# MULTIMODAL SPEAKER IDENTIFICATION WITH AUDIO-VIDEO PROCESSING

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

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

## ABSTRACT

In this paper we present a multimodal audio-visual speaker identi£cation system. The objective is to improve the recognition performance over conventional unimodal schemes. The proposed system decomposes the information existing in a video stream into three components: speech, face texture and lip motion. Lip motion between successive frames is £rst computed in terms of optical ¤ow vectors and then encoded as a feature vector in a magnitude-direction histogram domain. The feature vectors obtained along the whole stream are then interpolated to match the rate of the speech signal and fused with mel frequency cepstral coef£cients (MFCC) of the corresponding speech signal. The resulting joint feature vectors are used to train and test a Hidden Markov Model (HMM) based identi£cation system. Face texture images are treated separately in eigenface domain and integrated to the system through decision-fusion. Experimental results are also included for demonstration of the system performance.

## 1. INTRODUCTION

Biometric person identi£cation, in the most general case, refers to identi£cation of a person from a set of candidates using her/his biometric data. Different biometric signals such as face, voice, £ngerprints, signature strokes, iris and retina scans can be used to perform this identi£cation task. It is generally agreed that no single biometric technology will meet the needs of all potential identi£cation 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 identi£cation technologies [1].

A particular problem in multimodal biometric person identi£cation, which has a wide variety of applications, is the speaker identi£cation problem where basically two sources of information exist: audio signal (voice) and video signal. Speaker identi£cation, when performed over audio streams, is probably one of the most natural ways to perform person identi£cation. 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 movement, which is correlated with the audio stream. Most speaker identi£cation 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 signi£cantly degrade performance [3]. A better alternative is the use of all available modalities, i.e. audio, motion and texture, in a single identi£cation scheme.

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 the individual is speaking. Only few articles in the literature incorporate lip information for the speaker identi£cation problem [4, 5]. 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, 6]. In this paper we propose an HMM-based speaker identi£cation scheme for joint use of the lip movement and the audio signal of a speaking individual with data-fusion, i.e. early integration of audio and visual features [7]. In the joint feature vector, optical ¤ow vectors tranformed into a magnitude-direction histogram domain constitute the visual motion part and MFCCs constitute the audio part. Visual texture information, i.e. face images, expressed in eigenface domain is integrated to the system through decision-fusion.

## 2. MULTIMODAL SPEAKER IDENTIFICATION

In this study a text-dependent multimodal speaker identi£cation system is considered. The database consists of audio and video signals belonging to individuals of a certain population. Each person in this database utters a prede£ned secret phrase that varies from one person to another. The objective is, given the data of an unknown person, to £nd whether this person matches someone in the database or not. The system identi£es the person if there is a match and rejects if not. 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.

### 2.1. Face and Lip Detection

The £rst step in extracting visual features is to detect face and lip regions. We assume that the acquired images contain the face of a speaking person with a stationary background. A possibility here would be using a simple change detection algorithm. Such simple algorithms are computationally attractive; however they are usually very sensitive to noise, changing light and possible small camera movements. Thus we propose an optical ¤ow based detection

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

technique that gives more accurate and reliable results. Optical □ow vectors are £rst computed between successive frames of the video sequence [8]. The magnitudes of these vectors are accumulated in a buffer and then thresholded. The rectangular region enclosing the pixels survived after thresholding gives the face frame. Once face is detected, then in this region we search for the lip, assuming that the lip constitutes the largest portion of the face that dominates the overall movement. A second thresholding of optical □ow vector magnitudes in the detected face region, followed by morphological processing to £ll up small holes and eliminating small isolated regions provides us with moving, possibly separate portions of the lip. Around the center of gravity of these partial lip regions, we construct a £xed size window frame that we label as the lip region. In Fig.1, we demonstrate the performance of our detection method on a video sequence from our current database.



**Fig. 1**. Face and lip detection: The £rst image of the video sequence displayed with detected face and lip regions.

### 2.2. Extraction of Lip Motion Features

Optical □ow vectors computed for each lip frame as described in Section 2.1, provide the initial information for characterizing lip movement. This information has to be transformed into another domain so as to reduce the dimensionality. This process should also take into account the problem of invariance with respect to rotation, translation and scale. In order to achieve this, we transform the optical □ow vectors into a magnitude-direction histogram domain.

Let an optical □ow vector be denoted by $\mathbf{v}_j = (v_j^m, v_j^d)$, where $v_j^m$ and $v_j^d$ stand for the magnitude and the direction angle of the vector. For each lip frame, we perform the following basic tasks: A reference direction $\tilde{\theta}$ is £rst computed as the magnitude-weighted average of the optical □ow vector directions:

$$\tilde{\theta} = \frac{1}{M} \sum_j v_j^m v_j^d, \tag{1}$$

where $M = \sum_j v_j^m$. The reference direction $\tilde{\theta}$ can be considered as the average direction of the lip movement. The whole range of $2\pi$ degrees for direction angle is divided into $p$ equal sectors each denoted by $\Theta_k$, $k = 1, ..., p$. The direction $v_j^d$ of each optical □ow vector is re-adjusted with respect to the reference direction $\tilde{\theta}$ so as to give $\bar{v}_j^d = v_j^d - \tilde{\theta}$ and the magnitude $v_j^m$ is accumulated into the corresponding angular sector. The normalized $k$th coef£cient $w_k, k = 1, ..., p$, of the lip motion feature vector is then given by

$$w_k = \frac{1}{M} \sum_{\Theta_k} v_j^m \tag{2}$$

where the summation over $\Theta_k$ corresponds to accumulation of the magnitudes of □ow vectors with $k\frac{2\pi}{p} \leq \bar{v}_j^d < (k+1)\frac{2\pi}{p}$. The coef£cients $w_k^i$ constitute the lip motion feature vector for frame $i$, that will be denoted by $\mathbf{f}_m^i$:

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

As compared to eigenlip techniques that are often used in the literature to fuse lip information to identi£cation schemes [9], the described optical □ow based technique is a simple but robust method that does not require a very accurate localization of the lip contour. Note that the resulting feature vector is translation and scale invariant. As for the rotation, the representation is invariant to rotation along one axis of the face whereas small rotations along the other two axes can be tolerated up to a certain measure.

### 2.3. Fusion of Audio and Lip Motion Features

Mel frequency cepstral coef£cients (MFCC) give good discrimination of speech data; hence they are widely used to represent audio streams in HMM-based speech recognition and speaker identi£cation systems. The audio feature vector $\mathbf{f}_a^k$ for the $k$-th frame is formed as a collection of MFCCs denoted by $c_k$ along with the £rst and the second delta MFCCs [2]:

$$\mathbf{f}_a^k = [\mathbf{c}_k \ \Delta\mathbf{c}_k \ \Delta\Delta\mathbf{c}_k]. \tag{4}$$

The proposed audio-motion fusion scheme is based on the early integration model [7] where the integration is performed in the feature space to form a composite feature vector of acoustic and lip motion features. Classi£cation is implemented by using these composite vectors. The acoustic features that are chosen to be MFCCs and the motion features that are obtained by the optical □ow based technique, as explained in Section 2.2, are combined to form the joint audio-motion 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. 2.

As the audio features are extracted at a rate of 100 fps and the lip motion 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 motion features be represented at time instants $k\frac{1}{100}$ and $i\frac{1}{15}$ seconds, respectively, i.e.,

$$\mathbf{f}_a^k = \mathbf{f}_a(k\frac{1}{100}) \quad \text{for} \quad k = 0, 1, 2, \ldots \tag{5}$$

$$\mathbf{f}_m^i = \mathbf{f}_m(i\frac{1}{15}) \quad \text{for} \quad i = 0, 1, 2, \ldots. \tag{6}$$

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

$$\tilde{\mathbf{f}}_m^k = \tilde{\mathbf{f}}_m(k\frac{1}{100}) = (1-\alpha_k)\mathbf{f}_m^{i^*} + \alpha_k \mathbf{f}_m^{i^*+1}, \tag{7}$$

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

$$\mathbf{f}_{am}^k = [\mathbf{f}_a^k \ \tilde{\mathbf{f}}_m^k]. \tag{8}$$

### 2.4. HMM-based Recognition

Hidden Markov Models [10] are reliable structures to model human hearing system, and thus they are widely used for speech recognition and speaker identi£cation problems [2, 10, 11]. The temporal characterization of an audio-video stream can also successfully be modeled using an HMM structure, where state transitions model temporal correlations and Gaussian classi£ers model signal characteristics. In this work a word-level continuous-density HMM structure is built for the speaker identi£cation task using the HTK library [12]. 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 identi£cation 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 identi£ed to be the speaker with the best match based on a likelihood ratio test.

### 2.5. Fusion of Face Texture

The eigenface technique [3] 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. 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 $r$ eigenface coef£cients is obtained, that gives a parametrization for the distribution of the signal.

The eigenface coef£cients, when computed for every frame $i$ of a given sequence, constitute the face texture feature vector that we will denote by $\mathbf{f}_t^i$:

$$\mathbf{f}_t^i = [\omega_1, \omega_2, \cdots, \omega_r]. \tag{9}$$

The face images in the training set are used to obtain £rst the eigenspace and thereby an average feature vector to represent the world class of faces, that we will denote by $\tilde{\mathbf{f}}_t$. Then the likelihood ratio to be used in hypothesis testing of a face image $i$ with an image $j$ in the database is given by $\|\mathbf{f}_t^i - \mathbf{f}_t^j\| / \|\mathbf{f}_t^i - \tilde{\mathbf{f}}_t\|$. Note that a video sequence of a person contains a number of face images. Thus the best match of a test face sequence is determined by using a majority rule. The likelihood ratio of this best match provides us with a con£dence score that can be exploited during the fusion process.

As observed from Fig.2, the proposed overall scheme consists of two independent identi£cation tasks performed with audio-only and fused audio-motion-texture features. For the £nal decision, a Bayesian classi£er is incorporated to combine the two decisions obtained in this way. Bayesian classi£er uses likelihood ratios to measure the reliability of the two separate identi£cation results.

The fusion of audio and motion features is basically a data fusion process and results in a likelihood ratio $\rho_{am}$ obtained from the best match. This ratio can then be combined with the likelihod ratio $\rho_t$ provided separately by the face identi£cation process. The overall score of audio-motion-texture fusion is obtained by the weighted average of the two individual likelihood ratios, $G\rho_{am} + (1-G)\rho_t$, where the weight $G$, $0 \leq G \leq 1$, determines the contribution of each modality. Note that for $G = 0$, the overall system turns out to be an audio-texture identi£cation scheme.

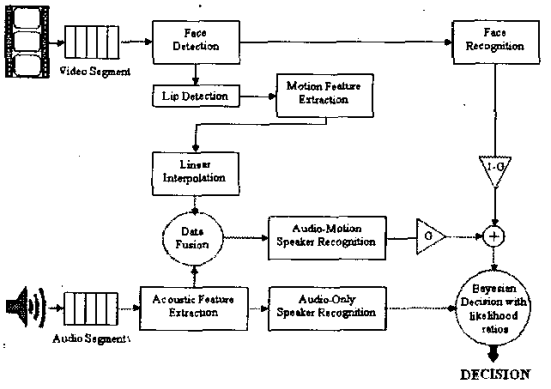



**Fig. 2.** Multimodal speaker identi£cation scheme.

## 3. EXPERIMENTAL RESULTS

The database used to test the performance of the proposed speaker identi£cation system includes 50 subjects where 8 of them are females. Training and testing are performed over two independent datasets each with £ve repetitions. A set of impostor data is also collected with each subject uttering £ve different names from the population. The audio-visual data have been acquired using a Sony DSR-PD150P video camera at Multimedia Vision and Graphics Laboratory of Koç University.

| Source Modality | EER (%) Noise Level (dB SNR) clean | 25 | 20 | 15 | 10 | 5 | 0 |
|---|---|---|---|---|---|---|---|
| Audio | 2.4 | 4.3 | 6.4 | 20.4 | 31.1 | 43.1 | 51.5 |
| Audio-Flow | 4.3 | 5.5 | 6.3 | 15.5 | 20.5 | 30.1 | 36.4 |
| Eigenface | 35.3 | | | | | | |
| | Bayesian Decision Fusion | | | | | | |
| G = 1.00 | 1.8 | 3.5 | 4.5 | 14.3 | 19.6 | 29.4 | 36.1 |
| G = 0.75 | 1.6 | 3.2 | 4.2 | 12.3 | 16.9 | 24.3 | 29.8 |
| G = 0.50 | 1.6 | 3.3 | 4.2 | 14.1 | 19.6 | 28.2 | 31.8 |
| G = 0.25 | 1.8 | 3.5 | 5.3 | 17.1 | 25.1 | 32.0 | 33.9 |
| G = 0.00 | 2.0 | 3.9 | 5.4 | 18.2 | 26.7 | 34.9 | 35.3 |

**Table 1.** Speaker identi£cation results: Equal error rates at varying noise levels.

The temporal characterization of the audio and the audio fused with the lip optical ¤ow have been obtained using the HTK tool version 3.0, where each speaker is represented by a 6-state left-to-right HMM structure. The acquired video data is £rst split into segments of secret phrase utterances. The visual and 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 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 identi£cation perfor-

mance 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, $c_k$, is formed from 13 cepstral coefficients including the 0th gain coefficient using 26 mel frequency bins. The resulting audio feature vector, $f_a^k$ of size 39, includes the MFCC vector along with the first and the second delta MFCC vectors.

Each video stream is at most 1 second in duration and results in 15 individual face and lip frames of sizes $370 \times 460$ and $120 \times 128$, respectively. The motion feature vectors $f_m^i$, which are used in both training and testing of the HMM-based classifier, are obtained as described in Section 2.2 with $p = 20$. As for the extraction of face feature vectors, an eigenspace of dimension $r = 20$ is computed using 5 pictures from each video sequence of the training set.

The identification results are shown in Table 1, where we observe the equal error rates at varying levels of acoustic noise. The first two rows display the equal error rates obtained for audio-only and audio fused with lip motion (audio-□ow). The third row presents the equal error rate for the texture-only identification system that is based on the eigenface method. Finally, the last five rows display the equal error rates obtained after the Bayesian decision fusion of the audio-only and the audio-motion-texture identification results, with varying values of $G$. The best equal error rate results are obtained when $G$ is 0.75, that is when audio-motion and texture-only schemes have 75% and 25% contributions to the decision fusion of likelihood ratios, respectively.

For the texture-only case, we have to point out that all the face images used for training have the same background whereas the background of the test images varies; this is why the texture-only identification performance may seem to be worse than expected, as observed in Table 1. In the audio-only case the identification performance degrades rapidly with decreasing SNR. However, when lip motion is fused with audio, the identification performance improves significantly at these low SNR values, due to the correlation existing between lip movement and speech.

The overall performance is further improved using the Bayesian decision fusion, especially at high noise levels. Thus the multimodal system seems less sensitive to noise level and the incorporation of the Bayesian classifier guarantees the overall performance to remain at least as good as the audio-only performance.

## 4. CONCLUSIONS

We have presented a multimodal audio-visual speaker identification system that improves the recognition performance over unimodal schemes. The data fusion of audio and lip motion information, to train a HMM-based classifier, has availed us with the possibility of fully exploiting the correlations existing between two modalities. The texture information decoupled from the video stream is incorporated via a score-based decision fusion mechanism to further improve the performance.

The optical □ow based technique used for characterizing lip motion appears to be a promising attempt in achieving a robust and practical multimodal speaker identification system. Such a 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. As compared to eigenlip techniques used commonly for the fusion of lip motion information, our optical □ow based method has attractive invariance properties that increases generality and robustness of the identification process.

There are problems and further issues to be addressed. First, 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 better invariance to such properties has to be explored. Secondly, the decision fusion mechanism can be improved, noting that there are many other ways of combining information coming separately from audio, motion and texture parts of the video sequence of a speaking person. 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] 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.

[6] 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.

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

[8] B. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," *Proc. of 7th International Joint Conference on Artificial Intelligence*, pp. 674–679, 1981.

[9] A. Kanak, E. Erzin, Y. Yemez, and A. M. Tekalp, "Joint audio-video processing for biometric speaker identification," *Proc. of the Int. Conf. on Acoustics, Speech and Signal Processing 2003 (ICASSP 2003)*, vol. II, pp. 377–380, 2003.

[10] 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.

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

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