Audiovisual Speaker Identity Verification Based on Lip Motion Features

Girija Chetty, Michael Wagner

School of Information Sciences and Engineering
University of Canberra, Australia
girija.chetty@canberra.edu.au, michael.wagner@canberra.edu.au

Abstract

In this paper, we propose the fusion of audio and explicit lip motion features for speaker identity verification applications. Experimental results using GMM-based speaker models indicate that audiovisual fusion with explicit lip motion information provides significant performance improvement for verifying both the speaker identity and the liveness, due to tracking of the closely coupled acoustic labial dynamics. Experiments performed on different gender specific subsets of data from the VidTIMIT and UCBN databases under clean and noisy conditions show that the best performance of 7%–11% EER is achieved for the speaker verification task and 4%–8% EER for the liveness verification scenario.

Index Terms: audiovisual, speaker verification, liveness verification, lip motion

1. Introduction

Including visual information from the lip region in a speaker verification system can improve the speaker verification performance as well as providing a verification of the liveness of the recording, as it would be more difficult for an impostor to imitate both audio and dynamical visual information simultaneously [1]. The speaking face is a kinematic-acoustic system in motion [2]. Hence, the extraction of explicit lip motion features might allow better speaking face modeling for speaker verification and liveness checks than previously used implicit lip motion features, based on lip geometry, lip intensity and lip opening ratio.

The use of explicit lip motion features for tracking the lip dynamics, instead of or in addition to implicit features based on lip intensity and/or geometry information attempts to address two issues: 1) Is explicit lip motion information useful; and 2) if so, what are the best lip motion features for the joint modeling of speaker liveness and speaker identity?

The paper is organized as follows: Lip motion feature extraction is described in Section 2. The audiovisual fusion techniques are discussed in Section 3. The evaluation of the explicit and implicit lip motion features and the fusion of the audio-lip motion features for the speaker identity verification (SIV) and the liveness verification (LV) scenarios are described in Section 4. The conclusions follow in Section 5.

2. Explicit Lip Motion Features

There generally exist three alternative representations for lip information: 1) lip intensity (texture); 2) lip geometry (shape) and 3) lip motion features [4]. The first alternative implicitly represents lip movements with texture. Texture information may carry useful discrimination information, but sometimes may degrade the verification performance since it is sensitive to acquisition conditions. Lip geometry requires tracking of the lip contour and fitting contour model parameters or computing geometric features such as horizontal and vertical openings etc. Lip tracking and contour fitting are challenging tasks, since the algorithms are generally sensitive to lighting conditions and image quality. Explicit lip motion features, on the other hand, are potentially easy to compute and are robust to lighting variations between the training and test data sets.

The explicit lip motion feature extraction scheme employs face detection, normalisation and lip region extraction before determining the lip-motion features. Grid-based motion features were extracted by estimating dense motion over a uniform grid of size $G_x \times G_y$ on the extracted lip region image. We use hierarchical block matching to estimate the lip motion with sub-pixel accuracy (quarter-pel) by interpolating the original lip image using the 6-tap Wiener and bilinear filters specified in H.264/MPEG-4 AVC [3]. The motion estimation procedure yields two 2D matrices, which contain the G_x and G_y components of the motion vectors at grid points, respectively. The first M DCT coefficients along the zigzag scan order, both for x and y directions, are combined to form a feature vector f of dimension 2M as depicted in Fig. 1. This feature vector representing the dense grid motion will be denoted by f_{GRD}.

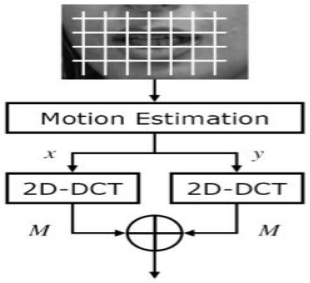

Figure 1: Grid-based lip motion extraction

For lip contour extraction, we employ the lip-geometric key points extracted in [5] and we fit polynomials on the outer lip contour based on a technique proposed in [6]. The technique is based on six designated key points detected on the lip contour. The algorithm fits additional points on the outer lip key points by guiding a "jumping snake" onto the upper-lip boundary [6]. The additional detected key points serve as the junction points of four cubic polynomials and two line segments to be fitted onto the lip contour via least-squares optimization. The DCT coefficients computed separately for the *x* and *y* directions are concatenated to form the feature vector that is denoted by f_{CTR}.

The contour-based lip motion feature vector f_{CTR} can further be fused with the lip shape parameters to improve the representation. We will denote the lip shape feature vector by f_{SHP}. The four cubic polynomial and two line segments, which are articulated, can therefore be represented by 14 points or 28 coordinates. The concatenation of lip shape parameters with contour-based motion information is illustrated in Fig. 2.

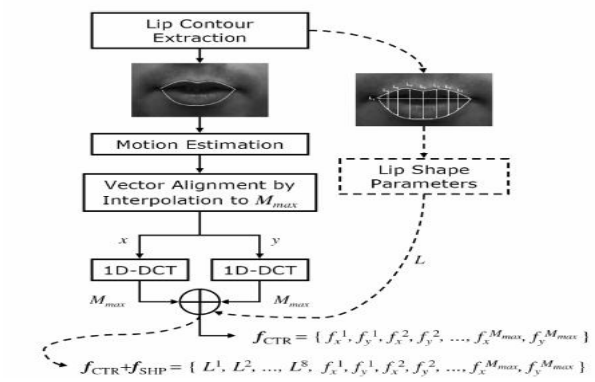

Figure 2 Lip shape and contour-based motion features.

3. Audiovisual Fusion

For audiovisual fusion, the audio features and lip motion (implicit and explicit) features were extracted separately and fused together. The explicit lip-motion features and the implicit lip motion features according to [5] were examined. The audio features were obtained by dividing the audio signal into frames using a Hamming window of length 30ms. A frame overlap of 10ms gives an audio frame rate, FA, of 50Hz. An acoustic feature vector was determined for each frame by warping 512 spectral bands into 30 mel-spaced bands, and computing the MFCCs. Cepstral mean normalization [7] was performed in order to compensate for the varying environmental noise and channel conditions. Since the audio and visual frame rates are different ($FA = 2 \times FV$), the visual features are duplicated for consecutive audio frames.

Every speaker is modeled with a single GMM model using the synchronized audiovisual fusion features. The test data are different for the SIV scenario and the LV scenario. For the SIV scenario, the impostor data come from the original audiovisual sequences of different speakers in the data subsets, whereas for the LV scenario, the impostor data for simulating a replay attack needs to be manufactured or synthesized.

4. Experiments

4.1. Experimental Setup

For the SIV scenario performance evaluation, different subsets of data from VidTIMIT [7] and UCBN were used. Gender-specific universal background models (UBMs) were developed using training data from Sessions 1 and 2 of the VidTIMIT corpus, and for testing Session 3 was used. Due to the type of data, only text-independent speaker verification experiments could be performed with the VidTIMIT database. This gave 1536s of training data for the male UBM and 576s of training data for the female UBM. A GMM topology with 10 mixtures was used for all the experiments. The number of mixtures was determined empirically to give the best performance. For the UCBN database, similar gender-specific UBMs were obtained using training data from text-dependent Subset 1 and text-independent Subsets 3 and 4. Ten sessions of the male and female speaking face data from these subsets were used for training and five for testing.

The approach used for the evaluation of the audiovisual features for the LV task was different compared to the SIV scenario. Here, an impostor attack is surreptitious replay of previously recorded data. The replay attack sequences have to be synthesized from the original audiovisual sequences in the VidTIMIT database.

Liveness verification experiments involved two phases, the training phase and testing phase. The training phase was similar to the SIV task, where a 10-mixture Gaussian model λ of a client's audiovisual feature vectors was built, reflecting the probability densities for the combined phonemes and visemes in the audiovisual feature space.

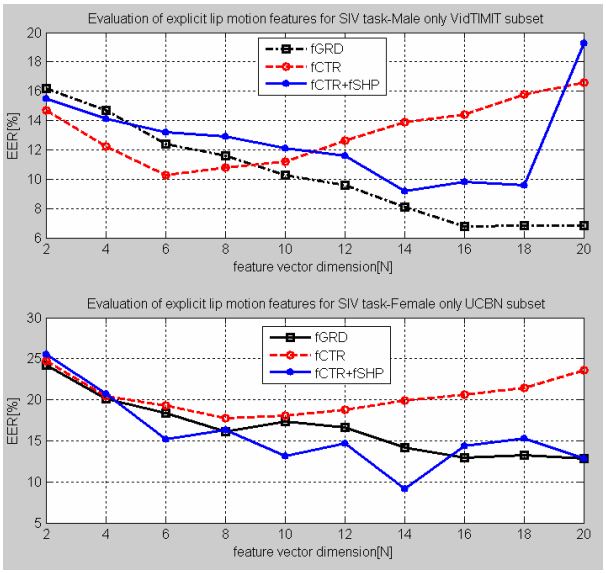

Figure 3 Explicit lip motion features for the SIV scenario

The testing phase was different, where the clients' live test recordings were first evaluated against the client's model λ by determining the log likelihoods log p(X|λ) of the time sequences X of audiovisual feature vectors under the usual assumption of statistical independence of successive feature vectors.

For testing static replay attacks, a number of "fake" or synthetic recordings were constructed by combining the sequence of audio feature vectors from each test utterance with ONE visual feature vector chosen from the sequence of visual feature vectors and keeping that visual feature vector constant throughout the utterance. Such a synthetic sequence represents an attack on the authentication system, carried out by replaying an audio recording of a client's utterance while presenting a still photograph to the camera. Four such fake audiovisual sequences were constructed from different still frames of each client test recording. Log-likelihoods log p(X'|λ) were computed for the fake sequences X' of audiovisual feature vectors against the client model λ. In order to obtain suitable thresholds to distinguish live recordings from fake recordings, detection error trade-off

(DET) curves and equal error rates (EER) were determined. Different sets of experiments were conducted to evaluate the performance of the implicit, explicit and fusion features in terms of DET curves and equal error rates (EER).

4.2. Experimental Results

The three explicit lip motion feature representations f_{GRD}, f_{CTR}, $f_{CTR} + f_{SHP}$ are tested on the male and female subsets of the VidTIMIT and UCBN corpora. Figure 3 shows that the grid-based features f_{CTR} yield the best EERs of about 7% for males and 13% for females in the SIV scenario with a feature space dimension of 16 and higher.

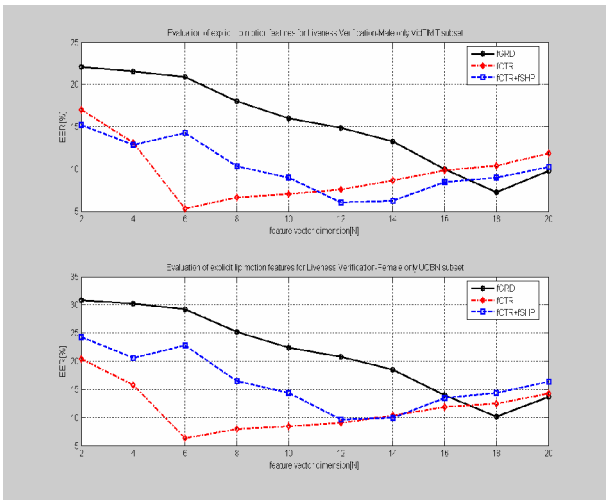

Figure 4 Explicit lip motion features for the LV scenario

Figure 4 shows the EER performance for different lip motion representations with varying feature dimension N for the VidTIMIT male subset and the UCBN female subset for the LV scenario. The contour-based lip features yield EERs of about 5% and 6% for the male and female subsets, respectively, at a feature space dimension of 6, whereas the grid-based features yield EERs of about 7% and 10%, respectively, for a feature space dimension of 18.

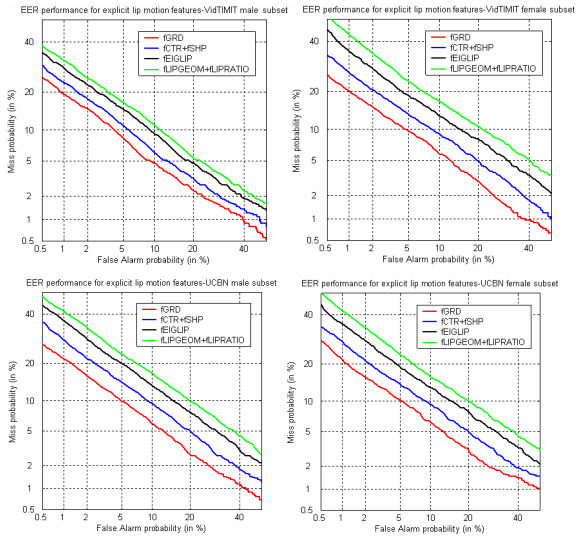

Figure 5 Audiovisual late fusion for the SIV scenario.

Next, we performed experiments to determine whether multimodal fusion of audio MFCC features and explicit lip motion (f_{GRD}, f_{CTR}, f_{SHP})-features, provides further performance gain in the SIV scenario. As shown in the DET curves of Figure 5, the best EER performance of 7% is achieved with late fusion of the audio and grid-based explicit lip motion features, f_{GRD}, for the VidTIMIT male subset.

The EER performance achieved with feature-level fusion of audio and implicit and explicit lip motion features is shown in the DET curves of Figure 6. Feature-level fusion leads to similar EER performance as late fusion. This is encouraging since we require feature fusion for liveness testing and since some earlier work [8] had reported that feature-level fusion leads to performance loss compared with late fusion.

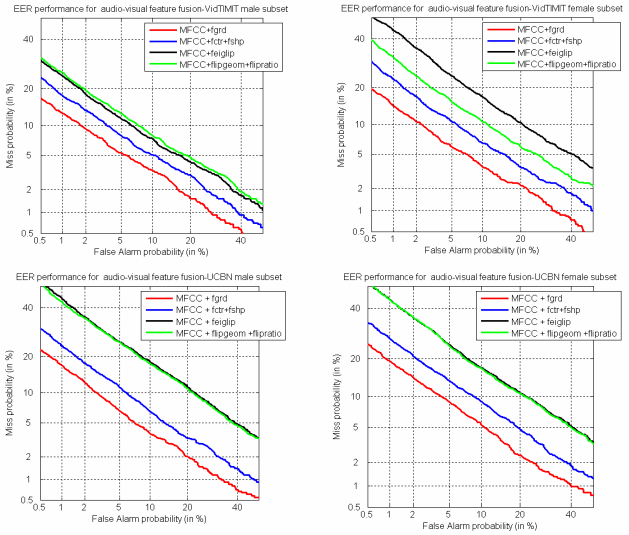

Figure 6 Audiovisual feature fusion for the SIV scenario.

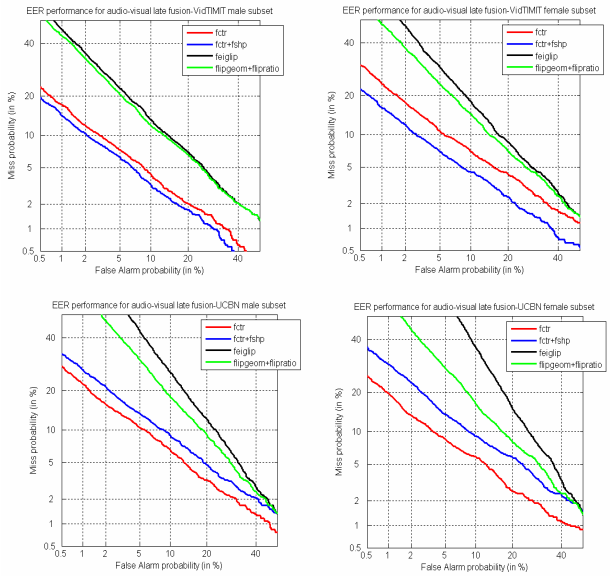

Figure 7 Audiovisual late fusion for the LV scenario.

Next, we performed liveness verification experiments to determine whether multimodal fusion of audio MFCC features and explicit lip motion features, provides any performance gain for the LV scenario. Similar to the SIV

scenario, the audiovisual fusion of explicit and implicit features resulted in synergistic fusion for LV scenario.

As can be seen in the DET curves in Figure 7, the late fusion of acoustic and visual lip motion features, improves the EER performance for both implicit and explicit lip motion features for the LV scenario. The performance gain with explicit fusion features is higher than for implicit fusion features, showing the better tracking of joint acoustic-lip dynamics by explicit features for modeling liveness.

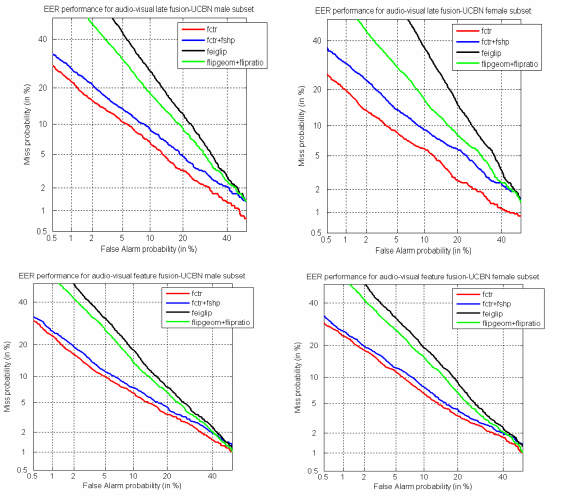

Figure 8 Audiovisual feature fusion for the LV scenario.

Finally, the EER performance with feature-level fusion for the LV scenario is shown in the DET curves of Figure 8. The fusion of audio and explicit lip motion features performs better compared with the fusion of audio and implicit lip motion features for the LV scenario. This could be due to the ability of explicit lip motion features to track the lip dynamics better.

Figures 7 and 8 show that for the LV scenario, feature fusion performs better than late fusion for both implicit and explicit lip motion features. Furthermore, the EER improvement achieved for explicit lip motion fusion is higher than for the implicit lip motion fusion features, which suggests that explicit lip motion features in feature fusion mode are more important for modeling liveness compared to modeling the identity of the speaker. Again, this might be due to better tracking of acoustic labial dynamics by explicit lip motion features in feature-level fusion mode compared to independent processing in late fusion mode.

5. Conclusion

The empirical results presented in this paper on audiovisual fusion based on lip motion features are quite promising, showing that the addition of a visual modality enhances the performance of a speaker identity verification system compared with audio-only SIV. It was also shown that explicit lip motion features perform better than implicit features and that the feature fusion approach of acoustic and lip motion features not only leads to better speaker identity verification performance, but also allows liveness verification to be performed. The use of explicit lip motion features and subsequent feature-level fusion with acoustic features allows continuous tracking of acoustic-labial dynamics for speaking faces. The fact that the fusion of audio and visual lip motion features yields different EER performance for the SIV scenario as compared to the LV scenario suggests that tracking the lip dynamics is more important for liveness verification than for speaker identity verification. The better performance of feature-fusion for the liveness verification task as compared to speaker identity verification further highlights the importance of the tracking of audiovisual speech dynamics for detecting static replay attacks. In conclusion, the results show that the addition of explicit lip motion features in conjunction with the feature fusion approach enhance the performance of the speaker identity verification system and equips the system with an ability to perform liveness checks.

6. References

[1] Poh, N., and J. Korczak, "Hybrid biometric person authentication using face and voice features," Proc. of Int. Conf. on Audio and Video-Based Biometric Person Authentication, Halmstad, Sweden, June 2001, pp. 348--353.

[2] Yehia Hani, Takaaki Kuratate, Eric Vatikiotic-Bateson, “Linking Facial Animation, Head Motion and Speech Acoustics”, Journal of Phonetics, Vol.30, Issue 3, 2002.

[3] Gary J. Sullivan, Pankaj Topiwala, and Ajay Luthra, “The H.264/AVC Advanced Video Coding Standard: Overview and Introduction to the Fidelity Range Extensions”, Proc. SPIE Conference on Applications of Digital Image Processing XXVII Special Session on Advances in the New Emerging Standard: H.264/AVC, August, 2004.

[4]E. Erzin, Y. Yemez, A. M. Tekalp, ``Multimodal Speaker Identification Using an Adaptive Classifier Cascade based on Modality Reliability," IEEE Transactions on Multimedia, Vol. 7, No. 5, pp. 840-852, October, 2005

[5] Chetty, G. and Wagner, M., “Automated lip feature extraction for liveness verification in audio-video authentication”, Proc. Image and Vision Computing 2004, New Zealand, pp 17-22.

[6] Nicolas Eveno, Alice Caplier, Pierre-Yves Coulon: “Jumping snakes and parametric model for lip segmentation”. ICIP (2) 2003: 867-870.

[7] Sanderson, C. and K.K. Paliwal, “Fast features for face authentication under illumination direction changes”, Pattern Recognition Letters 24, 2409-2419, 2003.

[8] Simon Lucey, Tsuhan Chen, Sridha Sridharan, Vinod Chandran, “Integration strategies for audio-visual speech processing: applied to text-dependent speaker recognition”. IEEE Transactions on Multimedia 7(3): 495-506, 2005.