, the EER performance in this case is more robust to selection of the threshold value.

In the next two rows of Table I, the decision fusion results of the audio-only and the correlated audio-lip based classifiers are presented. When the audio-only classifier is combined with the concatenated audio-lip classifier, $(\mathbf{f}_A + \tilde{\mathbf{f}}_{AL})$, the best EER performance is observed as 0.6%. Furthermore, the EER drops to 0.3% for the proposed fusion structure in Fig. 3, that is, for fusion of the audio-only classifier and the combined weak classifiers, $(\mathbf{f}_A + \sum \tilde{\mathbf{f}}_{ALi})$. Note that the performance saturates at this optimal EER value. Hence, the proposed

[2] The MVGL-AVD database has been acquired at MVGL Laboratory of Koç University and it is available upon request for academic purposes.

TABLE I

SPEAKER IDENTIFICATION RESULTS FOR MULTIMODAL FUSION USING CCA: EER FOR VARYING VALUES OF THE CORRELATION THRESHOLD (T_h) AND THE CORRESPONDING PROJECTION DIMENSION (M).

	EER (%) at (T_h, M)								
T_h	0.0	0.01	0.02	0.05	0.25	0.30	0.35	0.45	0.50
M	39	30	24	15	13	11	8	6	3
$\tilde{\mathbf{f}}_{AL}$	6.2	5.3	5.1	4.1	**3.8**	3.9	4.2	5.8	10.0
$\sum \tilde{\mathbf{f}}_{ALi}$	**3.8**	**3.8**	3.9	3.9	**3.8**	4.6	5.8	7.5	13.5
$\mathbf{f}_A + \tilde{\mathbf{f}}_{AL}$	0.9	0.8	**0.6**	**0.6**	**0.6**	**0.6**	0.8	1.0	2.7
$\mathbf{f}_A + \sum \tilde{\mathbf{f}}_{ALi}$	0.4	0.4	**0.3**	0.5	0.7	0.9	0.8	1.3	4.3
$\mathbf{f}_A + \mathbf{f}_L + \tilde{\mathbf{f}}_{AL}$	1.3	1.1	1.2	1.1	**1.0**	**1.0**	1.1	1.1	2.1
$\mathbf{f}_A + \mathbf{f}_L + \sum \tilde{\mathbf{f}}_{ALi}$	0.9	**0.8**	**0.8**	0.9	1.1	1.2	1.2	1.5	2.4

fusion scheme is also robust to selection of the threshold T_h, or equivalently, to selection of the optimal correlated audio-visual feature dimension M.

The last two rows of Table I present the EER performances when the lip-only classifier is further included in the final decision fusion. The optimal EER performance degrades to 1.0% and 0.8% with $\mathbf{f}_A + \mathbf{f}_L + \tilde{\mathbf{f}}_{AL}$ and $\mathbf{f}_A + \mathbf{f}_L + \sum \tilde{\mathbf{f}}_{ALi}$ decision fusion schemes, respectively. The performance degradation is due to the inclusion of the uncorrelated lip information which is noisy, mainly because the lip texture alone is very sensitive to lighting conditions during acquisition.

For benchmarking, we also present the EER results in Table II for unimodal and multimodal audio-visual speaker identification schemes in comparison with the best EER from Table I. We observe that the conventional early fusion by means of concatenation of audio-visual features, ($\mathbf{f}_{AL}$), does not bring any performance gain, and performs worse than the audio-only identification. The late integration of audio-visual classifiers, including $\mathbf{f}_A + \mathbf{f}_L$ and $\mathbf{f}_A + \mathbf{f}_{AL}$, brings performance gain over audio-only identification. The last two rows of Table II present the best EER obtained with fusion schemes using CCA. We observe that the best EER is achieved by using the proposed fusion structure in Fig. 3. Bayesian combination of weak classifiers after CCA outperforms Bayesian decision fusion of audio and combination (data fusion) of audio and lip motion features significantly. This is mainly due to fact that the features used in the Bayesian combination of weak classifiers are truly uncorrelated since they are output of the CCA analysis.

TABLE II

COMPARISON OF EER FOR VARIOUS AUDIO-VISUAL SPEAKER IDENTIFICATION STRATEGIES.

Strategy		EER (%)
Unimodal audio:	$\mathbf{f}_A$	1.1
Unimodal lip:	$\mathbf{f}_L$	7.0
Data fusion (concatenation):	$\mathbf{f}_{AL}$	6.4
Decision fusion:	$\mathbf{f}_A + \mathbf{f}_L$	0.7
Combined fusion (no CCA):	$\mathbf{f}_A + \mathbf{f}_{AL}$	0.8
Combined fusion using CCA:	$\mathbf{f}_A + \tilde{\mathbf{f}}_{AL}$	0.6
Combined optimal fusion:	$\mathbf{f}_A + \sum \tilde{\mathbf{f}}_{ALi}$	0.3

VI. CONCLUSIONS

We have presented new methods for multimodal synchronization and fusion using canonical correlation analysis. Experimental results show that the precise synchronization of modalities prior to fusion improves the speaker identification performance, and the proposed fusion strategy, following the proposed synchronization method, yields the best EER performance for the open-set audio-visual speaker identification. More specifically, we observe that i) in the late integration of weak classifiers, as the number of CCA transformed correlated audio-visual feature pairs increases, the equal error rate robustly drops to a minimum level and stays there, and ii) the best multimodal fusion strategy is constructed when the combination of weak classifiers is further integrated with the audio-only classifier.

Although the proposed fusion strategy using the CCA has only been demonstrated for the fusion of speech and lip modalities in the context of open-set speaker identification, it can indeed be applied to the fusion of any pair of modalities, which can be modeled as a mixture of correlated and uncorrelated components.

Appendix

Canonical Correlation Analysis

The canonical correlation analysis (CCA) is a linear statistical analysis technique, that provides a way of measuring how much and in what directions are two given multidimensional variables correlated. It was first proposed in [18], and then found applications in various fields [19],[20].

Let $\mathbf{x}$ and $\mathbf{y}$ be two jointly Gaussian, zero-mean multidimensional variables with dimensions N_x and N_y, respectively. CCA seeks two linear transformations $\mathbf{H}_x$ and $\mathbf{H}_y$, one for each multidimensional variable, that maximize the mutual information between the transformed variables $\mathbf{x}'$ and $\mathbf{y}'$

$$\begin{aligned} \mathbf{x}' &= \mathbf{H}_x \mathbf{x} \\ \mathbf{y}' &= \mathbf{H}_y \mathbf{y}, \end{aligned} \tag{10}$$

where the multidimensional variables are represented with column vectors. We will refer to the pair $(\mathbf{x}', \mathbf{y}')$ as the CCA transform of $\mathbf{x}$ and $\mathbf{y}$. The transformations $\mathbf{H}_x$ and $\mathbf{H}_y$ are represented by matrices of dimensions $N \times N_x$ and $N \times N_y$, respectively, where $N \leq \min(N_x, N_y)$:

$$\mathbf{H}_x = \begin{bmatrix} \mathbf{h}_{x1}^T \\ \mathbf{h}_{x2}^T \\ \vdots \\ \mathbf{h}_{xN}^T \end{bmatrix}, \mathbf{H}_y = \begin{bmatrix} \mathbf{h}_{y1}^T \\ \mathbf{h}_{y2}^T \\ \vdots \\ \mathbf{h}_{yN}^T \end{bmatrix} \tag{11}$$

The rows of each of these matrices, $\{\mathbf{h}_{xi}\}$ and $\{\mathbf{h}_{yi}\}$, $i = 1, 2, ..., N$, form an orthonormal basis for the corresponding transform space and are referred to as CCA basis vectors. The first pair of these basis vectors, $(\mathbf{h}_{x1}, \mathbf{h}_{y1})$, is given by the directions along which the projections are maximally correlated:

$$(\mathbf{h}_{x1}, \mathbf{h}_{y1}) = \arg \max_{(\mathbf{h}_x, \mathbf{h}_y)} \operatorname{Corr}(\mathbf{h}_x^T \mathbf{x}, \mathbf{h}_y^T \mathbf{y}) \tag{12}$$

The projections, $x_1' = \mathbf{h}_{x1}^T \mathbf{x}$ and $y_1' = \mathbf{h}_{y1}^T \mathbf{y}$, are the first pair of canonical components. The second pair of CCA basis vectors can then be extracted using the residuals left after removing the components along the first pair of basis vectors from the original variables. This is equivalent to maximizing the same correlation, but this time subject to the constraint that the projections are to be uncorrelated with the first pair of canonical components. The same procedure can be iterated to extract the remaining canonical pairs.

The CCA basis vectors are usually computed by solving an equivalent eigenvalue problem. The joint covariance matrix of the two random variables $\mathbf{x}$ and $\mathbf{y}$ is defined as:

$$\mathbf{C} = \begin{bmatrix} \mathbf{C}_{xx} & \mathbf{C}_{xy} \\ \mathbf{C}_{yx} & \mathbf{C}_{yy} \end{bmatrix} \tag{13}$$

where $\mathbf{C}_{xx}$ and $\mathbf{C}_{yy}$ are the within-set covariance matrices, $\mathbf{C}_{xy}$ is the between-set covariance matrix. These covariance matrices can be estimated using a sufficiently representative set of realizations of the random variables. The problem of CCA then becomes solving the following eigenvalue equations under the constraints $\mathbf{h}_x^T \mathbf{h}_x = 1$ and $\mathbf{h}_y^T \mathbf{h}_y = 1$,

$$\begin{aligned} \mathbf{C}_{xx}^{-1} \mathbf{C}_{xy} \mathbf{C}_{yy}^{-1} \mathbf{C}_{yx} \mathbf{h}_x &= \gamma^2 \mathbf{h}_x \\ \mathbf{C}_{yy}^{-1} \mathbf{C}_{yx} \mathbf{C}_{xx}^{-1} \mathbf{C}_{xy} \mathbf{h}_y &= \gamma^2 \mathbf{h}_y, \end{aligned} \tag{14}$$

where the eigenvectors correspond to the normalized CCA basis vectors and each associated eigenvalue γ_i, $i = 1, 2, ..., N$, is the canonical correlation between the components of the corresponding canonical pair, x_i' and y_i':

$$\gamma_i = E(x_i' y_i') \tag{15}$$

where $E(\cdot)$ is the expected value function. Since the two solutions of (14) are related by

$$\mathbf{C}_{xy} \mathbf{h}_y = \gamma \lambda_x \mathbf{C}_{xx} \mathbf{h}_x, \tag{16}$$

where

$$\lambda_x = \sqrt{\frac{\mathbf{h}_y^T \mathbf{C}_{yy} \mathbf{h}_y}{\mathbf{h}_x^T \mathbf{C}_{xx} \mathbf{h}_x}}, \tag{17}$$

it suffices to solve only one of the eigenvalue equations.

As a result, the CCA transform diagonalizes the between-set covariance matrix,

$$\mathbf{C}_{x'y'} = \mathbf{H}_x \mathbf{C}_{xy} \mathbf{H}_y \tag{18}$$

so that the diagonal entries of the resulting covariance $\mathbf{C}_{x'y'}$ correspond to the canonical correlations, γ_i. Similarly, the non-diagonal entries, which are all zero, are the cross-correlations,

$$E(x_i' y_j') = 0 \quad \text{for all } i \neq j. \tag{19}$$

Moreover, since the pairs of canonical components are uncorrelated with each other, we also have

$$E(x_i' x_j') = E(y_i' y_j') = 0 \quad \text{for all } i \neq j. \tag{20}$$

REFERENCES

[1] J. Kittler, M. Hatef, R. Duin, and J. Matas, "On combining classifiers," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, pp. 226–239, March 1998.

[2] T. Chen, "Audio-visual speech processing," *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 9–21, 2001.

[3] C. C. Chibelushi, F. Deravi, and J. S. Mason, "Audio-visual person recognition: an evaluation of data fusion strategies," in *European Conference on Security and Detection, ECOS 97*, April 1997, pp. 26–30.

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

[5] E. Erzin, Y. Yemez, and A. 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.

[6] 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.

[7] C. C. Chibelushi, F. Deravi, and J. S. D. Mason, "A review of speech-based bimodal recognition," *IEEE Transactions on Multimedia*, vol. 4, no. 1, pp. 23–37, March 2002.

[8] C. Broun, X. Zhang, R. Mersereau, and M. Clements, "Automatic speechreading with application to speaker verification," *Proc. of the Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP '02)*, vol. I, pp. 685–688, 2002.

[9] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. Senior, "Recent advances in the automatic recognition of audio-visual speech," *Proc. of the IEEE*, vol. 91, no. 9, September 2003.

[10] 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.

[11] Q.-S. Sun, S.-G. Zeng, P.-A. Heng, and D.-S. Xia, "Feature fusion method based on canonical correlation analysis and handwritten character recognition," in *IEEE Int. Conf. on Control, Automation, Robotics and Vision Conference (ICARCV)*, vol. 2, 2004, pp. 1547–1552.

[12] K. Choukri, G. Chollet, and Y. Grenier, "Spectral transformation through canonical correlation analysis for speaker adaptation in ASR," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP '86)*, 1986, pp. 2659–2662.

[13] S.Nakamura, K.Kumatani, and S.Tamura, "Multi-modal temporal asynchronicity modeling by product HMMs for robust audio-visual speech recognition," in *Proc. of the Int. Conf. on Multimodal Interfaces (ICMI '02)*, 2002, pp. 305–309.

[14] M. Slaney and M. Covell, "Facesync: A linear operator for measuring synchronization of video facial images and audio tracks," in *Proc. Neural Information Processing Systems*, 2000, pp. 814–820.

[15] 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 (EUSIPCO'05)*, 2005.

[16] 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.

[17] H. Cetingul, Y. Yemez, E. Erzin, and A. Tekalp, "Discriminative lip-motion features for biometric speaker identification," *IEEE Int. Conf. on Image Processing (ICIP '04)*, pp. 2023–2026, October 2004.

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

[19] M. Borga, "Learning multidimensional signal processing," *PhD Thesis, Linkoping University, Sweden*, Dissertation No 531, 1998.

[20] D. R. Hardoon, S. Szedmak, and J. S. Taylor, "Canonical correlation analysis: An overview with application learning," *Technical Report, Department of Computer Science, University of London*, CSD-TR-03-02, 2003.