ROBUST LIP-MOTION FEATURES FOR 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 robust lip-motion features for audio-visual open-set speaker identification problem. We consider two alternatives for initial lip motion representation. In the first alternative, the feature vector is composed of the 2D-DCT coefficients of the motion vectors estimated within the detected rectangular mouth region whereas in the second, lip boundaries are tracked over the video frames and only the motion vectors around the lip contour are taken into account along with the shape of the lip boundary. Experimental results of the HMM-based identification system are included for performance comparison of the two lip motion representation alternatives.

1. INTRODUCTION

It has been a common practice to use lip information for speech recognition applications [1, 2, 3]. This is justified by the observation that the 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 the principal components of the lip movement and to establish 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 the identity of an individual as well as what the individual is speaking. [4] is one of the recent works showing the improved performance of speech-lip fused systems over those of speech-only systems. For the speaker identification problem however, the use of lip motion is a more sophisticated issue and has been addressed in only few works such as [4, 5, 6]. The main reason for this is that the principal components of the lip movement are not usually sufficient to well discriminate the biometric properties of a speaker. High frequency or non-principal components of the 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. The success of a lip-based speaker identification system depends very much on the accuracy and precision of lip tracking and/or lip motion estimation procedure.

In audiovisual speech/speaker recognition literature, there exist basically three alternatives for initial representation of lip-motion features: 1) the use of raw intensity values on the rectangular grid of the mouth region [2, 4], 2) the use of motion vectors instead of intensity values [5, 7] and 3) the use of lip shape parameters [6]. The first option represents the lip motion only implicitly along with some texture information that might sometimes carry useful discrimination information; but in many occasions the texture may also corrupt the identification task since it is very sensitive to acquisition conditions. Moreover, texture information can more adequately be incorporated to a recognition system via a multimodal fusion system [4, 5]. Thus, in this paper we rather focus on the second and third options where the lip motion is more explicitly and adequately represented. The last option seems to be the most powerful one, but only with the condition that the lip contour can accurately be tracked. However, this is a very challenging task especially in adverse circumstances since lip contour tracking algorithms are in general very sensitive to lighting conditions and image quality; in such cases, detection of the rectangular mouth region is relatively an easier task to accomplish.

In our work, we consider two different scenarios. In the first one, the rectangular mouth region is first to be detected and then the mouth movement is represented by the motion vectors computed on a predefined rectangular grid within this region. Thus, no explicit information about the lip shape is included in the feature vector. The main disadvantage of this strategy is that some irrelevant noisy motion vectors may show up especially inside the inner lip boundary as parts of this region are occluded or uncovered during the speaking act. In the second scenario, the lip boundary has to be tracked over time and only motion of lip boundary pixels are taken into account. In this way, noisy motion vectors are mostly eliminated at the cost of disregarding some useful motion information around the lip. One advantage of this strategy is that extracted lip shape information can explicitly be included and exploited in the feature set. However, as stated before and demonstrated in our experiments, robustness issue in lip contour tracking is still an unsolved problem.

The success of a lip-based speaker identification system eventually depends on how much of the obtained precision, that is useful for discrimination, is then included in the reduced low-dimensional feature set. In this work, the dimension of the initial lip feature vector is reduced by using the 2D-DCT and may be further subjected to a two-stage discrimination analysis so as to exploit both temporal and spatial correlations [7].

2. LIP-MOTION FEATURE EXTRACTION

2.1. Lip Contour Tracking

There are a number of approaches such as splines, active contours, and parametric models in the literature in order to represent and extract the lip contour. Classical active contours and splines suffer from complex parameter tuning and they are mostly unable to perfectly fit to the characteristic lip parts such as Cupidon's bow because of the erroneous gradient information due to illumination

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

differences.

Eveno et al. [8] proposed to fit cubic polynomials on the outer lip contour using the color information of the lip image. In this technique a preprocessing stage to find predefined 6 key points on the lip is followed by an optimization stage in which four cubic polynomials and two lines are fitted to the outer lip contour.

The preprocessing stage includes computing the pseudo hue component $h(x, y)$ at pixel (x, y) for lip segmentation,

$$h(x,y) = \frac{R(x,y)}{G(x,y) + R(x,y)} \tag{1}$$

where $R(x, y)$ and $G(x, y)$ are the red and green color components at pixel (x, y), respectively. Furthermore, the "hybrid edges" concept for upper and lower lip localization is introduced in the preprocessing,

$$\begin{aligned} R_{top} &= \nabla[h_N(x,y) - L_N(x,y)] \\ R_{mid} &= \nabla[L_N(x,y)] \cdot h_N(x,y) \\ R_{low} &= \nabla[h_N(x,y) + L_N(x,y)] \end{aligned} \tag{2}$$

where $h_N(x, y)$ and $L_N(x, y)$ represent the normalized pseudo hue and the normalized intensity components respectively.

Two key points on the lip corners, three points for the Cupidon's bow and the last one on the lower lip together constitute the 6 points to characterize the lip shape. In our implementation, the three points on the Cupidon's bow are automatically found by estimating the Cupidon's bow boundary by an edge tracking algorithm and then using the local maxima and minima of the function:

$$d(x) = \sum_{y=y_{top}-\delta}^{y_{top}+\delta} [h_N(x,y) - L_N(x,y)] \tag{3}$$

as in [8], where $[y_{top} - \delta, y_{top} + \delta]$ locates the upper lip boundary strip. Furthermore, the lip corners are also placed in the same manner using the minima of the intensity component computed along each vertical pixel group.

The technique proposed in [8] works well only under some assumptions on the acquisition environment and illumination conditions. However in many practical conditions such as ours, these assumptions do not hold. When tested on our visual database, the algorithm fails in about one-third of the sample video sequences. The lack of discriminative color information, especially on the lower lip boundary, becomes occasionally so severe that even a human eye can hardly make a distinction. Thus we use a quasi-automatic strategy that needs user interaction. In cases where the algorithm fails, the tracking task is assisted with some hand-labeled points on the lip boundary.

The modeling stage is then basically a least-squares (LS) optimization task on the color information to find the four cubic polynomials, two for the upper lip boundary and the other two for the lower lip boundary [8]. More specifically, when there is not any assistant point, the LS stage finds the best fitting polynomial using only the end points $m(x, y)$, $n(x, y)$, and $k(x, y)$. However, if the user needs to put an assistant point for one of the polynomials, the optimization procedure also uses this additional point $a(x, y)$:

$$y = c_1 x^3 + c_2 x^2 + c_3 x + c_4 \tag{4}$$

In other words, the best cubic polynomial of the form 4 is found using the points set $S_1 = \{m, k\}$ or $S_2 = \{m, a, n\}$.

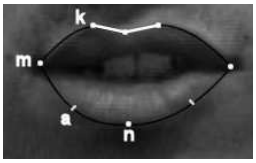

Fig. 1. The LS optimization for 2 polynomials to be fitted between [m,k] and [m,a,n] and the fitted lines on the Cupidon's bow (white lines).

This stage is completed by forming the Cupidon's bow. Figure 1 illustrates these two optimization cases. In Figure 2, some lip tracking results are presented illustrating the key points found on the lip contour and the fitted parametric models under different illumination conditions.

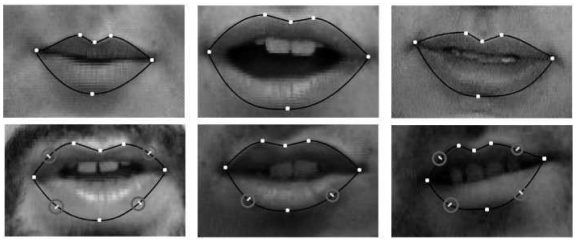

Fig. 2. Lip contour extraction by parametric model fitting. Characteristic points are shown as dots and the hand-labeled "assistant" points are marked with circles.

2.2. Lip Motion Representation

The first scenario is the use of a uniform grid of size $N \times M$ on the intensity lip image. The grid sizes used in our work are given in Section 4. This grid definition allows us to analyze the information content of both the lip area and the non-lip area inside the rectangular mouth region. The motion vectors are computed on this grid by a two-level hierarchical block-matching algorithm. The l_1-norm is used to match the blocks. The best matches with l_1 distance larger than a certain threshold are eliminated to avoid erroneous motion vectors. The threshold is automatically adjusted via a histogram analysis. Following the motion estimation, the x and y components of the motion vectors are separately transformed via 2D-DCT.

The extracted parametric lip contour is in fact a rough sketch of the real lip and does not contain sufficiently detailed information to characterize discriminative biometrics of different speakers. Discriminative information can be provided by incorporating the motion vectors computed along the parametric lip contour. Thus in the second scenario, only the motion vectors estimated on the pixels of the extracted lip contour are taken into account and the rest is discarded. Since this time we do not have a 2D grid, the two sequences of x and y motion components on the contour pixels are separately transformed using 1D-DCT. The final lip feature vector is formed by concatenating these DCT coefficients along with lip shape parameters.

The lip shape is represented in the feature vector with 4 parameters: maximum horizontal distance, and the 3 vertical distances

from the Cupidon's bow points to the lower lip boundary. Figure 3 shows these lip shape parameters. These parameters are then added to the feature vector in order to consider the effect of the lip shape to the identification performance.

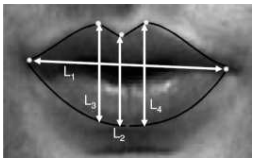

Fig. 3. The 4 lip shape parameters added to the lip-motion feature vector.

There exist a number of subspace representation techniques that can be used as a solution to the dimensionality problem of recognition systems. We introduced a two-stage discriminative feature selection technique in [7], where the Bayesian discriminative feature selection stage takes into account the intra-class and inter-class distribution of individual single-frame feature vectors whereas the second stage, i.e. LDA-based discrimination reveals the temporal correlations. The reader may refer to [7] for details.

3. SPEAKER IDENTIFICATION SYSTEM

Biometric speaker identification experiments are conducted using the audio-visual database MVGL-AVD [9]. 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 [10]. Then the face images are warped according to these calculated parameters. After this alignment, the motion vectors from the lip frames of size 128×80 are extracted using hierarchical block-matching technique. The blocks used in the estimation are of size 15×9 and 15% of the best matches with the largest l_1-norm are eliminated. The hierarchical block-matching allows a block of mid-point (x_m, y_m) to move $[y_m-7, y_m+7]$ vertically and $[x_m-3, x_m+3]$ horizontally. When working with the rectangular grid, the lip-motion vectors on x and y directions are separately transformed into DCT domain and the first C 2D-DCT coefficients of the zig-zag scan both on x and y directions are combined to form a feature vector $\boldsymbol{F}$ of dimension $2 \times C$. In case of contour processing, after interpolating both of the motion vectors on x and y directions to vectors of maximum allowable length in the database the first C_{max} 1D-DCT coefficients of the motions vectors are combined with possible concatenation of the lip shape parameters. This feature extraction procedure is illustrated in Figure 4.

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 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 training data of the

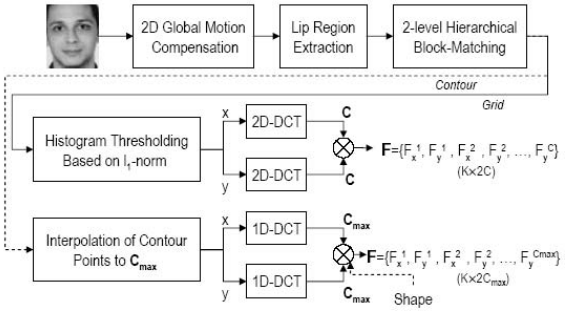

Fig. 4. Feature extraction methodology.

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.

4. EXPERIMENTAL RESULTS

The performance analysis of the open-set speaker identification [7] 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,

$$\begin{aligned} \text{FAR} &= 100 \times \frac{\text{number of false accepts}}{N_a + N_r} \\ \text{FRR} &= 100 \times \frac{\text{number of false rejects}}{N_a} \end{aligned} \quad (5)$$

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 5 presents the equal error rate (EER) performance of the rectangular grids 64×40, 32×20, and 16×10 at which lip-motion features are computed at varying dimensions. The EER performances of the grids of sizes 64×40, 32×20, 16×10 are maximized at 8.4%, 8.4%, and 8.9% at dimensions 37, 43, and 37, respectively. Thus, it is concluded that lower dimensions, the EER performance of the grid of size 64×40 is slightly better than the results of the other grids.

Figure 6 presents the EER performance of the features extracted from the lip contour with/without the lip shape information at varying dimensions. The EER performances of the lip contour without/with shape information are maximized at 11.2% and 8.4% at dimensions 20 and 21, respectively. Therefore, it is clear that the EER performance gain is 2.8% if the lip shape information (i.e.

only 4 parameters L_1 through L_4) is added to the feature vector extracted from the lip contour. For the sake of completeness, another experiment has been performed using feature vectors formed by fusing 2D-DCT coefficients of the grid of size 64×40 and the 4-parameter lip shape information. Figure 7 presents the EER performance of the features extracted from the grid with/without the lip shape information at varying dimensions. The EER performances of the grid without/with shape information are maximized at 8.4% and 7.6% at dimensions 37 and 41, respectively. Our conclusion on the EER performance increase by fusing lip shape information with other available vectors has been justified.

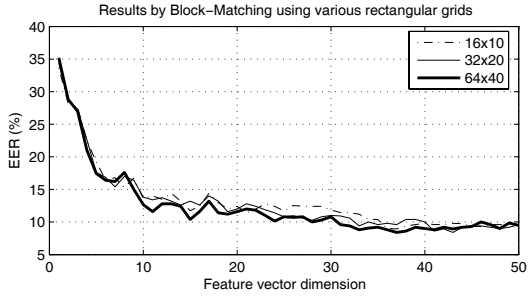

Fig. 5. EER results for rectangular grids of size 64×40, 32×20, and 16×10.

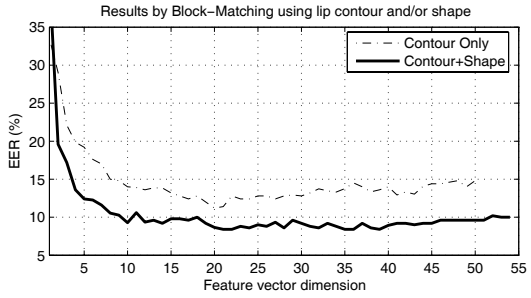

Fig. 6. EER results for lip contour and lip shape information.

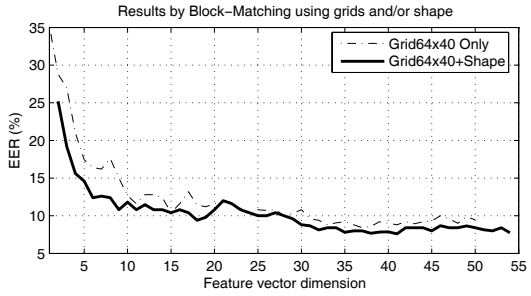

Fig. 7. EER results for a 64×40 grid and lip shape information.

5. CONCLUSION

A quasi-automatic system to extract and analyze robust lip-motion features is presented for the open-set speaker identification problem. It is concluded that the utilization of the grid points for motion vector computation is better than using only lip contour points. This shows the importance of the skin region for identification even if it introduces some erroneous vectors. Moreover, it is worth noting that the additional shape information greatly improves the EER performance and becomes indispensable for speaker identification problem. Therefore, if available, accurate and robust lip-motion information is an asset to improve the performance of unimodal (i.e. speech-only) systems, which are mostly corrupted by noise in real-life.

Further studies will primarily investigate the effect of discriminative feature selection techniques proposed in [7] on the EER performance, the performance of the overall speaker identification system using one specific secret phrase (i.e. password) and features fused with corresponding speech modality.

6. 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.

[3] D. G. Stork and M. E. Hennecke, *Speechreading by Humans and Machines: Models, Systems, and Applications*, Springer-Verlag, 1996.

[4] E. Erzin, Y. Yemez, and A. M. Tekalp, "Multimodal speaker identification using an adaptive classifier cascade based on modality reliability," *accepted for publication on IEEE Transactions on Multimedia*, 2004.

[5] 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.

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

[8] N. Eveno, A. Caplier, and P. Y. Coulon, "A Parametric Model for Realistic Lip Segmentation," *Proc. of 7th Int. Conf. on Control, Automation, Robotics and Vision (ICARV2002)*, December 2002.

[9] 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, October 2004.

[10] 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.