DISCRIMINATIVE LIP-MOTION FEATURES FOR BIOMETRIC SPEAKER IDENTIFICATION

H. E. Çetingül, Y. Yemez, E. Erzin and A. M. Tekalp

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

ABSTRACT

This paper addresses the selection of best lip motion features for biometric open-set speaker identification. The best features are those that result in the highest discrimination of individual speakers in a population. We first detect the face region in each video frame. The lip region for each frame is then segmented following registration of successive face regions by global motion compensation. The initial lip feature vector is composed of the 2D-DCT coefficients of the optical flow vectors within the lip region at each frame. The discriminant analysis is composed of two stages. At the first stage, the most discriminative features are selected from the full set of DCT coefficients of a single lip motion frame by using a probabilistic measure that maximizes the ratio of intra-class and inter-class probabilities. At the second stage, the resulting discriminative feature vectors are interpolated and concatenated for each time instant within a neighborhood, and further analyzed by LDA to reduce dimension, this time taking into account temporal discrimination information. Experimental results of the HMM-based speaker identification system are included to demonstrate the performance.

1. INTRODUCTION

It has been a common practice to use lip-motion for speech recognition applications [1, 2]. This is justified by the observation that lip movement is highly correlated with the audio signal, and the speech content can be revealed through lip-reading. As far as speech is concerned, it is usually sufficient to extract principal components of the lip movement; there are various techniques in the literature to implement this approach. One possibility is the use of eigenlips from lip intensity or lip contour shape, that eventually corresponds to a principal component analysis (PCA). Another possibility is the use of low frequency 2D-DCT coefficients of lip-motion vectors.

It is quite natural to assume that lip movement would also characterize the identity of an individual as well as what the individual is speaking. However, for the speaker identification problem, the principal components of the lip movement are not usually sufficient to well discriminate the biometric properties of a speaker; principal component analysis is actually based on the criterion to minimize the mean-square error of reconstruction of a signal. In a recognition problem, the criterion should rather be to minimize the recognition error, and in this sense high frequency or non-principal components of a signal should also be valuable especially when the objective is to model the biometrics, i.e. specific lip movements of an individual rather than what is uttered. A possibility is the use of the linear discriminant analysis (LDA) so as to map the high dimensional feature vector to a subspace of reduced dimension that best describes the discrimination among classes. LDA has a major drawback due to the required matrix inversion operation: Prior to the LDA analysis, the dimension of the original feature vector has to significantly be reduced, using for instance PCA analysis, in order to lessen the computational complexity of the inversion process or due to the limited number of available training samples.

Only few articles in the literature incorporate lip information for the speaker identification problem [3, 4]. In [4], the full set of optical flow DCT coefficients are used for identification with no discrimination analysis whereas in [3] the lip contour shape is modeled by PCA, followed by linear discriminant analysis. In this work, we propose a two-stage discriminative lip-motion feature extraction technique that can be used in multimodal speaker identification systems. The proposed technique, as described in Section 2, takes into account the temporal discrimination information as well as the intra-class and inter-class distribution of individual single-frame feature vectors. Open-set speaker identification problem is briefly described in Section 3 and then the overall system is discussed in detail in Section 4. The experimental results are presented in Section 5 and finally in Section 6, conclusions and comments on future work are provided.

2. DISCRIMINATION ANALYSIS

There exist a number of subspace representation techniques that can be used as a solution to the dimensionality problem of recognition systems. Linear discriminant analysis (LDA) is a well-known dimension reduction and data analysis method to achieve discrimination among classes and has proven its success in speech recognition [2]. Another possibility, as we propose, is to achieve discrimination in the Bayesian sense using a probabilistic measure that maximizes the ratio of intra-class and inter-class probabilities. The overall discrimination analysis that we propose in this paper employs both of these techniques that are discussed in detail in the subsections 2.1 and 2.2.

This work has been supported by TUBITAK under the project EEEAG-101E038 and by the European FP6 Network of Excellence SIMILAR (http://www.similar.cc).

2.1. Bayesian Discriminative Feature Selection

The identification problem is often formalized within a probabilistic framework. The decision rule selects the class λ_i given an observation $\boldsymbol{f}_k$ with maximum posterior probability $P(\lambda_i|\boldsymbol{f}_k)$. The posterior probability can be written in terms of class conditional probability distributions:

$$\begin{aligned} P(\lambda_i|\boldsymbol{f}_k) &= \frac{P(\boldsymbol{f}_k|\lambda_i)P(\lambda_i)}{P(\boldsymbol{f}_k)} \\ &= \frac{P(\boldsymbol{f}_k|\lambda_i)P(\lambda_i)}{P(\boldsymbol{f}_k|\lambda_i)P(\lambda_i)+\sum_{j\neq i}P(\boldsymbol{f}_k|\lambda_j)P(\lambda_j)} \\ &= \left[1+\frac{\sum_{j\neq i}P(\boldsymbol{f}_k|\lambda_j)P(\lambda_j)}{P(\boldsymbol{f}_k|\lambda_i)P(\lambda_i)}\right]^{-1} \end{aligned} \quad (1)$$

The maximum likelihood estimator, that maximize the class conditional probability $P(\boldsymbol{f}_k|\lambda_i)$, becomes maximum mutual information estimator (MMIE) [5] by maximizing the likelihood ratio $\boldsymbol{l}(\lambda_i|\boldsymbol{f}_k)$,

$$\boldsymbol{l}(\lambda_i|\boldsymbol{f}_k) = \frac{P(\boldsymbol{f}_k|\lambda_i)P(\lambda_i)}{\sum_{j\neq i}P(\boldsymbol{f}_k|\lambda_j)P(\lambda_j)} \quad (2)$$

where this ratio can clearly serve as a measure of discrimination between the class λ_i and all other classes for the corresponding observation $\boldsymbol{f}_k$.

When the posterior probability distributions are available for some K independent observations $\{\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_K\}$, one can compute the discriminative power of the independent observation $\boldsymbol{f}_k^i$ that belongs to class λ_i using $\boldsymbol{l}(\lambda_i|\boldsymbol{f}_k^i)$. When the ratio $\boldsymbol{l}(\lambda_i|\boldsymbol{f}_k^i)$ is larger for an observation, it is clear that the observation is discriminative, that is the class conditional probability for its own class is high and the average of the class conditional probabilities for all other classes are small. Here, we can conveniently assume that the class probabilities, $P(\lambda_i)$, are equally likely. The posterior probability distributions are generally computed over some training data using expectation-maximization type of algorithms and assuming some underlying probability distribution. Let us refer this training data as $\boldsymbol{f}_k^i(n)$, that is a collection of the k-th independent observation from the i-th class, which are available for all independent observations and for all classes. We propose to represent the discriminative power $\boldsymbol{d}(\boldsymbol{f}_k)$ of the observation $\boldsymbol{f}_k$ as,

$$\boldsymbol{d}(\boldsymbol{f}_k) = \sum_i \frac{1}{M}\sum_{n=0}^{M-1}\boldsymbol{l}(\lambda_i|\boldsymbol{f}_k^i(n)) \quad (3)$$

where M is the number of the k-th independent observations in each class λ_i. The discriminative power of each independent observation creates an ordering between different observations. Hence, this ordering could be used to select the most discriminative features, or similarly to filter out the least discriminative features from the list of observations.

Under Gaussian distribution assumption, if we de-correlate the observation vector, each observation coefficient becomes independent from the rest of the observation coefficients. One can use DCT or PCA to de-correlate the observation vectors. After the de-correlation transformation, traditionally the low indexed coefficients (*FirstN*) are used as the representative features as they yield the best reconstruction for the original observations. However, a discriminative set of features (*DiscriminativeN*) can be selected using the discriminative power ordering. The posterior probability distributions of each transform domain coefficient can be modeled by Gaussian mixture densities (GMM), so that the discriminative power of each coefficient can be calculated as in Eq. 3. Then, the discriminative power ordering is performed, and the ordering of the first N discriminating coefficients is fixed and these coefficients are selected as the *DiscriminativeN* features.

2.2. Temporal Discriminative Feature Selection

The Bayesian discriminative feature selection technique described above takes into account the intra-class and inter-class distribution of individual single-frame feature vectors. However, a proper lip feature discrimination analysis should also exploit the temporal correlations existing between successive lip frames specific to a speaker class. One possibility here is to successively concatenate the resulting lip feature vectors so as to create a new sequence of higher dimensional feature vectors, each centered at the current frame instant. In that case, these higher dimensional feature vectors become subject to further analysis to discriminatively reduce dimension.

Potamianos et al. [2] used LDA to reduce the dimension of the feature vector composed of low-indexed intensity-based DCT coefficients resulting from lip images. They concatenated a number of consecutive feature vectors centered at the current frame so as to capture dynamic speech information. A similar step is performed in our work to exploit temporal correlations for the speaker identification problem as described in detail in Section 4.

3. OPEN-SET SPEAKER IDENTIFICATION

In the open-set speaker identification problem, the decision is taken by the maximization of a posteriori probability $P(\lambda_i|\boldsymbol{f})$ over all possible classes λ_i, i.e. for each speaker's model. The feature vector $\boldsymbol{f}$ is then assigned to the class λ^* that maximizes the a priori probability:

$$\lambda^* = \arg\max_{\lambda_i} P(\boldsymbol{f}|\lambda_i) \quad (4)$$

In an open-set speaker identification scheme, a reject mechanism is also required due to possible impostor identity claims. A possible reject strategy is thus to refer a reject (imposter) class $\lambda_{\bar{i}}$, so that a likelihood ratio $\rho(\boldsymbol{f}|\lambda_i)$ in log domain is used for the accept or reject decision:

$$\rho(\boldsymbol{f}|\lambda_i) = \log\frac{P(\boldsymbol{f}|\lambda_i)}{P(\boldsymbol{f}|\lambda_{\bar{i}})} = \log P(\boldsymbol{f}|\lambda_i) - \log P(\boldsymbol{f}|\lambda_{\bar{i}}) \quad (5)$$

Ideally, the imposter class model should be constructed by using all possible imposter observations for class λ_i, which is practically infeasible to achieve. In this paper we use the universal background model which is estimated by using all available training data regardless of which class they belong to. The final decision strategy can be stated as follows:

$$\begin{array}{ll} \text{if } \rho(\boldsymbol{f}|\lambda^*) \geq \tau & \text{accept} \\ \text{otherwise} & \text{reject} \end{array} \quad (6)$$

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

4. SPEAKER IDENTIFICATION SYSTEM

Biometric speaker identification experiments are conducted using the audio-visual database MVGL-AVD [6]. 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 collected with each subject in the population uttering five different names from the population.

Before the lip-motion feature extraction, each face image frame is aligned 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 [7]. Then the face images are warped according to these calculated parameters. After this alignment, the optical flow vectors from the lip frames are extracted using hierarchical Lucas-Kanade technique. The lip-motion vectors on x and y directions are separately transformed into DCT domain and the first 500 DCT coefficients of the zig-zag scan both on x and y directions are combined to form a feature vector $\boldsymbol{F}$ of dimension 1000 as presented in Figure 1.

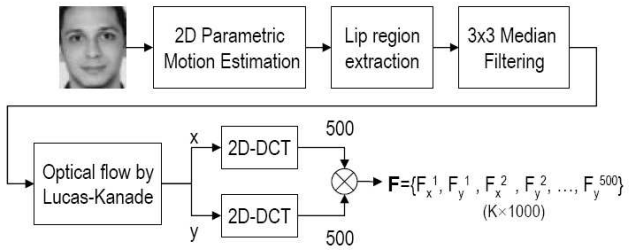

Fig. 1. Feature extraction through DCT

Figure 2 shows the next phase of the feature selection system. There are two reference feature sets in Figure 2 for comparison purposes. The first N coefficients of $\boldsymbol{F}$ excluding dc-terms (*FirstN*) form the first reference feature set representation $\boldsymbol{F_A}$ for the lip-motion modality. A second reference feature set, $\boldsymbol{F_B}$, is formed by discriminative feature selection as described in section 2.1. The posterior probability distributions of the $\boldsymbol{F}$ vector are modeled using 3 mixture GMM structures with diagonal covariance vectors for each speaker. Then, the discriminative power ordering is performed in the training data, and the ordering of the first N discriminating coefficients is fixed and these coefficients are selected as the *DiscriminativeN* ($\boldsymbol{F_B}$) features.

Later, as outlined in Figure 3, LDA is applied to a window of features to capture the temporal correlations between the observations as in [2]. The observations are linearly interpolated by 4 prior to concatenation of temporal windows. In the interpolated temporal domain each feature set at time instant k, together with the previous and next six sets of features, are concatenated, and LDA is performed on this concatenated vector of dimension $13N$ to reduce the feature vector dimension to 49 for 50 classes. The LDA modeling is applied to both of the feature sets, *FirstN* and *DiscriminativeN*. Correspondingly, the resulting feature sets are represented by $\boldsymbol{F_C}$ and $\boldsymbol{F_D}$ after the LDA analysis of the features $\boldsymbol{F_A}$ and $\boldsymbol{F_B}$.

The temporal characterizations of the lip-motion modality is performed using Hidden Markov Models (HMM). Word-level continuous-density HMM structures are built for the speaker identification task. Each speaker in the database population is modeled using a separate HMM and is represented with the feature sequence that is extracted over the lip stream while uttering the secret phrase. First a world HMM model is trained over the whole

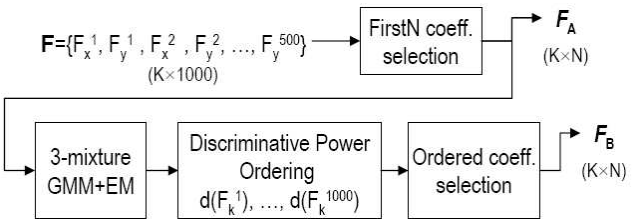

Fig. 2. Feature extraction through $FirstN$ DCT and Bayesian discriminative feature selection

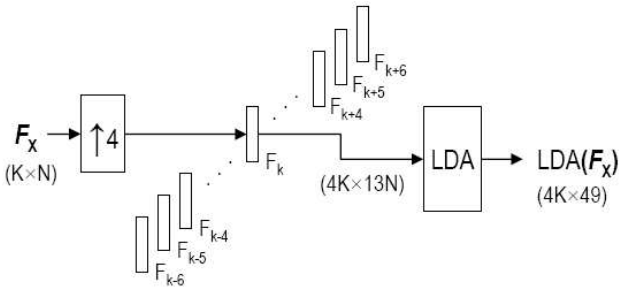

Fig. 3. Feature extraction through LDA. A concatenated set of features centered at time instant k of dimension $13N$ is applied to LDA.

training data of the population. Then each HMM associated to a speaker is trained over some repetitions of the lip-motion streams of the corresponding speaker. In the identification process, given a test feature set, each HMM structure associated with speakers and the world class produces a likelihood. 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 system.

5. EXPERIMENTAL RESULTS

The performance analysis of the open-set speaker identification system is done 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). False accept and false reject rates are defined as,

$$\begin{aligned} \text{FAR} &= 100 \times \frac{\text{number of false accepts}}{N_a + N_r} \\ \text{FRR} &= 100 \times \frac{\text{number of false rejects}}{N_a} \end{aligned} \quad (7)$$

where N_a and N_r are the total number of trials for the true and imposter clients in the testing, respectively.

Let $\boldsymbol{D}_T$ represents the whole database for the true clients. The $\boldsymbol{D}_T$ database is partitioned into two sets namely $\{\boldsymbol{D}_{T_A}$ and $\boldsymbol{D}_{\bar{T}_A}\}$, where $\boldsymbol{D}_{T_A}$ and $\boldsymbol{D}_{\bar{T}_A}$ are mutually exclusive sets each having five repetitions from each subject in the database. The subsets $\boldsymbol{D}_{T_A}$ and $\boldsymbol{D}_{\bar{T}_A}$ are used for training and testing respectively. As there are 50 subjects and five repetitions for each true and imposter client tests, the resulting total number of trials becomes as $N_a = 250$ and $N_r = 250$.

Figure 4 presents the equal error rate (EER) performance of the *FirstN* ($\boldsymbol{F}_{\mathbf{A}}$) and *DiscriminativeN* ($\boldsymbol{F}_{\mathbf{B}}$) features at varying dimensions. It is clear that at lower dimensions, EER performance of *DiscriminativeN* is better than *FirstN* EER performance.

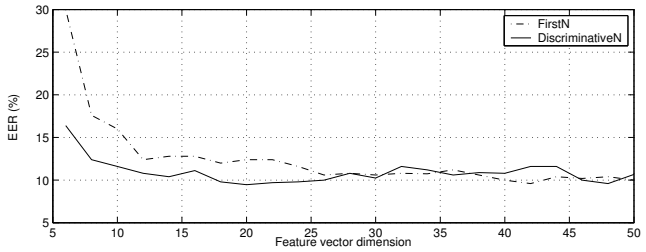

Fig. 4. The EER performance of the *FirstN* and *DiscriminativeN* features at varying feature vector dimensions.

Figure 5 presents the equal error rate (EER) performances of three different experiments, where the DCT based dimensions are picked as $N = 40$, 30, and 20, for LDA-reduced feature sets $\boldsymbol{F}_{\mathbf{C}}$ and $\boldsymbol{F}_{\mathbf{D}}$. The dimension of the concatenated feature vector is reduced to 49 using LDA and the EER performance is plotted for varying dimensions (6-49) of the LDA-reduced features. It is seen that the EER performance in general increases as the feature vector dimension increases and the overall EER performance is maximized at 5.2% rate with $\boldsymbol{F}_{\mathbf{D}}$ features for the second experiment when $N = 30$. Note that the best $\boldsymbol{F}_{\mathbf{C}}$ feature EER performance is 6.8% again in the second experiment when $N = 30$.

6. CONCLUSION

A discriminative lip feature selection method has been presented for the open-set speaker identification problem. The proposed method can select the most discriminative feature components from the full set of DCT coefficients. The identification performance in EER scale is improved at relatively low dimensions. A further EER performance gain is achieved using LDA over a window of features that captures more temporal information. It is worth to note that LDA followed by Bayesian discriminative feature selection outperforms solely applying LDA. Hence, the Bayesian discriminative feature selection with the LDA method provides performance maximization by optimally selecting discriminative lip-motion features for speaker identification.

The proposed lip features can be used in conjunction with audio to improve the performance of the multimodal speaker identification systems. Another direction of the future work will be to improve the feature extraction phase via lip tracking and motion estimation.

7. REFERENCES

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

[2] 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.

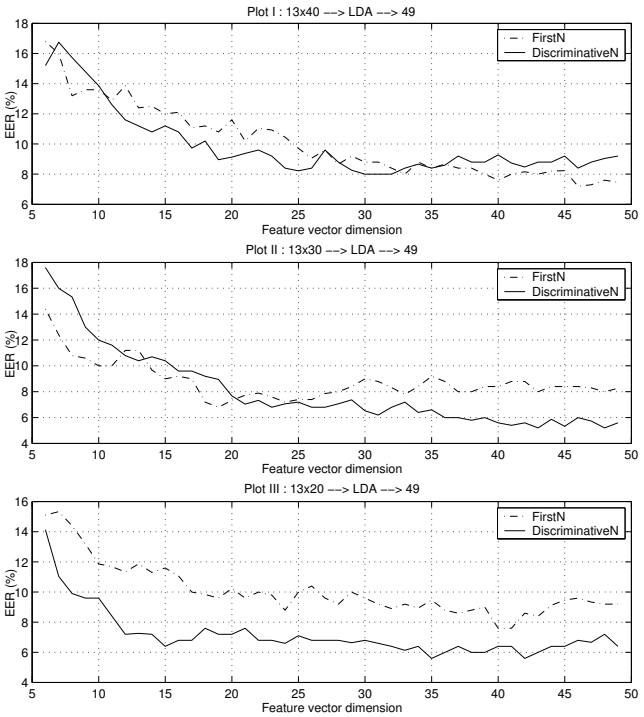

Fig. 5. The EER performances of the *FirstN* and *DiscriminativeN* features at varying feature vector dimensions. The concatenated feature vectors have base dimensions of $N = 40, 30$, and 20 from top-to-bottom, respectively.

[3] 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.

[4] 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.

[5] X. Huang, A. Acero, and H-W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*, Prentice Hall, 2001.

[6] E. Erzin, Y. Yemez, and A. M. Tekalp, *DSP in Mobile and Vehicular Systems*, chapter Joint Audio-Video Processing for Robust Biometric Speaker Identification in Car, Kluwer Academic Publishers, forthcoming.

[7] 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.