FUSION OF TALKING FACE BIOMETRIC MODALITIES FOR PERSONAL IDENTITY VERIFICATION

Ulises Ramos Sanchez and Josef Kittler

Centre for Vision, Speech and Signal Processing
University of Surregy
Guildford GU2 7XH
United Kingdom
J.Kittler@surrey.ac.uk

ABSTRACT

We describe a personal identity verification system based on lip dynamics biometric. The lip shape is represented in terms of a B-spline model, tracked over time. The coordinates of the 11 control points of the B-spline model are used as features for each frame. An utterance consisting of N frames produces a sequence of 22 dimensional feature vectors that is matched to the template using dynamic time warping. The verification error rate achived by the systems on the XM2VTS database is about 14%. By fusing the system with face and voice biometrics the error rate is reduced to a fractiopn of one percent.

1. INTRODUCTION

References to lip-reading applications for speech recognition and synthesis are abundant in the literature but, other than the work done by Lüttin [1] or Mason *et al* (*e.g.* [3]), there does not seem to be that much research on lip-based speaker verification (recognition). However, as suggested in [2], even very coarse lip features can be used as behavioural biometric characterisation of the speaker or as a means for detecting the lip shape status which in turn can serve as a control information for face coding or recognition. The latter was demonstrated in [9], where a B-spline lip tracking system was used to provide control information regarding the state of the lip shape which is used by a conventional eigenface-based face verification system to confirm or reject a claimed personal identity. The performance of the system tested on the M2VTS database [4] showed a promising improvement over the unimodal approach. This improvement derives from the achieved reduction in the population entropy of the models, thus minimising the probability of impostor acceptance.

In this paper it will be shown how the information supplied by the lip tracker can be advantageously used to implement a text-dependent speaker verification system based exclusively on lip shape features. This verification modality is thereafter combined with other visual and vocal experts, resulting in improved overall performance. The experiments reported in this section were carried out in the XM2VTS database [5] according to the Lausanne Protocol [6]. The results obatined are illustrative of the kind of improvements to expect.

This project was partially supported by EU Project Biosecure

2. DESCRIPTION OF THE XM2VTS DATABASE AND THE LAUSANNE PROTOCOL

The XM2VTS database contains synchronised image and speech data as well as sequences with views of rotating heads. The database includes recordings of 295 subjects taken at one month intervals. In each session two recordings were made, each one consisting of a speech shot and a head rotation shot. The speech shot consisted of a frontal face recording of each subject during speech production, namely the utterance of three speaking sequences: two digit sequences, and a sentence.

The Lausanne protocol is a published evaluation proposal for the XM2VTS database and two protocol configurations were defined.

- Configuration I: The assumption is good expert training using data from three different sessions, and inferior fusion training using data from the same shots that were used for expert training
- Configuration II: The assumption is inferior expert training using data from only two different sessions, and good fusion training using data from shots that were not used for expert training

Each shot being used consists of the 2 audio digit sequences and of one image. The 295 subjects were divided into three sets: 200 clients, 25 impostors for evaluation, and 70 impostors for independent testing. The impostors in the evaluation set allow to train a supervisor with impostors that were never seen by the experts. The evaluation set serves for the evaluation of experts, the determination of the verification threshold, and for the training of the supervisor.

This leads to the following statistics:

- Client training examples: 3 per client in Configuration I, 4 per client in Configuration II.
- Evaluation samples (clients): 600 in Configuration I, 400 in Configuration II.
- Evaluation samples (impostors): 40000 ($25 \times 4 \times 2 \times 200$).
- Test client accesses: 400 (200×2).
- Test impostor accesses: 112000 ($70 \times 4 \times 2 \times 200$).

3. EXTRACTION OF LIP FEATURES AND MATCHING STRATEGY

The lip tracker described was used to extract lip features from the two audio digit sequences available in each shot. In the current text dependent lip-based verification that will be described below, the two audio sequences are concatenated, resulting in a single, bigger sequence. The lip tracking initialisation was based on a colour clustering procedure.

As far as tracking performance itself is concerned, its importance is very much acknowledged, but lacking an objective and meaningful quality metric, such an analysis will be omitted and, as previously pointed out by Jourlin *et al* [7], it is the combined performance of tracking and verification that will be evaluated through the verification experiments that will be described in the coming sections. Nonetheless, the subjective impression is quite good for most of the speakers, which is quite remarkable in view of the broad ethnical background coverage of the XM2VTS, and the fact of having to cope occasionally with significant motion. The results also show the good generalisation capability of the eigenlips estimated from another database. Problems have, however, been detectd a) with some speakers wearing dark beards and/or moustaches, b) with speakers where lip colour is hardly distinguishable from the surrounding skin, specially if relatively reddish areas occur in the surrounding skin area, and c) unusual degree of motion. The presence of surrounding skin together with moustaches/beards already violates the working assumption of estimating the lip contour as the boundary between two homogeneous regions, with the added difficulty in these cases that the estimation of a unimodal colour model of the area surrounding the lips (by merging quite different chromaticiy clusters results into a single one) results in a model which happens to be closer to the lips model than any of the constituting clusters. In case a), although tracking remained stable it failed to accurately follow the lip contour outline. Case b) was less severe, and tracking failure was restricted more often to temporary distractions, typically involving the lower contour of the lips. As far as case c) is concerned, even significant degrees of motion were generally well tolerated, although in a few cases allowing for a temporary, partial loss of tracking (for instance affecting a corner of the mouth), prior to recovery.

The lip tracker supplies a set of eigenlip coefficients and affine transform parameters for each frame. By warping the linear combination of eigenlips with the affine transform parameters, a 22-dimensional feature vector is obtained that consists of the geometrical coordinates of the 11 control points used for characterising the lip contour. Accordingly, an utterance consisting of N frames is represented by a sequence of control point vectors $\mathbf{u}_1, \ldots, \mathbf{u}_N$ which define a trajectory in a 22-dimensional space. Verification tests are operated by matching the trajectory under test $T = \{\mathbf{u}_1, \ldots, \mathbf{u}_N\}$ against a reference template $R = \{\mathbf{v}_1, \ldots, \mathbf{v}_M\}$ corresponding to the claimed identity using a Dynamic Time Warping Algorithm (DTW) [8].

A framewise dissimilarity metric is given by

$$d(\mathbf{u}_i, \mathbf{v}_j) = (\mathbf{u}_i - \mathbf{v}_j)^T \mathbf{H} (\mathbf{u}_i - \mathbf{v}_j) \quad (1)$$

where $\mathbf{H}$ is the metric matrix converts control point distances into real shape distances.

Unless additional information becomes available, it is generally not possible to establish whether affine variation just corresponds to pose variations (*e.g.* different lips size due to posing at a different distance from the camera) or they are really characteristic of a given identity. Furthermore, even small, perceptually irrelevant changes in scale can have a stronger impact on the metric considered than proper shape variation. This is why eventually all control point vectors are normalised for translation, scale and rotation. Hence the control point constellation for each frame is translated to the origin of coordinates and the point set is rotated so that the points corresponding to the mouth corners on the B-spline contour are aligned with the x axis. Finally, to account for scale normalisation, whilst allowing for relative size changes in the sequence of frames corresponding to a single utterance, the mouth width mode $w_0 = mod_{i=1}^{N} w_i$ is computed in a first pass and then the control point constellation of each frame is homogeneously (both horizontally and vertically) scaled by a factor $\frac{w_{ref}}{w_0}$ in a second pass, where w_{ref} is a predefined width value. As a result of this process, the mouth shapes of every single utterance will have their most common width set to this reference value.

The distribution of matching scores for client and imposter claims obtained on the evaluation set of the XM2VTS database are shown in Figure 1. The error rates for Protocol Configuration I are presented in Table 1.

4. FUSION EXPERIMENTS ON THE XM2VTS DATABASE ACCORDING TO THE LAUSANNE PROTOCOL

The combination of a number of experts can potentially improve -and sometimes significantly- the performance attained

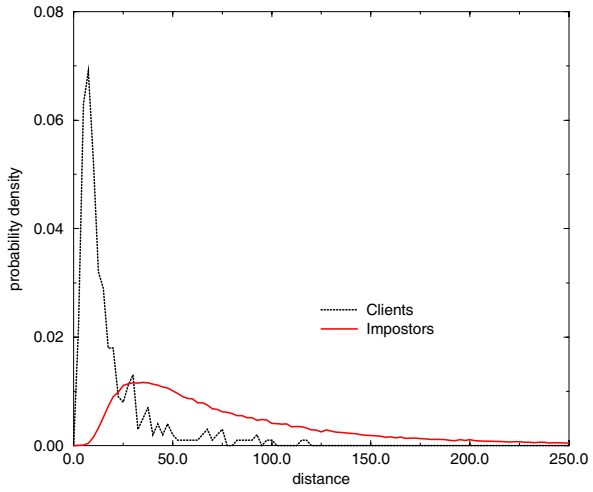

Fig. 1. Distance probability density functions for client and imposter claims.

by the best individual modality. This was the theme of the M2VTS project [11], and more recently, also in the context of the XM2VTS database and the Lausanne protocol, succesful integration result were reported in [12].

The fundamentals underpinning fusion is that by drawing on several independent sources of information, an adequate combination of them can overcome the shortcomings and limitations of each of the individual modalities. The converse is also possible (performance degradation) and from that it follows the importance of developing appropriate information integration strategies.

Typical fusion strategies consist of simple combination rules: maximum, minimum, median, average score, and product of scores. Conditions under which such schemes perform well are theoretically understood and have been shown to hold in applications [13]. However, in a very similar fusion scenario [12] (in fact some of the experts combined are also used here) combining high performance speech verification modules and a medium vision module (face recognition), the conditions were violated and none of the aforementioned fusion schemes performed better than the best individual expert.

In the light of those considerations, and the succesful performance obtained with a linear weighted combination rule [7], this was eventually the fusion strategy adopted for these experiments. According to this integration paradigm, a verification score v is obtained as a linear combination of the scores of the m modalities to fuse ($v = w_1 v_1 + \ldots + w_m v_m$) and then compared with a threshold τ_0. The optimal weights $w_1 \ldots w_m$ and the acceptance threshold τ_0 are chosen using the evaluation set.

Algorithm	threshold	FRR	FAR
SURREYL (lips)	0.50	14.00 %	12.67 %
SURREY1 (face)	0.50	7.25 %	7.78 %
SURREY2 (face)	0.21	5.00 %	4.45 %
IDIAP2 (voice)	0.50	7.00 %	1.42 %
IDIAP3 (voice)	0.50	0.00 %	1.48 %
AUT1 (face)	0.50	6.00 %	8.12 %

Table 1. Performance of modalities on test set (Configuration I).

Apart from the described DTW-based text dependent verification system based on lip features (SURREYL), 3 face recognition algorithms (SURREY1: based on robust correlation [15], SURREY2: Linear Discriminant Analysis [14] and AUT1 developed at the Aristotle Technical University), and 2 voice-based modalities [12] (IDIAP2 -sphericity and IDIAP3 -HMMs) were considered for the fusion experiments. Their individual performances (Configuration I) on the test set are shown in Table 1.

The following fusion experiments were considered:

1. Lips and face (SURREY2)
2. Lips and voice (IDIAP3)
3. Face (SURREY2) and voice (IDIAP3)
4. Lips, face (SURREY2) and voice (IDIAP3)
5. SURREY1, SURREY2, IDIAP2, IDIAP3 and AUT1
6. All: SURREYL, SURREY1, SURREY2, IDIAP2, IDIAP3 and AUT1

The results, as well as the corresponding optimal combination weights and acceptance threshold can be seen in Table 2. It is interesting to see how in all cases the trained linear weighted classifier performs better than the best individual expert. It is also worth remarking how the 4th fusion strategy (lips, face and voice) does perform slightly worse than the 3rd one (face and voice), which can be put down to overtraining since the former did yield a lower FAR figure during evaluation. Eventually the best results among the 6 test scenarios considered are obtained when all 6 modalities are combined, although it can be seen how the weights attributed to some of them are quite low, or even zero. In order to see to what extent lip features do represent a positive contribution to the overall performance, the results for a trained classifier combining the other 5 modalities, leaving aside the lips, are shown as well. Getting further improvements at low error rates is very difficult and lips reduce the error rate of the 5 modality case by roughly 40%.

Modalities	weights	threshold	FRR	FAR
lips and face	0.42, 0.58	0.38	4.50 %	0.73 %
lips and voice	0.41, 0.59	0.54	0.00 %	1.39 %
face and voice	0.58, 0.42	0.42	0.00 %	1.25 %
lips, face and voice	0.27, 0.23 0.49	0.51	0.00 %	1.31 %
5 modalities (no lips)	0.00, 0.02 0.87, 0.05 0.06	0.50	0.00 %	0.52 %
all 6 modalities	0.03, 0.00 0.01, 0.89 0.03, 0.04	0.50	0.00 %	0.29

Table 2. Fusion results (Configuration I).

5. CONCLUSIONS

A text-dependent DTW-based person identity verification system using lip features during speech production has been presented. The system builds upon the tracking results generated by a shape-constrained chromaticity-based lip tracker which was run for the more than two thousand audio sequences of the XM2VTS database.

The verification performance of this lip-based modality was tested according to the Lausanne protocol, with error rates of about 14% on average in both configurations. More importantly, it has been demonstrated, how a 'weak' verification modality brings in additional discriminatory information that can result in improved overall performance when combined with other verification experts. Experiments carried out with a trained weighted linear classifier combining different verification modalities (face, voice, lips) showed, in all cases considered, better verification performance than the best individual modality being combined.

6. REFERENCES

[1] J. Lüttin. *Visual Speech and Speaker Recognition*. PhD thesis, University of Sheffield, 1997.

[2] M. U. Ramos Sánchez, J. Matas, and J. Kittler. Lip shape modelling and tracking for security and video coding applications. In *Seventh Spanish Symposium on Pattern Recognition and Image Analysis*, pages 73–78, 1997.

[3] J. S. D. Mason, J. Brand, R. Auckenthaler, F. Deravi, and C. Chibelushi. Lip signatures for automatic person recognition. In *IEEE Third Workshop on Multimedia Signal Processing*, pages 457–462, 1999.

[4] S. Pigeon and L. Vandendorpe. The M2VTS multimodal face database. In *International Conference on Audio- and Video-Based Biometric Person Authentication*, pages 403–409, 1997.

[5] K. Messer, J. Matas, J. Kittler, J. Lüttin, and G. Maître. XM2VTS: The extended M2VTS database. In *Second International Conference on Audio and Video-Based Biometric Person Authentication*, 1999.

[6] J. Lüttin and G. Maître. Evaluation protocol for the XM2FDB database (Lausanne Protocol). Technical report, IDIAP, 1998.

[7] P. Jourlin, J. Lüttin, D. Genoud, and H. Wassner. Acoustic-labial speaker verification. *Pattern Recognition Letters*, 18(9):853–858, 1997.

[8] M. Pandit and J. Kittler. Feature selection for a DTW-based speaker verification system. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 769–772, 1998.

[9] J. Kittler, Y. P. Li, J. Matas, and M. U. Ramos Sánchez. Lip-shape dependent face verification. In *International Conference on Audio and Video-Based Biometric Person Authentication*, pages 61–68, 1997.

[10] J. L. Wayman. Error-rate equations for the general biometric system. *IEEE Robotics & Automation Magazine*, 61:35–48, March 1999.

[11] G. Richard, Y. Mengay, I. Guis, N. Suaudeau, J. Boudy, P. Lockwood, C. Fernandez, F. Fernandez, C. Kotropoulos, A. Tefas, I. Pitas, R. Heimgartner, P. Ryser, C. Beumier, P. Verlinde, S. Pigeon, G. Matas, J. Kittler, J. Bigün, Y. Abdeljaoued, E. Meurville, L. Besacier, M. Ansorge, G. Maitre, J. Luettin, S. Ben-Yacoub, B. Ruiz, and K. Aldama. Multi modal verification for teleservices and security applications (M2VTS). In *IEEE International Conference on Multimedia Computing and Systems*, volume 2, pages 1061–1064, 1999.

[12] S. Ben-Yacoub, J. Lüttin, K. Jonsson, J. Matas, and J. Kittler. Audio-visual person verification. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 580–585. IEEE, 1999.

[13] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239, March 1998.

[14] Y. Li, J. Kittler, and J. Matas. Effective implementation of linear discriminant analysis for face recognition and verification. In *8th International Conference on Computer Analysis of Images and Patterns*, pages 234–242, 1999.

[15] J. Matas, K. Jonsson, and J. Kittler. Fast face localisation and verification. In *British Machine Vision Conference*, volume 1, pages 152–161, 1997.