FEATURE EXTRACTION FROM TALKING MOUTHS FOR VIDEO-BASED BI-MODAL SPEAKER VERIFICATION

Hua Ouyang, Tan Lee and W.N. Chan

Department of Electronic Engineering
The Chinese University of Hong Kong
Shatin, N.T., Hong Kong
Email: {houyang, tanlee, wnchan}@ee.cuhk.edu.hk

ABSTRACT

As the low-cost video transmission becomes popular, video-based bi-modal (audio and visual) authentication has great potential in various applications that require access control over handheld terminals. In this paper, we propose to use the averaged mouth image (AMI) for speaker verification. The AMI is computed by averaging properly aligned mouth images over the whole video sequence. Despite its simplicity, the AMI not only contains appearance information but also describes stylistic articulation gestures of individual speakers. The AMI is found to be fairly invariant against the spoken content. The experimental results show that the AMI based features are very effective in discriminating speaking persons. Explicit and precise extraction of lip contours or other feature points are not required. For bi-modal verification, the proposed video features are found to be highly complementary to the audio features.

1. INTRODUCTION

Automatic face recognition has been an active area of research for over 30 years [1]. The initial focus was on the use of still images. Recently, video-based face recognition has attracted great attention [2][3][4]. The intended applications are access control or restricted login through handheld terminals, on which low-cost video acquisition and transmission become increasingly popular recently.

This research addresses the problem of video-based bi-modal speaker authentication. It aims at verifying a person's identity, given a video that contains, at least, his/her face. The video is composed of synchronously recorded visual and audio signals, and thus provides both face and voice biometrics of the speaker. These two aspects of information are considered to be highly complementary to each other, especially in terms of their robustness against adverse conditions. For example, the change of illumination doesn't much affect the effectiveness of voice biometrics. Audio-visual bi-modal authentication has been studied extensively. In [5], frontal face images and voice were used in combination by a linear weighted classifier as well as Support Vector Machine (SVM). In [6], a coupled hidden Markov model (HMM) was used to combine lips, face and voice features for person identification. Video-based speaker authentication also benefits from additional information that stems from inter-frame changes. The spatial-temporal changes in a video sequence provide a lot more identity-related information than snapshot images. A tracking-for-verification system was proposed in [7]. It showed that the trajectory pattern of tracked feature points from the same face is more coherent than those from different faces. In [8], a 3-D point distribution model and a shape-and-pose-free texture model were constructed from 2-D face images to represent the depth information.

On handheld devices, the captured face image sequences are usually of low quality and low resolution because of bandwidth limitation. Many discriminative details may be lost when the image size becomes small. From applications point of view, face recognition tends to be vulnerable to hair style variation, cosmetic makeup, presence of glasses and facial expressions. When the mouth area is chosen as the primary region of interest (ROI), these variations are less problematic in practical applications.

In video-based speaker verification, non-rigid deformation of the articulation-related area provides very useful information that differentiates one person from another. Lips, as one of the most important viewable articulators, can be shaped differently to produce different sounds. Different styles of articulatory gesture are reflected in the movement of lips and the so caused visual features of other partially visible articulators like teeth and tongue. Since most salient movement of a talking face belongs to the mouth area, we believe that robust and discriminative speaker-specific visual features can be extracted from this area. Lip movement features can be captured by the optical flow technique [9]. In [10] and [11], motion cues of lips during speech were exploited based on extracted lip contours or motion vectors.

In this paper, we propose to use the averaged mouth image (AMI) for speaker verification. Our conjecture is that the precise frame-by-frame visual movement of lips and other articulators may contain much detail information that are redundant or even causing confusion in speaker discrimination. The long-term average of the articulation images tends to better reflect personalized articulation styles. The same idea was used in recognizing speakers using long-term average spectrum [12]. Given a sequence of talking face images, the ROI is located to roughly cover the mouth area and then properly aligned across all frames. The AMI is computed by averaging the ROI images over the video sequence. Despite its simplicity in computation, the AMI not only contains appearance information but also describes stylistic articulation gestures of individual speakers. AMI is found to be fairly invariant against the spoken content. The dimension of AMI is reduced by Multiple Discriminant Analysis (MDA) and the resulted feature vectors are recognized by Euclidean distance based pattern matching. Experimental results show that the proposed video based features are very effective in discriminating speaking persons. The video sub-system is combined with a state-of-the-art audio based speaker verification system via linear score fusion.

2. THE XM2VTS AUDIO-VIDEO DATABASE

The XM2VTS database [13] is used in our bi-modal speaker verification experiments. It is by far the largest audio-video databases designed for multi-modal biometric research. Face image sequences and speech were simultaneously recorded. The database involves 295 subjects. Each of them has 4 sessions recorded with an inter-session interval of one month. Each session consists of two video shots, in each of which a string of 20 English digits was spoken. The 295 subjects were randomly divided into 200 clients, 25 evaluation imposters and 70 test imposters. In our experiments, the recommended Configuration II is adopted [13]. That is, the first two sessions of the clients (4 shots per client) are used for training of client models, the third session is for system tuning and the fourth session is used as test data.

Originally the audio signals in XM2VTS were sampled at 32kHz. In our experiments, they were down-sampled to 16kHz. In addition to the original audio, two sets of noise-corrupted audio data are created. White noise is digitally added to the original audio at the SNRs of 10dB and 5dB.

All video clips were captured with a blue background at the frame rate of 25 frames per second (fps). The original resolution is 720*576 pixels with 24-bit colors. In order to simulate the video quality of typical handheld devices, the images are Gaussian blurred (radius: 1 pixel) [14] and down-sampled to 360*288 pixels. Furthermore, the color images are converted into 8-bit gray-scale images for subsequent processing.

3. AVERAGED MOUTH IMAGES FOR SPEAKER CHARACTERIZATION

3.1. Mouth Corner Detection

In the intended application, the video captures the frontal face of a speaker. To extract articulation-related visual features, we firstly need to locate the ROI, which is a small area around the mouth. Localization of the mouth from a face image has been a research topic for many years. State-of-the-art techniques of lip localization can achieve good performance if there is no substantial illumination variation. In this research, we assume that the ROI for feature extraction can be reliably detected. Without tackling the localization problem in depth, we adopt a semi-automatic method of mouth corner detection to localize the ROI. For the first frame of a video sequence, the two mouth corners are manually marked. Given a subsequent frame, the mouth corners are determined around the neighboring regions of the corner positions of the previous frame. The method described in [15] is adopted to perform corner searching. The major steps are: 1) Canny edge detection - every detected edge is regarded as a candidate for corner lines; 2) selection of corner line pairs using a small moving window; 3) clustering of nearby corners; and 4) search for the best corner candidate based on a cost function that measures the dissimilarity of inner and outer mouth corner areas.

Figure 1 shows some example results of mouth corner detection. This method seems not to be robust to faces with complex textures, or with the mouth corners covered by bushy beard. There are 31 subjects (out of 295) that have such problems. They are not used in our experiments. In practice, the skin texture or the beard itself can be a distinctive feature for speaker recognition, but this is beyond the scope of this paper. The remaining 264 subjects include 182 clients, 20 evaluation imposters and 62 test imposters.

Based on the corner locations, the facial pose in each frame is normalized by rotation and a rectangle of 60*50 pixels is cropped to give an image of the ROI.

Fig. 1. Example results of mouth corner detection

3.2. Averaged Mouth Images

The information conveyed by a sequence of talking mouth images can be categorized into those being related to the speaker identity and those describing the residual situations, such as pose, ambient lighting, expressions and linguistic contents. Our motivation is to discard the speaker-irrelevant variations and keep speaker-dependent physical and behavioral features by exploring all contents in the image sequence. The physical features refer mainly to the geometric shape of the lips while the behavioral features are related to the movement of lips and other articulators in correspondence to speech production.

The geometric features of lips can be estimated from individual images. However, the resulted features would depend on the content being spoken at that moment. To remove such text-dependent variation, we can use the average image over many frames in the video sequence. If the video sequence covers different articulatory gestures, the resulted average image would give a more robust representation of the geometric features than the individual images.

An intuitive understanding of features related to the movement of articulators is that different people have different articulatory styles when speaking the same utterance. For examples, when producing the same phoneme or word, some speakers tend to widely open their mouths while the others may only open slightly; some speakers tend to move the lips in a specific direction, and for some speakers, the teeth are always visible when speaking. The time-averaged image described above also contains such stylistic features of articulation.

Given a video sequence of aligned mouth images, the average intensity level at pixel (i, j) is computed as

$$\overline{x}_{i,j} = \frac{1}{M} \sum_{m=1}^{M} x_{i,j,m} \tag{1}$$

where $x_{i,j,m}$ is the intensity value of pixel (i, j) in frame m, and M is the total number of frames in the video clip. $\overline{x}_{i,j}$ is used to construct the averaged mouth image (AMI).

Figure 2 shows some examples of AMI computed from four speakers in the XM2VTS database. The four rows are taken from four different time sessions. The same content was spoken among all sessions. The ROIs are localized as described in Section 3.1. Subsequently the images are rotated and aligned so that the intra-class variations caused by head movement and small pose changes can be reduced. As seen from the figure, good inter-session consistency can be attained in the AMIs from the same speaker.

In Figure 3, three AMIs from the same speaker with different spoken contents are compared. In all cases, the duration of video is about 5 seconds, in which more than 30 English phonemes are uttered and various articulation gestures are covered. It can be seen

that the AMIs of the same speaker remain fairly invariant for different spoken contents.

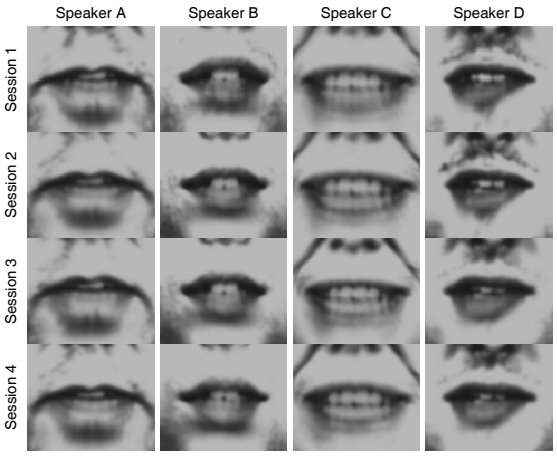

Fig. 2. Examples of Averaged Mouth Images. Each column contains the AMIs of the same speaker in four different recording sessions

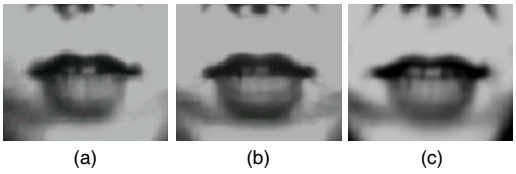

Fig. 3. AMIs of the same speaker uttering different contents. (a) English digit string "0 1 2 3 4 5 6 7 8 9" (b) English digit string "5 0 6 9 2 8 1 3 7 4" (c) English sentence "Joe took fathers green shoe bench out" [13]

Figure 4 demonstrates how the speaker-specific stylistic features of articulation are represented by AMI. Figure 4(a) and (b) are snapshot mouth images of two different speakers. They are cropped from the first frames of the video, when the speakers are silent. These images reflect the static appearance of their mouths and related articulators. Figure 4(c) and (d) are the AMIs of these two speakers. Although the mouths of the two speakers look very similar, there are noticeable differences being revealed in their AMIs. The teeth of speaker A is more visible than speaker B. Speaker B tends to round his lower lip while speaking, while the lower lip of speaker A is more like a rectangle.

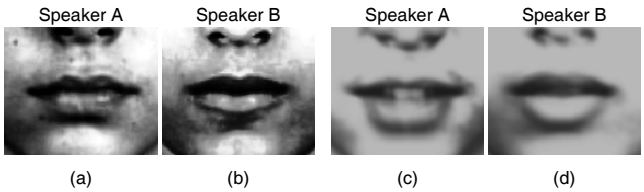

Fig. 4. Comparison between still mouth images and AMIs.

3.3. Multiple Discriminant Analysis and Pattern Matching

Each AMI can be regarded as a still image, which has the same size as the cropped mouth images (60*50 pixels). A dimension reduction method is needed for pattern classification. Principle Component Analysis (PCA) and Linear Discriminant Analysis (LDA) are two commonly used linear transformation methods that project high-dimensional data to a low-dimensional space. PCA is often used for data representation, while LDA is effective in discriminative classification [16]. They have been widely used for face feature extraction. Since the dissimilarities of mouth geometries among different persons are not as significant as that among their faces, holistic methods such as PCA may not be effective. This was observed in our preliminary experiments on the analysis of AMI. The method of Fisherfaces based on Multiple Discriminant Analysis (a multi-class generalization of LDA) [17] is used to transform an AMI into a feature vector of 181 dimensions. This feature vector is used for pattern matching based on Euclidean distance. For each client, four reference templates are established from the four video clips of the training sessions. Given an incoming feature vector, its distance with respect to the client is the average Euclidean distance with respect to the four reference templates.

4. BI-MODAL VERIFICATION EXPERIMENTS

4.1. Audio Sub-system

An audio based speaker verification system is developed using the state-of-the-art GMM-UBM technique [18]. The acoustic features include 13 Mel-frequency cepstral coefficients (MFCC) and their first and second-order derivatives. The universal background model (UBM) is a Gaussian mixture model (GMM) with 256 mixture components. It is trained with 2 sessions of the 182 clients. Speaker models are adapted from the UBM by Maximum a Posteriori (MAP) estimation. For each test utterance, the log-likelihood ratio, ℓ_a is computed as the difference between the log likelihoods of the claimed speaker model and the background model.

4.2. Score Fusion of Audio and Video Sub-systems

The verification scores produced by the two sub-systems are fused by simple linear combination, i.e.

$$s = \beta \cdot \Psi(\frac{1}{d_v}) + (1-\beta) \cdot \Psi(\ell_a) \qquad (2)$$

where β is the weighting factor trained from the evaluation session (the third session), d_v is the Euclidean distance produced by the video sub-system, ℓ_a is the log-likelihood ratio produced from the audio sub-system. Ψ is a normalization operator, which attempts to equalize the dynamic ranges of the two sets of scores.

4.3. Results and Discussion

Three experiments have been conducted for the cases of "original audio", "noisy audio 10dB" and "noisy audio 5dB" respectively (see Section 2). Figure 5 shows the detection error tradeoff (DET) curves attained by the bi-modal system as well as the uni-modal sub-systems. For the clarity of presentation, only "noisy audio 5dB" is shown in the figure. The DET curve for "noisy audio 10dB" resides between those of "original audio" and "noisy audio 5dB". Table 1 gives the performance in terms of Equal Error Rates (EER).

It can be seen that the bi-modal system takes the complementary advantages of the audio and video sub-systems. With the presence

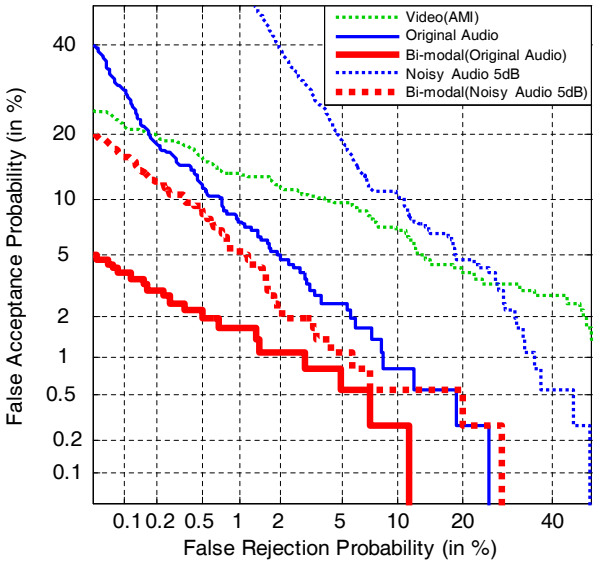

Fig. 5. Performance (DET curves) of uni-modal and bi-modal speaker verification

Table 1. Performance (EERs) of uni-modal and bi-modal speaker verification

	Video	Audio	Bi-Modal
Original Audio	7.69%	3.02%	1.37%
Noisy Audio 10dB		7.42%	1.65%
Noisy Audio 5dB		10.16%	1.92%

of audio noise, the performance of the audio sub-system degrades significantly but the bi-modal system is not affected much. In particular, although the EERs of the two sub-systems are 7.69% and 10.16% respectively (in the case of 5dB audio SNR), the bi-modal system can attain a much lower EER of 1.92%. This clearly shows that the video based features are highly complementary to the audio based ones.

5. CONCLUSIONS

AMI is very similar to long-term spectral average that has been used in text-independent audio-based speaker verification. While the use of AMI has attained a high performance level, it doesn't require explicit and precise extraction of lip contours or other feature points. We have also evaluated the AMI features in the task of video-based speaker identification and achieved an accuracy that is comparable to the methods of using the entire face. It is also clear that the AMI features are highly complementary to the acoustic features. The continuation of this work will be an investigation on the use of phoneme-dependent AMI, where the phoneme boundaries are detected by automatic speech recognition techniques.

6. ACKNOWLEDGEMENT

This research was partially supported by a Central Allocation Grant (Ref. CUHK1/02C) from the Hong Kong Research Grants Council.

The authors would like to thank Dr. Kenneth Lam for allowing us to use his software programs of mouth corner detection.

7. REFERENCES

[1] W. Zhao, R. Chellappa, and P.J. Phillips, "Face Recognition: A Literature Survey," *ACM Computing Surveys*, vol. 35, no. 4, pp. 399–458, 2003.

[2] S.J. Mckenna and S. Gong, "Recognising Moving Faces," in *Face Recognition: From Theory to Applications*, NATO Advanced Study Institute, Nov. 1998, pp. 578–687.

[3] S. Zhou, V. Kruger, and R. Chellappa, "Probabilistic Recognition of Human Faces from Video," *Computer Vision and Image Understanding*, vol. 91, pp. 214–245, 2003.

[4] X. Tang and Z. Li, "Video Based Face Recognition Using Multiple Classifiers," in *Proc. FGR'04*, 2004.

[5] S. Yacoub, J. Luettin, and J. Kittler, "Audio-Visual Person Verification," in *Proc. CVPR'99*, 1999.

[6] A.V. Nefian, L. Liang, T. Fu, and X. Liu, "A Bayesian Approach to Audio-Visual Speaker Identification," in *Proc. AVBPA'03*, 2003.

[7] B. Li and R. Chellappa, "Face Verificatioin through Tracking Facial Features," *Journal of the Optical Society of America*, vol. 18, no. 12, pp. 2969–2981, 2001.

[8] Y. Li, S. Gong, and H. Liddell, "Constructing Facial Identity Surfaces in a Nonlinear Discriminating Space," in *Proc. CVPR'01*, 2001.

[9] R.R. Frischholz and U. Dieckmann, "BioID: A Multimodal Biometric Identification System," *Computer*, vol. 33, no. 2, pp. 64–68, 2000.

[10] L.L. Mok, W.H. Lau, S.H. Leung, and S.L. Wang, "Lip Features Selection with Application to Person Authentication," in *Proc. ICASSP'04*, 2004.

[11] H.E. Cetingul, Y. Yemez, E. Erzin, and A.M. Tekalp, "Robust Lip-Motion Features for Speaker Identification," in *Proc. ICASSP'05*, 2005.

[12] J.P. Campbell, "Speaker Recognition: A Tutorial," *Proceedings of the IEEE*, vol. 85, no. 9, pp. 1437–1462, 1997.

[13] K. Messer, J. Kittler, J. Luettin, and G. Maitre, "XM2VTSDB: The Extended M2VTS Database," in *Proc. AVBPA'99*, 1999.

[14] M.C. Cheung, M.W. Mak, and S.-Y. Kung, "Intramodal and Intermodal Fusion for Audio-Visual Biometric Authentication," in *Proc. ISIMP'04*, 2004.

[15] K.M. Lam and H. Yan, "Locating and Extracting the Eye in Human Face Images," *Pattern Recognition*, vol. 29, no. 5, pp. 771–779, 1996.

[16] R.O. Duda, P.E. Hart, and D.G. Stork, *Pattern Classification*, Wiley-Interscience, 2001.

[17] P.N. Belhumeur, J.P. Hespanha, and D.J. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection," *IEEE Trans. on PAMI*, vol. 19, no. 7, pp. 711–720, 1997.

[18] D.A. Reynolds, T.F. Quatieri, and R.B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, 2000.