interspeech 2008 incorporating SST 08

Audio-Visual Multilevel Fusion for Speech and Speaker Recognition

Girija Chetty and Michael Wagner

National Centre for Biometric Studies, University of Canberra, Australia

girija.chetty@canberra.edu.au , michael.wagner@canberra.edu.au

Abstract

In this paper we propose a robust audio-visual speech-and-speaker recognition system with liveness checks based on audio-visual fusion of audio-lip motion and depth features. The liveness verification feature added here guards the system against advanced spoofing attempts such as manufactured or replayed videos. For visual features, a new tensor-based representation of lip motion features, extracted from an intensity and depth subspace of 3D video sequences, is fused used with the audio features. A multilevel fusion paradigm involving first a Support Vector Machine for speech (digit) recognition and then a Gaussian Mixture Model for speaker verification with liveness checks allowed a significant performance improvement over single-mode features. Experimental evaluation for different scenarios with AVOZES, a 3D stereovision speaking-face database, shows favourable results with recognition accuracies of 70-90% for the digit recognition task, and EERs of 5% and 3% for the speaker verification and liveness check tasks respectively.

Index Terms: lip motion, biometrics, speaker verification, speech recognition, liveness verification.

1. Introduction

Most of the commercial speaker recognition systems currently deployed model a speaker based on unimodal audio information. In typical interactive applications with a need for robust speech and speaker recognition, current audio-based speech and speaker recognition systems degrade significantly in environments with low signal-to-noise ratios (SNR) [1]. Using a visual modality such as a 3D face or a 2D lip region in addition to the voice information, can make systems more robust against such noise degradation. However, visually- based speaker modelling is susceptible to pose and illumination variations, occlusion and poor image quality [2]. Moreover, 2D visual speech features of the lip region cannot on their own be used to model a speaker's face in its entirety, but need to be used along with other biometrics.

However, the lip region contains important liveness-related information, which can be used to detect fraudulent replay attacks with a still photo and audio recording of the speaker or a synthesized speaking face. Further, the audio and the lip motion during speech production are partially correlated, comprising mutually dependent (correlated) and mutually independent (uncorrelated) components. The use of 3D face dynamic information in addition allows better liveness checks as we can better quantify the differences between two persons' facial feature variations in 3D than in 2D face images. The facial movements during speech production comprise a complex sequence of muscle activations, and it is extremely difficult to imitate another person's facial speech gestures, which are highly characteristic to an individual [3, 4].

To enhance the robustness of the system against noise mismatch and fraudulent replay attacks, an approach based on multilevel fusion of audio-lip motion and depth information was proposed in one of our previous works [5]. In this paper, we extend this work further by reporting the investigations on a new representation of lip motion features – for a recognition paradigm involving speech (digit) recognition first, followed by speaker verification and liveness checks next.

The paper is organised as follows. The next section describes the audio and the new lip motion features. Section 3 describes the statistical speech and speaker modeling approaches used in this work. The details of the experiments and some results obtained are described in Section 4 and the paper concludes in Section 5 with conclusions and plans for further research.

2. Audio-Lip Motion Features

The lip-features were based on structure tensor approach involving eigenvalue analysis of the multidimensional structure tensor proposed by Bigun et al. in [9]. The method is briefly described below.

A line in the image plane translated with a certain velocity in the normal direction will generate a plane in the spatiotemporal image. The normal unit vector is denoted as $\boldsymbol{k} = (k_x, k_y, k_t)^T$ and the projection of the normal vector to the *x-y* coordinate axes represents the direction vector of the line's motion. The normal **k** of the plane will then relate to the velocity vector v_a as follows:

$$v = v_a = \frac{k_t}{k_x^2 + k_y^2}(k_x, k_y)^T$$

$$-\frac{1}{\left(\frac{k_x}{k_t}\right)^2 + \left(\frac{k_y}{k_t}\right)^2}\left(\frac{k_x}{k_t}, \frac{k_y}{k_t}\right)^T \quad (1)$$

where **v** is the optical flow normal. The optical flow normal estimation problem becomes a problem of solving the tilts ($\tan \gamma_1 = k_x / k_t$) and ($\tan \gamma_2 = k_y / k_t$) of the motion plane in the *x-t* and *y-t* manifolds, which is obtained from an eigenvalue analysis of the structure tensor [9]. Using complex numbers and smoothing, the angles of the eigenvectors are given effectively by:

$$\tilde{u}_1 = \iint \left(\frac{\partial f}{\partial x} + i\frac{\partial f}{\partial t}\right)^2 dxdt \quad (2)$$

and

$$\tilde{u}_2 = \iint \left(\frac{\partial f}{\partial y} + i \frac{\partial f}{\partial t} \right)^2 dxdt \quad (3)$$

Here, $\tilde{u}_1$ is complex valued ($i = \sqrt{-1}$) such that its magnitude is the difference between the eigenvalues of the local structure tensor in the *x-t* manifold, whereas its argument is twice the angle of the most significant eigenvector approximating $2\gamma_1$. The interpretation of $\tilde{u}_2$ is analogous to that of $\tilde{u}_1$. The function *f* represents the continuous local image, whose sampled version can be obtained from the observed image sequence. Thus, the *Total Least Square* estimations of γ_1 and γ_2 in the local 2D manifolds *x-t* and *y-t*, in the double-angle representation [10], leads to following estimated velocity components:

$$\frac{k_x}{k_t} = \tan \gamma_1 = \tan \left(\frac{1}{2} \arg(\tilde{u}_1) \right)$$

$$\Rightarrow \tilde{v}_x = \frac{\tan \gamma_1}{\tan^2 \gamma_1 + \tan^2 \gamma_2} \quad (4)$$

$$\frac{k_y}{k_t} = \tan \gamma_2 = \tan \left(\frac{1}{2} \arg(\tilde{u}_2) \right)$$

$$\Rightarrow \tilde{v}_y = \frac{\tan \gamma_2}{\tan^2 \gamma_1 + \tan^2 \gamma_2} \quad (5)$$

The tildes over v_x and v_y indicate that these quantities are estimates of v_x and v_y. With the calculated 2D-velocity feature vectors $(v_x, v_y)^T$ in each lip-region frame (128 × 128 pixels), we have dense 2D-velocity vectors.

For feature reduction a mean approximation is then performed. The 2D-velocity feature vectors $(v_x, v_y)^T$ at each pixel are quantized to three expected directions of the motion of 0, +45, and -45 degrees with all vectors pointing in the same direction then forming a particular region. The motion vectors within each region thus become real scalars, which take positive or negative signs depending on which direction they point to relative to their expected spatial directions:

$$f(p,q) = \left\| v_x(p,q), v_y(p,q) \right\| \times \mathrm{sgn}(\angle(v_x(p,q), v_y(p,q))) \quad (6)$$

with $0 \le p,q \le 127$.

Why use three spatial directions in six regions? This is because local lip motions are not completely free but must follow physical constraints. Mase and Pentland [11] and Yamamoto et al. [12] investigated lip articulation during speech by means of motion detection around different people's mouths. It is possible to conclude from these and other observations that the articulation of the lips progresses in a constrained manner during lip movement. For instance, when making the sound */o/*, the lip articulators deform so that the right and left sides of the mouth move toward each other while the upper and lower areas move up and down, respectively.

The next step is to quantize the estimated velocities from arbitrary real scalars to a more limited set of values. The quantized velocities are obtained from the data by calculating the mean value as follows:

$$g(l,k) = \sum_{p,q=0}^{N-1} f(Nl + p, Nk + q) \quad (7)$$

Here, *p,q=0 ...N-1* and *l,k = 0 ...(M-1),* where *N=10* and *M=12* represent the window size of the regions and number of regions respectively. The resulting mean values are used as 144-dimensional (*M*×*M*) feature vectors containing the statistics of lip-motion. It is worth noting that the original dimension before reduction is 128×128×2 = 32,768.

The method for extracting normal velocity features (structure tensors) for the intensity images is used for depth images obtained from the 3D AVOZES database, resulting in two sets of lip features, lip motion-intensity tensor features f_{litf} and lip-motion-depth tensor features f_{ldtf}. The details of extracting the depth image for AVOZES database faces are described in several works on 3D face reconstruction from stereo faces [6].

Mel-frequency cepstral coefficients *(MFCC)* were used to represent speakers' voice information. The voice signal is first pre-emphasized and divided into one 25-ms frame every 10 ms. A Hamming window is applied and a mel-warped log-amplitude spectrum is computed for each frame. Audio vectors are determined from the mel spectra by a discrete cosine transform and cepstral filtering. Each vector comprises 12 cepstral coefficients and the normalized log energy plus first and second derivatives (delta coefficients and delta-delta coefficients) for a total 39 components. The new lip motion features (f_{litf}, f_{ldtf}), were then fused with the acoustic features (f_{mfcc}) and used for building the speaker and digit models for speaker and speech recognition. Since the audio rate is twice the video frame rate, video frames were upsampled by 2 to match the audio frame rate.

3. Speech and Speaker Models

The speech and speaker models were built using Gaussian Mixture Models (GMM) and Support Vector Machine (SVM) models. A brief description of GMM and SVM is given below.

3.1. GMM

This statistical model can be understood as a weighted sum of multivariate Gaussian distributions:

$$p(x \mid \lambda) = \sum_{j=1} p_j b_j(x) \quad (8)$$

Here, $\boldsymbol{x}$ is a D-dimensional feature vector, and p_j and $b_j(x)$ represent the mixture weights and multivariate Gaussian component densities, respectively. The weights p_j represent the probability that identity λ is represented by features from a specific region of the feature space. For speech recognition, we build phoneme level models using a hidden Markov model (HMM) with five states and three mixtures in each state.

3.2. SVM

The SVM formulation is based on the Structural Risk Minimization principle, which minimizes an upper bound on the generalization error, as opposed to Empirical Risk Minimization [10, 17]. An SVM is a discrimination-based

binary method using a statistical algorithm. It is frequently used in pattern-recognition and information-retrieval tasks because of its ability to generalize well. The background idea in training an SVM system is finding a hyperplane $\boldsymbol{w}.\boldsymbol{x}+b=0$ as a decision boundary between two classes. Further details of SVM are given in [17].

When conducting speech-classification experiments, we need to choose between multiple classes. The best method of extending the two-class classifiers to multiclass problems appears to be one against all, consisting of building SVM classifiers equal to the number of classes. We train each SVM with one of the classes against the rest of the classes. The one-against-one approach simply constructs, for each pair of classes, an SVM classifier that separates those classes. All tests here were performed for the speech features only (f_{mfcc}), the visual features only ($f_{litf} + f_{ldtf}$) and the fused audio-visual features ($f_{mfcc} + f_{litf} + f_{ldtf}$).

4. Experiments and Results

The AVOZES 3D stereovision database [6] was used for all the experiments described in this paper. Different sets of experiments were performed to quantify the performance of the new lip motion (intensity tensor and depth tensor) features, and the fused audio and lip motion features. Also, different protocols were used for speech (digit) recognition, speaker verification, and liveness checks, as described in the next subsections.

4.1. SVM Based Speech Recognition

For this set of experiments, we specified our own protocol *(Protocol B)*, and we used the digit subset of the AVOZES database. The digit sequences were segmented manually to extract each digit, and single-digit HMM models for digits 0-9 were built. During the manual segmentation, we found that the words 4, 5, and 8 contained less visual information than the other digits. The training and test groups comprised 10 speakers each, and the training data were separate from test data.

Table 1: Speech (digit) recognition accuracy for all digits using Protocol B

	Features			
Digit	f_{mfcc}	$f_{litf}+f_{ldif}$	$f_{mfcc}+f_{litf}$	$f_{mfcc}+f_{litf}+f_{ldif}$
0	69%	60%	82%	83%
1	80%	67%	90%	93%
2	76%	70%	79%	81%
3	80%	65%	86%	88%
4	79%	45%	75%	79%
5	80%	40%	73%	75%
6	90%	80%	90%	94%
7	88%	85%	90%	95%
8	81%	44%	73%	76%
9	80%	39%	75%	78%

We then performed speaker-independent speech recognition (digit identification) according to this protocol. Table 1 shows that we obtained the best recognition accuracy of 75-95% for the fused features ($f_{mfcc} + f_{litf} + f_{ldtf}$), compared with 69-88% for audio-only features f_{mfcc} and 39-85% for visual-only features ($f_{litf} + f_{ldtf}$).

These results vary considerably, one cause being that there is insufficient information (especially visual information) for certain digit utterances. This is not surprising given the small size of the AVOZES database, which is a pilot 3D stereovision audio-visual database, but not a full-fledged database for conclusive speech and speaker recognition experiments. More importantly, the utility of the visual information for the digit recognition varies considerably between the digits depending on the distinctiveness of the visemes. The audio-visual fusion for some digits was catastrophic with digits 5, 8 and 9 giving worse identification accuracy in the fusion modes than with either the audio or the video input. Those two factors have negatively influenced fusion performance, which presently assumes that the quality of information is uniform.

During the manual segmentation, we could verify that the digits 5 and 8 were pronounced with shorter duration by most of the speakers in the database, and hence there was notably less visual data than for other digits.

4.2. GMM Based Speaker Verification

We specified a different protocol (*Protocol A*) for this set of experiments, which was inspired by the Lausanne protocol (Configuration 1) as defined by the M2VTS consortium for standardizing person-recognition experiments. It suggests splitting the database into training, evaluation, and test groups. The evaluation set is used to quantify client and impostor access performance after training. It is used to find the threshold for accepting or rejecting a person at predefined operation points. Finally, the test data is used to quantify how well the audio and visual features perform with respect to the desired performance once the thresholds are fixed. In our Protocol A, the training group contained 10 subjects as clients, the evaluation group contained an additional 5 subjects as impostors, and the testing group contained yet another 5 subjects as impostors.

For clients, we used the AVOZES Module-4-CVC-subset for training, the Module-4-VCV-subset for evaluation, and the Module-5-digit-subset for testing. From the AVOZES database documentation [6], these three subsets appear to be recorded in different sessions, making them suitable as training, evaluation and test sets for the speaker verification tasks. Further, we manually segmented the actual CVC-VCV words from all the sentences in Module 4 and the actual digits from the sentences in Module 5. This left the carrier phrase "*You grab beer*", resulting in all the training, evaluation and test groups to be the same, hence facilitating text-dependent speaker recognition tests to be performed.

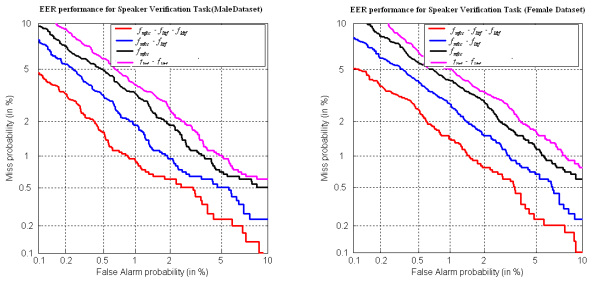

Figure 1. DET curves for speaker verification: (a) males; (b) females. The curves from top-right to bottom-left are: magenta = $f_{litf}+f_{ldif}$, black = f_{mfcc}, blue = $f_{mfcc}+f_{litf}$, red = $f_{mfcc}+f_{litf}+f_{ldif}$.

Figure 1 shows the DET curves for GMM-based speaker verification for the audio, visual and fused audio-visual features for the male and female subgroups. The equal-error rates (EER) were 2.0-2.5% for visual-only features (f_{litf} + f_{ldtf}), 1.9-2.3% for audio-only features (f_{mfcc}), and 0.9-1.2% for fused audio and lip motion features (f_{mfcc} + f_{litf} + f_{ldtf}). Thus the fusion of audio with tensor-based lip motion features resulted in synergistic fusion for this scenario.

4.3. GMM Based Liveness Checks

For this scenario, we extended Protocol A for testing the liveness assurance performance of the new visual features. We created a fraudulent-speaker video database for the speakers in AVOZES Module 4 using a set of computer animation tools [5] and used it for testing. The synthetic speaking-face videos in the fraudulent database emulate a sophisticated-replay-attack scenario. More details of our liveness check protocols are described in some of our previous works [5].

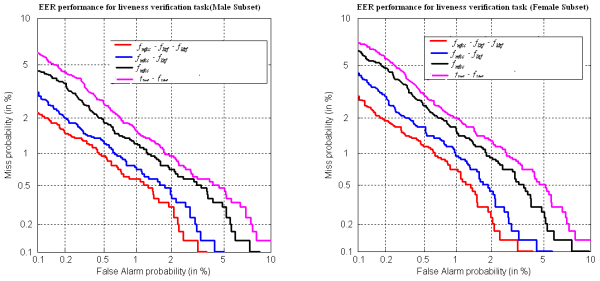

Figure 2. DET curves for liveness verification: (a) males; (b) females. The curves from top-right to bottom-left are: magenta = f_{litf}+f_{ldif}, black = f_{mfcc}, blue = f_{mfcc}+f_{litf}, red = f_{mfcc}+f_{litf}+ f_{ldif}.

For this scenario, the gender-specific client models were built using the training data from the AVOZES Module-4-CVC-subset and the thresholds were determined using the Module-4-VCV-subset – similarly to the speaker verification scenario. For testing, the artificially created fraudulent-speaking-face videos were used. The DET curves in Figure 2 show the performance achieved for audio, lip-motion and fused audio-visual features for this scenario.

The EERs were 1.3-1.4% for visual-only features (f_{litf} + f_{ldtf}), 1.1-1.3% for audio-only features (f_{mfcc}), and 0.6-0.8% for fused audio and lip motion features (f_{mfcc} + f_{litf} + f_{ldtf}). Thus the fusion of audio with tensor based lip motion features for this scenario also resulted in synergistic fusion.

5. Conclusions

In this paper we have proposed a robust audio-visual speech-and-speaker-recognition system with liveness checks, based on audio-visual fusion of audio features and visual lip-motion features. The liveness verification feature added here guards the system against advanced spoofing attempts such as replayed or synthesised videos. New tensor-based representations of lip motion features are extracted from the intensity and depth subspaces of 3D video sequences and used as visual feature vectors, which are then fused with the audio features. This new representation of lip features allows better modeling of the visual speech dynamics. First, a support vector machine was used for speech (digit) recognition, and then a Gaussian Mixture Model was used for speaker verification with liveness checks. Experimental evaluation on a 3D stereovision speaking-face database, AVOZES, show favourable results for different scenarios with a recognition accuracy of 75-95% for a speech (digit) recognition task, and equal-error rates of 0.9-1.2% and 0.6-0.8% for speaker verification and liveness check tasks respectively.

6. References

[1] G. G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. W. Senior, "Recent Advances in the Automatic Recognition of Audiovisual Speech," Proceedings of the IEEE, vol. 91, pp.1306-1324, Sept. 2003.

[2] R. Brunelli and D. Falavigna, "Person Identification Using Multiple Cues," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 17, pp. 955-966, Oct. 1995.

[3] A. Santi, P. Servos, E. Vatikiotis-Bateson, T. Kuratate & K. Munhall (2003). "Perceiving Biological Motion: Dissociating Talking from Walking". Journal of Cognitive Neuroscience. 15, 800-809.

[4] D. Callan, J.A. Jones, K.G. Munhall, C. Kroos, A. Callan & E. Vatikiotis-Bateson (2003) "Neural Processes Underlying Perceptual Enhancement by Visual Speech Gestures," Neuroreport, 14, 2213-2218.

[5] G. Chetty and M. Wagner, "Audio-Visual Speaker Identity Verification Using Lip Motion Features, Proc. Interspeech-2007.

[6] R. Goecke and J.B. Millar. "The Audio-Video Australian English Speech Data Corpus AVOZES," Proceedings of the 8th International Conference on Spoken Language Processing INTERSPEECH 2004 - ICSLP, Volume III, pages 2525-2528, Jeju, Korea, 4 - 8 October 2004. University Science, 1989.

[7] B. Horn and B. Schunck, "Determining Optical Flow," J. Artificial Intelligence, vol. 17, no. 1, pp. 185-203, 1981.

[8] B.D. Lucas and T. Kanade, "An Iterative Image Registration Technique with an Application to Stereo Vision," Proc. Int'l Joint Conf. Artificial Intelligence, pp. 674-679, 1981.

[9] J. Bigun, G. Granlund, and J. Wiklund, "Multidimensional Orientation Estimation with Applications to Texture Analysis of Optical Flow," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 13, no. 8, pp. 775-790, Aug. 1991.

[10] G.H. Granlund, "In Search of a General Picture Processing Operator," Computer Graphics and Image Processing, vol. 8, no. 2, pp. 155-173, 1978.

[11] K. Mase and A. Pentland, "Automatic Lip-Reading by Optical-Flow Analysis," Systems and Computers in Japan, vol. 22, no. 6, pp. 67-76, 1991.

[12] E. Yamamoto, S. Nakamura, and K. Shikano, "Lip Movement Synthesis from Speech Based on Hidden Markov Models," J. Speech Comm., vol. 26, no. 1, pp. 105-115, 1998.

[13] D. Reynolds and R. Rose, "Robust Text-Independent Speaker Identification Using Gaussian Mixture Models," IEEE Trans. Speech and Audio Processing, vol. 3, no. 1, pp. 72-83, 1995.

[14] M. Schmidt and H. Gish, "Speaker Identification via Support Vector Classifiers," Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing (ICASSP '96), pp. 105-108, 1996.

[15] V. Wan and W. Campbell, "Support Vector Machines for Speaker Verification and Identification," Proc. IEEE Signal Processing Soc. Workshop Neural Networks for Signal Processing X, vol. 2, pp. 775-784, 2000.

[16] V.N. Vapnik, The Nature of Statistical Learning Theory. Springer, 1995.

[17] C.J. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," Data Mining and Knowledge Discovery, vol. 2, no. 2, pp. 121-167, 1998.