IDIAP RESEARCH REPORT

# Joint Bi-Modal Face and Speaker Authentication using Explicit Polynomial Expansion

Sébastien Marcel [a]

IDIAP–RR 07-14

March 2007

submitted for publication

[a] IDIAP Research Institute

**IDIAP Research Institute** **www.idiap.ch**
Rue du Simplon 4 P.O. Box 592 1920 Martigny – Switzerland
Tel: +41 27 721 77 11 Fax: +41 27 721 77 12 Email: info@idiap.ch

IDIAP Research Report 07-14

# JOINT BI-MODAL FACE AND SPEAKER AUTHENTICATION USING EXPLICIT POLYNOMIAL EXPANSION

Sébastien Marcel

MARCH 2007

SUBMITTED FOR PUBLICATION

# 1 Introduction

Multi-modal person authentication methods are mainly based on fusion algorithms (merging the output of many biometric modules). Most of the proposed fusion techniques operate at the score or decision levels only. However, there should be clear correlation, at the frame level, between the video and audio streams.

The principal problem is to deal with asynchronous streams of data. Let us note, $x_1^T$ the sequence of audio frame features of length $T$ where $x_t \in \mathbb{R}^m$, $y_1^S$ the sequence of video frame features of length $S$ where $y_s \in \mathbb{R}^n$. Audio and video frame features (1) are not synchronised, (2) doesn't have the same length (here we assume $S < T$) and (3) doesn't have the same frame size. Recently, Asynchronous Hidden Markov Models (AHMMs) have been proposed [1]. An AHMM takes into account temporal correlations jointly between the audio and video streams, and has shown to be robust to noise compared to other techniques.

In this extended abstract, we propose an alternative technique for modelling jointly audio and video streams at the feature level using an explicit polynomial expansion kernel. We concatenate audio and video frame features after the kernel expansion and then perform the classification using a linear SVM. Experiment results have been performed on a audio/visual database and compared to other techniques.

# 2 The Proposed Approach

We propose to merge the audio and video frame features at the access level. We will adapt the polynomial expansion technique proposed in [2] for speaker authentication to the case of two streams of information. The two streams would thus be concatenated at the access level, but in a very high dimensional space, thanks to the kernel expansion.

Let us note $\Psi^{audio}(x)_1^T$ and $\Psi^{video}(y)_1^S$ the polynomial expansion sequences of audio/video frame features $x_1^T$ and $y_1^S$, where $\Psi^{audio}$ and $\Psi^{video}$ are polynomial expansion functions such that $\Psi^{audio} : \mathbb{R}^m \mapsto \mathbb{R}^M$ and $\Psi^{audio} : \mathbb{R}^n \mapsto \mathbb{R}^N$ with $M >> m$ and $N >> n$. At the access level, audio and video frames features become $\hat{x} = \frac{1}{T}\sum_1^T \Psi^{audio}(x)_t$ and $\hat{y} = \frac{1}{S}\sum_1^S \Psi^{video}(x)_s$.

Finally, the decision is taken on the basis of the joint audio/video vector $\hat{z} = (\hat{x}, \hat{y})$. [2] proposed to use Support Vector Machine (SVM) for that task. However, any other statistical machine learning algorithm can be used for that task such as Multi-Layer Perceptrons (MLPs).

# 3 Results

We report (Table 1) baseline results (GMM, AHMM) on the M2VTS database according to the original protocol described in [1], and we performed experiments using the proposed approach on the same database and using the same features (lips features of frame size $48$ for video, and voice features of frame size $33$ for audio). We separately tested lips and voice

Table 1: Comparative results obtained on the M2VTS using baseline systems (GMM, AHMM) and the proposed approach. In the notation Poly $^d_m$, $d$ is the degree of the polynomial expansion and $m$ is the number of features after expansion.

| System | HTER (4-fold average) |
|---|---|
| AHMM [1] | 13.05 |
| GMM voice [1] | 3.73 |
| Poly $^2_{1225}$ lips SVM | 9.35 |
| Poly $^2_{595}$ voice SVM | 1.86 |
| Poly $^2_{1820}$ lips+voice SVM | 7.8 |

with different polynomial degree (but we report here only the best configuration) and then jointly by concatenating the features.

# 4 Conclusion

We proposed an alternative technique for modelling jointly audio and video streams at the feature level using an explicit polynomial expansion kernel. Results have shown that the polynomial expansion technique performs very well on voice (coherent with [2]), and that the proposed approach performs better than AHMMs.

We hope that this work will open new research directions in joint bi-modal authentication, such as the joint extraction of both video and audio features using expansion techniques, or the use of other facial features in addition to lips, or the use of alternative kernels.

# References

[1] Bengio, S.: Multimodal authentication using asynchronous HMMs. In: 4th International Conference on Audio- and Video-Based Biometric Person Authentication, AVBPA, Springer-Verlag (2003) 770–777

[2] Campbell, W.M.: A sequence kernel and its application to speaker recognition. In Dietterich, T.G., Becker, S., Ghahramani, Z., eds.: Advances in Neural Information Processing Systems 14, Cambridge, MA, MIT Press (2002) 1157–1163