JOINT AUDIO-VIDEO PROCESSING FOR BIOMETRIC SPEAKER IDENTIFICATION

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

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

ABSTRACT

In this paper we present a bimodal audio-visual speaker identification system. The objective is to improve the recognition performance over conventional unimodal schemes. The proposed system exploits not only the temporal and spatial correlations existing in speech and video signals of a speaker, but also the cross-correlation between these two modalities. Lip images extracted for each video frame are transformed onto an eigenspace. The obtained eigenlip coefficients are interpolated to match the rate of the speech signal and fused with mel frequency cepstral coefficients (MFCC) of the corresponding speech signal. The resulting joint feature vectors are used to train and test a Hidden Markov Model (HMM) based identification system. Experimental results are also included for demonstration of the system performance.

1. INTRODUCTION

Biometric person identification, in the most general case, refers to identification of a person from a set of candidates using her/his biometric data. Different biometric signals such as faces, voice, fingerprints, signature strokes, iris and retina scans can be used to perform this identification task. It is generally agreed that no single biometric technology will meet the needs of all potential identification applications. Although the performance of several of these biometric technologies have been studied individually, there is little work reported in the literature on the fusion of the results of various biometric identification technologies [1].

A particular problem in multi-modal biometric person identification, which has a wide variety of applications, is the speaker identification problem where basically two modalities exist: audio signal (voice) and video signal. Speaker identification, when performed over audio streams, is probably one of the most natural ways to perform person identification. However, video stream is also an important source of biometric information, in which we have still images of biometric features such as face and also the temporal motion information such as lip, which is correlated with the audio stream. Most speaker identification systems rely on audio-only data [2]. Even assuming ideal noiseless conditions, such systems are far from being perfect for high security applications. The same observation is also valid for systems using only visual data, where poor picture quality or changes in lighting conditions significantly degrade performance [3]. A better alternative is the use of both modalities in a single identification scheme.

This work has been supported by TUBITAK under the project EEEAG-101E038.

Lip movement is a natural by-product of the speaking act. Information inherent in lip movement has so far been exploited mostly for the speech recognition problem, establishing a one-to-one correspondence with the phonemes of speech and the visemes of lip movement. It is quite natural to assume that lip movement would also characterize an individual as well as what that individual is speaking. Only few articles in the literature incorporate lip information for the speaker identification problem [4, 5, 6]. Although these works demonstrate some improvement over unimodal techniques, they use a decision-fusion strategy and hence do not fully exploit the mutual dependency between lip movement and speech [4, 5]. A recent work addresses data-fusion between face and multiple speech features [7], but does not investigate the correlation between different modalities. In this paper we propose an HMM-based speaker identification scheme for joint use of the lip sequence and the audio signal of a speaking individual with data-fusion, i.e. early integration of audio and visual features [8]. In the joint feature vector, transform domain coefficients in an eigenspace of lip images constitute the visual part and MFCCs constitute the audio part.

The paper is organized as follows. Section 2 describes the joint audio-video processing scheme that we propose for robust speaker identification. The problem of extracting features from audio-only and video-only data, the methodology employed in fusion of these unimodal features as well as the description of our HMM-based classifier are all addressed in this section. Performance results of the proposed system is demonstrated in Section 3 and finally concluding remarks are given in Section 4.

2. BIMODAL SPEAKER IDENTIFICATION

In this study a text-dependent bimodal speaker identification system is considered. The bimodal database consists of audio and video signals belonging to individuals of a certain population. Each person in this database utters a predefined secret phrase that may vary from one person to another. The objective is, given the data of an unknown person, to find whether this person matches someone in the database or not. The system identifies the person if there is a match and rejects if not. Hence the problem can be thought of as an N-class recognition problem, including also a reject class. Such an identification system addresses problems of unauthorized use of computer and communication systems and multilevel access control. In the case of a secret phrase specific to each individual, the system provides two levels of security, thus seems to be more reliable. However, the proposed system should also be

robust against false identity claims. Our goal is to fully exploit the spatial and temporal correlations existing in a video stream and thereby to characterize the biometric properties of a speaker. We consider two different scenarios: in the first one each speaker utters her/his name whereas in the other each speaker utters the same 6-digit number.

Hidden Markov Models [9] are reliable structures to model human hearing system, and thus they are widely used for speech recognition and speaker identification problems [2, 9, 10]. The temporal characterization of an audio-video stream can also successfully be modeled using a Hidden Markov Model (HMM) structure, where state transitions model temporal correlations and Gaussian classifiers model signal characteristics. In this work a word-level continuous-density HMM structure is built for the speaker identification task using the HTK library [11]. Each speaker in the database population is modeled using a separate HMM and is represented with the feature sequence that is extracted over the audio-video stream while uttering the secret phrase. First a world HMM model is trained over the whole training data of the population. Then using the world HMM model as the initial state, each HMM associated to a speaker is trained over some repetitions of the audio-video utterance of the corresponding speaker. In the identification process, hypothesis testing is performed between the best match of the population and the world model for the given audio-video utterance of an unknown subject. The subject is either rejected or identified to be the speaker with the best match based on a likelihood ratio test.

2.1. Extraction of Visual Features

The eigenface technique [3], or more generally the principal component analysis, has proven itself as an effective and powerful tool for recognition of still faces. The core idea is to reduce the dimensionality of the problem by obtaining a smaller set of features than the original dataset of intensities. In principal component analysis, every image is expressed as a linear combination of some basis vectors, i.e. eigenimages that best describe the variation of intensities from their mean. When a given image is projected onto this lower dimensional eigenspace, a set of eigenimage coefficients is obtained, that gives a parametrization for the distribution of the signal.

In our case, each original dataset is a lip image obtained from video sequences of speaking individuals. We will represent each lip image by a set of *eigenlip* coefficients as referred in [12]. Obtaining principal components of the lip signal, i.e. eigenlips, can be thought of as an eigenvalue problem. Suppose that the training set consists of M mean-removed lip image vectors, $\boldsymbol{\ell}_0, \boldsymbol{\ell}_1, ..., \boldsymbol{\ell}_M$. Then the eigenlips $\mathbf{u}_i$ for $m = 0, 1, ..., M$, can be computed as the eigenvectors of the following covariance matrix $\mathbf{U}$,

$$\mathbf{U} = \frac{1}{M} \sum_{m=1}^{M} \boldsymbol{\ell}_m \boldsymbol{\ell}_m^T. \tag{1}$$

Each eigenlip $\mathbf{u}_m$ is associated to an eigenvalue λ_m. Principal components are the first p eigenlips that have $\lambda_m >> 0$. Usually the reduced dimension p is much smaller than M and the j-th eigenlip coefficient ω_j is obtained by the following projection:

$$\omega_j = \mathbf{u}_j^T \boldsymbol{\ell} \quad \text{for} \quad j = 0, 1, ..., p. \tag{2}$$

The eigenlip coefficients, when computed for every frame i of a given lip sequence, constitute the visual feature vector that will be denoted by $\mathbf{f}_v^i$:

$$\mathbf{f}_v^i = [\omega_1, \omega_2, \cdots, \omega_p]. \tag{3}$$

Beside being an efficient and powerful representation, the advantage of the eigenlip approach is that it works simply on intensity values. This improves the robustness of the overall scheme as compared to techniques that require more sophisticated methods such as lip tracking for extraction of some geometric features, e.g. lip contours as in [13]. With the eigenlip approach, it suffices to employ a simple lip detection process for extracting lip frames from face images. The disadvantage of this approach is that it is generally sensitive to rotation and lighting conditions, though small rigid motions of the head and small changes in illumination can be tolerated up to a certain measure.

2.2. Fusion of Audio and Visual Features

Mel frequency cepstral coefficients (MFCC) give good discrimination of speech data; hence they are widely used to represent audio streams in HMM-based speech recognition and speaker identification systems. MFFCs can be computed from the log filterbank magnitudes using the Discrete Cosine Transform (DCT). The MFCC vector c_k at time index k is defined as,

$$\mathbf{c}_k = \text{IDCT}(\{\mathbf{Y}_i | i = 0, 1, ..., I-1\}), \quad \text{and} \tag{4}$$

$$\mathbf{Y}_i = \sum_n \log |\mathbf{S}_k(n)| \mathbf{H}_i(n) \tag{5}$$

where $\mathbf{S}_k(n)$ is the n-th Fourier transform coefficient of the k-th speech frame and $\mathbf{H}_i$ is the i-th mel frequency window. The audio feature vector $\mathbf{f}_a^k$ for the k-th frame is formed as a collection of MFCCs, the first and the second delta MFCCs [2]:

$$\mathbf{f}_a^k = [\mathbf{c}_k \;\; \Delta\mathbf{c}_k \;\; \Delta\Delta\mathbf{c}_k]. \tag{6}$$

The proposed audio-visual fusion scheme is based on the early integration model [8] where the integration is performed in the feature space to form a composite feature vector of acoustic and visual features. Classification is implemented by using these composite vectors. This model makes the assumption of conditional dependence between acoustic and visual data. The acoustic features that are chosen to be MFCCs and the visual features that are obtained by the eigenlip approach, as explained in Section 2.1, are combined to form the joint audio-visual features. Thus we expect to better exploit the temporal correlation of audio-video streams for robust performance, especially in the presence of environmental noise. The structure of the fusion scheme is outlined in Fig. 1. The synchronous audio and video streams are processed separately. The lip areas are cropped from the video stream to form a stream of lip images at a rate of 15 fps using the hand labeled localization information. A subset of these lip images is used to create an eigenspace of dimension p. The eigenlip coefficients are computed by projecting every lip image of a given video stream onto the eigenspace.

As the audio features are extracted at a rate of 100 fps and the visual features are extracted at a rate of 15 fps, a rate synchronization should be performed prior to the data fusion. Let the audio and the visual features be represented at time instants $k\frac{1}{100}$ and

$i\frac{1}{15}$ seconds, respectively, i.e.,

$$\mathbf{f}_a^k = \mathbf{f}_a(k\frac{1}{100}) \quad \text{for} \quad k = 0, 1, 2, \ldots \tag{7}$$

$$\mathbf{f}_v^i = \mathbf{f}_v(i\frac{1}{15}) \quad \text{for} \quad i = 0, 1, 2, \ldots. \tag{8}$$

The visual features can be computed using linear interpolation over the $\mathbf{f}_v^i$ sequence to match the 100 fps rate,

$$\tilde{\mathbf{f}}_v^k = \tilde{\mathbf{f}}_v(k\frac{1}{100}) = (1-\alpha_k)\mathbf{f}_v^{i^*} + \alpha_k \mathbf{f}_v^{i^*+1}, \tag{9}$$

where $i^* = \lfloor \frac{3k}{20} \rfloor$ and $\alpha_k = \frac{3k}{20} - i^*$. Hence the joint audio-visual feature $\mathbf{f}_{av}^k$ is formed by combining the MFCCs, the first and second delta MFCCs and the interpolated visual features $\tilde{\mathbf{f}}_v^k$ for the k-th audio-visual frame:

$$\mathbf{f}_{av}^k = [\mathbf{f}_a^k \;\; \tilde{\mathbf{f}}_v^k]. \tag{10}$$

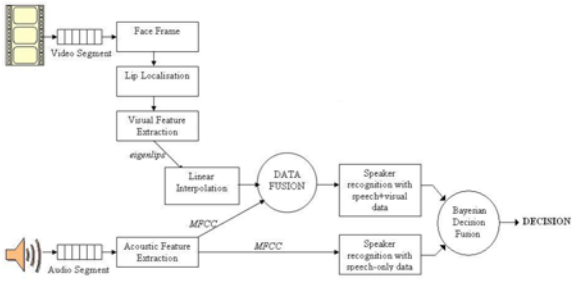

Fig. 1. Feature extraction and fusion flow.

As observed from Fig.1, the overall proposed scheme consists of two independent identification tasks performed with audio-only and fused audio-visual features. For the final decision, a Bayesian classifier is incorporated to combine the two decisions obtained in this way. Bayesian classifier uses likelihood ratios to measure the reliability of the two separate identification results.

3. EXPERIMENTAL RESULTS

The database that have been used to test the performance of the proposed speaker identification system is briefly described in Table 1, where the distributions of the test and training data collected disjointly for two different scenarios are given. The audio-visual data have been acquired using Sony DSR-PD150P video camera at Multimedia Vision and Graphics Laboratory of Koç University. Samples from this database are displayed in Fig. 2. In addition to the database displayed in Table 1 a set of impostor data is collected for the first scenario. In the collection of impostor data each subject utters five different names from the population. However in the second scenario all the utterances not belonging to the subject are used as impostor data.

The presented experimental results have all been obtained using the HTK tool version 3.0, each speaker being represented by a 6-state left-to-right HMM structure. The acquired video data is first split into segments of secret phrase utterances. The visual and the audio streams are then separated into two parallel streams, where the visual stream has gray-level video frames of size 720×576 pixels containing the frontal view of a speaker's

Scenario	Subjects		Repetitions		Total
	male	female	train	test	
Subject's name	31	7	10	10	760
6-digit number	31	7	4	6	380

Table 1. The distribution of the data collected for the two different scenarios.

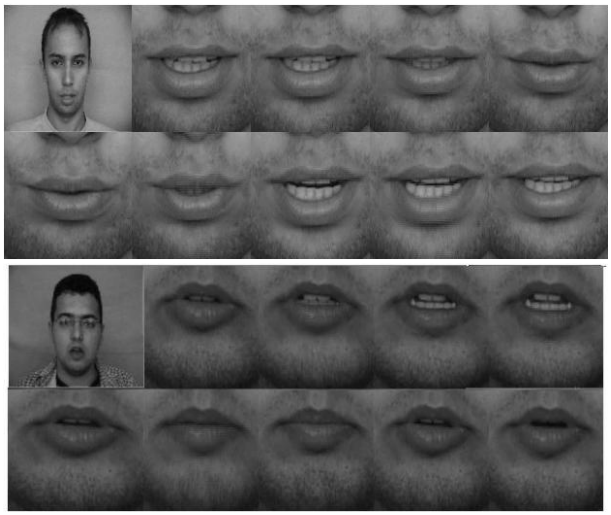

Fig. 2. Sample images from the acquired audio-visual database.

head at a rate of 15 fps and the audio stream has 16 kHz sampling rate. The acoustic noise, which is added to the speech signal to observe the identification performance under adverse conditions, is picked to be a mixture of office and babble noise.

The audio stream is processed over 10 msec frames centered on 25 msec Hamming window. The MFCC feature vector, $\boldsymbol{c}_k$, is formed from 13 cepstral coefficients including the 0th gain coefficient using 26 mel frequency bins. The resulting audio feature vector, $\boldsymbol{f}_a^k$ of size 39, includes the MFCC vector and the first and the second delta MFCC vectors.

Each video stream is at most 1 second in duration and during this time it is assumed that the subject does not considerably move her/his head. Hence, hand labeled lip regions can be used to crop 120×128 lip frames to form the lip sequence for each visual stream. An eigenspace of dimension p is computed using the training part of the lip sequence set. The eigenspace dimension is set to be $p = 20$, as this value is observed to be sufficient for the desired performance of the identification system. The system's performance starts to degrade as p is further decreased. The visual feature vectors $\boldsymbol{f}_v^i$, which are used in both training and testing of the HMM-based classifier, are obtained by projecting each lip image of the database onto this eigenspace and thereby computing the eigenlip coefficients.

The identification results are shown in Table 2, where we observe the equal error rates at varying levels of acoustic noise for the two scenarios: the secret phrases are the person's name and the fixed 6-digit number (348572), respectively. The unimodal (video-only and audio-only) and bimodal (audio-visual with and without Bayesian classifier) equal error rates are displayed on the

	EER (%)			
Noise Level	Audio Only	Video Only	Audio Visual	Bayesian Decision
	Scenario 1: Subject's name			
clean	2.99	20.21	3.17	2.58
15 dB	4.28	20.21	4.12	3.67
10 dB	8.86	20.21	4.91	4.10
5 dB	18.32	20.21	6.56	5.7
0 dB	34.01	20.21	10.87	8.5
	Scenario 2: Fixed 6-digit number			
clean	1.61	5.55	2.11	1.41
21 dB	4.31	5.55	2.14	1.61
18 dB	8.82	5.55	2.18	1.75
15 dB	24.06	5.55	2.22	1.90
12 dB	44.04	5.55	2.77	2.40

Table 2. Speaker identification results: equal error rates at varying noise levels.

same table to better observe the improvement obtained by audio-visual data fusion. In the audio-only case the identification performance degrades rapidly with decreasing SNR. This degradation is stronger for the fixed 6-digit number scenario in which identification task is expected to be harder since all impostor speakers utter the same 6-digit number. In the video-only case the identification performance is observed to be poorer for the first scenario. This is mainly due to the sensitivity of this scenario to the impostor data since in this case the HMM structure models not only the personal biometric lip movements but also the lip movements corresponding to the speech content.

The overall performance is improved with the incorporation of visual information, though this improvement is lower at high SNR levels. The bimodal identification performance is significantly higher than the audio-only and the video-only cases at lower SNR levels. Thus the bimodal system seems less sensitive to noise level. The identification performance of the bimodal system is further improved with the incorporation of the Bayesian classifier; this improvement does not seem to be significant, however the Bayesian approach guarantees the overall performance to remain at least as good as the audio-only performance.

4. CONCLUSIONS

We have presented a bimodal (audio-visual) speaker identification system that improves the recognition performance over unimodal schemes. The data fusion of audio and video information, to train a HMM-based classifier, has availed us with the possibility of fully exploiting the correlations existing between two modalities. Hence the proposed technique, as verified with the experiments, seems to be more reliable for security systems as compared to unimodal approaches. Furthermore, the use of eigenlip coefficients computed directly from intensity values appears to be a promising attempt in achieving a robust and practical speaker identification system. Such an eigenlip representation avoids inevitable robustness problems of the systems relying rather on geometric features that require sophisticated and mostly unreliable image analysis tasks, such as segmentation and lip tracking.

There are problems and further issues to be addressed. First, currently lip frames are extracted from video stream by hand labeling; this process should be automatized by a simple but effective lip detection method using spatial and motion information. Second, the training and test database should be enriched both in terms of total population and variety for a more reliable performance analysis. The variety in database refers mainly to changing environmental conditions such as lighting and background, and to including video sequences where the head of the speaker may undergo arbitrary rigid motion. This would allow us to better measure the tolerance of our system to head rotation and changing illumination. In this respect, methodologies that would enforce the overall scheme for invariance to such properties has to be explored. All these issues and problems are currently under investigation.

5. REFERENCES

[1] N.K. Ratha, A. Senior, and R.M.Bolle, "Automated biometrics," *ICAPR*, pp. 445–474, May 1997.

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

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

[4] 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.

[5] P. Jourlin, J. Luettin, D. Genoud, and H. Wassner, "Acoustic-labial speaker verification," *Pattern Recognition Letters*, vol. 18, no. 9, pp. 853–858, 1997.

[6] T. Wark, S. Sridharan, and V. Chandran, "The use of temporal speech and lip information for multi-modal speaker identification via multi-stream HMM's," *Proc. of the Int. Conf. on Acoustics, Speech and Signal Processing 2000 (ICASSP 2000)*, pp. 2389–2392, 2000.

[7] C. Sanderson and K. K. Paliwal, "Noise compensation in a person verification system using face and multiple speech features," *Pattern Recognition*, vol. 36, no. 2, pp. 293–302, February 2003.

[8] D. D. Zhang, *Automated Biometrics*, Kluwer Academic Publishers, 2000.

[9] L.R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, February 1989.

[10] J. Luettin and S Dupont, "Continous audio-visual speech recognition," *Technical Report IDIAP*, 1997.

[11] S. Young, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, "The HTK-hidden markov model toolkit v2.1," *Entropic Research,Cambridge*, 1997.

[12] C. Bregler and Y. Konig, "Eigenlips for robust speech recognition," *Proc. of IEEE Conf. on Acoustics, Speech and Signal Processing*, pp. 669–672, 1994.

[13] T.J. Wark, S. Sridharan, and V. Chandran, "An approach to statistical lip modeling for speaker identification via chromatic feature extraction," *Int. Conf. on Pattern Recognition*, vol. 1, pp. 123–125, 1998.