Aliveness Detection Using Coupled Hidden Markov Models

Enrique Argones Rúa[1], Carmen García Mateo[1], Hervé Bredin[2], and Gérard Chollet[2] *

[1] UVigo, ST Group, STC Dept., Vigo (Spain)
[2] GET-ENST, Dépt. TSI, Paris (France)

Abstract. A biometric system must verify the identity of a person. Furthermore, it should ensure that the biometric cues have actually been acquired from that person at the moment of the identity verification. The aliveness check ensures that the acquired biometric cue is actually acquired from a live person actually present at the time of capture. This paper compares different techniques to check the aliveness by measuring the synchrony between speech and lip movement in an audio-visual framework. This statistical relationship between speech and lip movement is checked with four different statistical tests based on coupled hidden Markov models.

1 Introduction

It is well known that oral communication between people is intrinsically multimodal. Speech, lip movements and even gestures can help to understand the message. Since gestures usually depend a lot on the person who is talking, the useful information about the message is usually concentrated on the speech itself and on the lip movements. Blind people can obviously understand speech since they can listen to it, and deaf people can understand speech since they can lip-read.

Multimodal biometric systems based on face verification and speaker verification usually make a score level fusion of the face expert and speaker expert outputs. Nonetheless, some of them try to use visual speech information to improve the overall verification performance [1]. One of the major weakness in multimodal biometric systems based on face verification and speaker verification is that they do not take into account realistic impostor attacks scenarios. If a previously recorded segment of speech uttered by the user is used jointly with a photograph of the user's face, even a perfect speaker verifier and a perfect face verifiers fused at the score level would be easily cheated.

Liveness detection based on the synchrony detection of lip movements and speech has been recently proposed in the literature [2]. On the other hand, Coupled Hidden Markov Models (CHMM) have been used for audio-visual speech

* This project has been partially supported by Spanish MEC under the project PRESA TEC2005-07212 and the European Union through the NoE BioSecure and K-Space

recognition [3–5], since they are well suited to model dynamic relationships between several signals. This paper is organized as follows. In section 2 the audio-visual features and further processing necessary to adapt the features to the CHMM framework are presented. In section 3 the CHMM audio-visual modelling is shown. In section 4 four different hypothesis tests to perform the asynchrony detection based on CHMM are presented. The experimental framework is explained in section 5. Results are shown in section 6, and the paper is drawn to conclusion in section 7.

2 Audio-visual Feature Extraction

Any aliveness check based on the link between the lip movement and the speech produced needs at least two information streams. One of them must encode the acoustic information whilst the other must encode the lip movement information.

2.1 Acoustic Features

Mel-Frequency Cepstral Coefficients (MFCC) are classical acoustic speech features in automatic speech processing. They are state-of-the-art features in many applications, including automatic speech recognition and speaker verification. Every 10 ms, a 20 ms long window is extracted from the acoustic signal and 12 MFCCs and the signal energy are computed, in order to get 13-dimensional acoustic speech features. First and second order time-derivatives are then appended. Finally, a 39-dimensional feature vector is extracted every 10 ms.

2.2 Lip Features

The mouth detection algorithm described in [6] was used to locate the lip area, as shown in figure 1. A Discrete Cosine Transform (DCT) is then applied on the gray level size-normalized ROI, and the first 30 DCT coefficients (in a zig-zag manner, corresponding to the low spatial frequency) are kept as the visual speech features. In the same way as acoustic features, first and second order derivatives are appended to the static visual features, and finally 90-dimensional visual features are produced every video frame. Visual features have been linearly interpolated in order to equilibrate visual and acoustic sample rates. After the interpolation, both acoustic and visual features have a sample rate of 100 Hz.

2.3 Coinertia Analysis Transform

The CoInertia Analysis (CoIA) was first introduced by Doledec et al. [7] to solve statistical problems in ecology. CoIA aims at providing a two sets of axes, one for each data stream, on which the projections of the data maximize the covariance of the projections. Given two multivariate random variables $X \in \mathbb{R}^n$ and $Y \in \mathbb{R}^m$ of covariance matrix $C_{XY} = \mathrm{E}\{(X - \mu_X)(Y - \mu_Y)^t\}$, where the operator $\mathrm{E}\{\cdot\}$ is the expectation operator, CoIA finds the orthogonal vectors $\{\mathbf{a_1}, \ldots, \mathbf{a_d}\}$

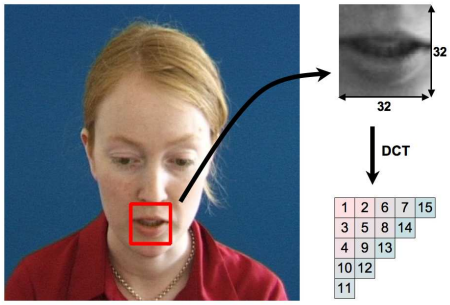

Fig. 1. Appearance-based features extraction

and $\{\mathbf{b_1}, \ldots, \mathbf{b_d}\}$, where d is the rank of C_{XY}, which maximize the covariance between projections of X and Y. These vectors are learnt from a training subset and then applied to all the features. These projected variables with maximum covariance are $\mathcal{X} = \{\mathbf{a_1}^t X, \ldots, \mathbf{a_d}^t X\}$ and $\mathcal{Y} = \{\mathbf{b_1}^t Y, \ldots, \mathbf{b_d}^t Y\}$, sorted by the covariance. Covariance is a compromise between correlation, maximized by the CANonical CORrelation (CANCOR) and variance [2], maximized by the Principal Component Analysis (PCA), and hence between inter-set and intra-set modelization. CoIA, as a compromise between PCA and CANCOR, provides us with a mechanism to reduce the dimension of both visual and acoustic streams while keeping the most covariance as possible just keeping the K first components of the transformed features $\mathcal{X}$ and $\mathcal{Y}$. This is necessary to alleviate the curse of dimensionality in the CHMM training.

3 Dynamic Modelling

A CHMM can be seen as a collection of HMM where the state at time t for every HMM in the collection is conditioned by the states at time $t-1$ of all the HMM in the collection. This is illustrated in figure 2.A CHMM can be completely described by the parameters $\boldsymbol{\lambda} = \{\lambda^i\} = \{\pi^i_{s^i}, a^i_{s^i|\mathbf{r}}, b^i_{s^i}\}$, for every stream $i \in \{1, \ldots, N_h\}$, where N_h is the number of streams; $\mathbf{q_t} = \{q_t^1, \ldots, q_t^{N_h}\}$ is the composite state at time t, where $q_t^i \in \{1, \ldots, NS_i\}$ is the state of stream i and NS_i is the number of possible states for that stream; $\pi^i_{s^i}$ is the initial probability of the state s^i for the stream i; $a^i_{s^i|\mathbf{r}}$ is the state transition probability for the stream i and state s_i from the composite state $\mathbf{r} = \{r^1, \ldots, r^{N_h}\}$; and $b^i_{s^i}$ is the output distribution for stream i and state s^i. The transition probabilities for the stream i are defined as:

$$a^i_{s^i|\mathbf{r}} = \mathrm{P}\left(q_t^i = s^i | q_{t-1}^1 = r^1, \ldots, q_{t-1}^{N_h} = r^{N_h}\right) \tag{1}$$

The output distribution function for every state s^i and stream i is a gaussian mixture model (GMM) with $M^i_{s^i}$ mixtures. Let o_t^i be the observation of the stream i at time t. The output distribution can be written as:

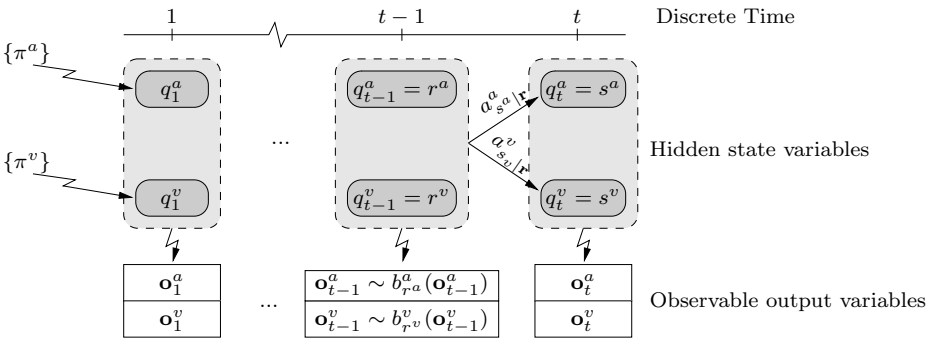

Fig. 2. CHMM state sequence depends for every HMM on the state of all the HMM in the CHMM. Only 2 streams, denoted as a and v, are used.

$$b^i_{s^i}(o^i_t) = \mathrm{P}\left(o^i_t | q^i_t = s^i\right) = \sum_{m=1}^{M^i_{s^i}} w^i_{s^i,m} \mathcal{N}(o^i_t; \mu^i_{s^i,m}, \sigma^i_{s^i,m}) \tag{2}$$

The initial states for the training sequences are obtained using the 5 internal states of an energy-based voice activity detector (VAD), applied to the most informative acoustic and visual features $\mathcal{X}_1$ and $\mathcal{Y}_1$ as described in subsection 2.3. The state transition probabilities $a^i_{s^i|\mathbf{r}}$ are initially estimated from the state transitions obtained by the VAD evolution for all the training sequences: $a^i_{s^i|\mathbf{r}} = n^i_{s^i|\mathbf{r}}/n^i_{\mathbf{r}}$, where $n^i_{s^i|\mathbf{r}}$ is the number of transitions to the state s_i of the stream i from the composite state $\mathbf{r} = \{r^1, \ldots, r^{N_h}\}$, and $n^i_{\mathbf{r}}$ is the total number of times that the CHMM visits the composite state $\mathbf{r}$ before the last sample for every training sequence. The initial state probabilities $\pi^i_{s^i}$ can be estimated as $\pi^i_{s^i} = n^i_{s^i}/ns$, where $n^i_{s^i}$ are the number of training sequences of which first state of the stream i is the state s^i, and ns is the total number of training sequences.

The Baum-Welch algorithm adapted to the CHMM framework is iterated 20 times to train the CHMM. The Viterbi algorithm is used to calculate the sequence of states for every stream and the frame loglikelihoods. This framework has been derived in previous works such as [3].

4 Asynchrony Detection

In order to detect the asynchrony between the acoustic and visual streams X and Y, a hypothesis test can be performed with the following hypothesis:

- $\mathcal{H}_\mathbf{0}$: Both streams are likely produced synchronously, and thereby there is a dependence of the state evolution of one stream with the other one. This hypothesis is represented by the CHMM $\boldsymbol{\lambda}$.
- $\mathcal{H}_\mathbf{1}$: Both streams are produced by independent sources, and hence there is not any dependence between both streams' state sequences. This hypothesis is represented by a two stream HMM as described in [8], namely $\boldsymbol{\lambda'}$.

Four different hypothesis have been derived within this framework:

1. The *first approach* is a slight modification of the classical Bayesian test:

$$\mathcal{H}_{\mathbf{0}} \text{ is accepted } \iff \frac{\mathrm{P}\left(\mathcal{X},\mathcal{Y},Q|\boldsymbol{\lambda}\right)}{\mathrm{P}\left(\mathcal{X},\mathcal{Y},Q'|\boldsymbol{\lambda'}\right)} > \theta, \tag{3}$$

   where Q and Q' are the most likely state sequence. These likelihoods are provided by the Viterbi algorithm. This test approximates the classical Bayesian test whether there is a state sequence much more likely than the others.
2. The *second approach* is derived from the previous one:

$$\mathcal{H}_{\mathbf{0}} \text{ is accepted } \iff \frac{\mathrm{P}\left(Q|\boldsymbol{\lambda}\right)}{\mathrm{P}\left(Q'|\boldsymbol{\lambda'}\right)} > \theta, \tag{4}$$

   since $\mathrm{P}\left(\mathcal{X},\mathcal{Y},Q|\boldsymbol{\lambda}\right) = \mathrm{P}\left(\mathcal{X},\mathcal{Y}|Q,\boldsymbol{\lambda}\right)\mathrm{P}\left(Q,\boldsymbol{\lambda}\right)$. This test eliminates the mismatch due to the differences in the trained output distributions of $\boldsymbol{\lambda}$ and $\boldsymbol{\lambda'}$.
3. The *third approach* performs the test:

$$\mathcal{H}_{\mathbf{0}} \text{ is accepted } \iff \frac{\mathrm{P}\left(\mathcal{X},\mathcal{Y},Q|\boldsymbol{\lambda}\right)}{\mathrm{P}\left(\mathcal{X},\mathcal{Y},Q'|\boldsymbol{\lambda'_u}\right)} > \theta, \tag{5}$$

   where $\boldsymbol{\lambda'_u}$ is an uncoupled version of $\boldsymbol{\lambda}$.
4. The *fourth approach* is a combination of the second and the third one:

$$\mathcal{H}_{\mathbf{0}} \text{ is accepted } \iff \frac{\mathrm{P}\left(Q|\boldsymbol{\lambda}\right)}{\mathrm{P}\left(Q'|\boldsymbol{\lambda'_u}\right)} > \theta, \tag{6}$$

   where $\boldsymbol{\lambda'_u}$ is an uncoupled version of $\boldsymbol{\lambda}$.

The two stream HMM $\boldsymbol{\lambda'_u}$ used in the third and fourth approaches shares the parameters $\{\pi^i_{s^i}\}$ and $\{b^i_{s^i}(o^i_t)\}$ with $\boldsymbol{\lambda}$. The state transition vectors $\{a^i_{s^i|r_i}\}$ are generated from the CHMM $\boldsymbol{\lambda}$ parameters abiding the following relation:

$$\begin{aligned} a^i_{s^i|r^i} &= \mathrm{P}\left(q^i_t = s^i | q^i_{t-1} = r^i\right) \\ &= \sum_{\mathbf{q_{t-1}}|q^i_{t-1}=r^i} \mathrm{P}\left(q^i_t = s^i|\mathbf{q_{t-1}} = \mathbf{r}\right) \prod_{j=1,j\neq i}^{N_h} \mathrm{P}\left(q^j_{t-1} = r^j\right) \\ &= \sum_{r_1=1}^{NS_1} \cdots \sum_{r^{i-1}=1}^{NS_{i-1}} \sum_{r^{i+1}=1}^{NS_{i+1}} \cdots \sum_{r^{N_h}=1}^{NS_{N_h}} a^i_{s^i|\mathbf{r}} \prod_{j=1,j\neq i}^{N_h} \mathrm{P}\left(q^j_{t-1} = r^j\right) \end{aligned} \tag{7}$$

The probability $\mathrm{P}\left(q^i_t = r^i\right)$ can be calculated. It depends on the time, but it is not desirable to work with time-dependent state transition probabilities. Therefore, since the quantity $\lim_{t\to\inf} \mathrm{P}\left(q^i_t = r^i\right)$ converges fastly for ergodic models, it is computed following this iterative procedure:

1) Initialization: for $t = 1$,	$\mathrm{P}\left(q^i_1 = s_i\right) = \pi^i_{s^i}$
2) Induction:	$\mathrm{P}\left(q^i_t = s^i\right) = \sum_{\mathbf{r}} a^i_{s^i\|\mathbf{r}} \prod_{j=1}^{N_h} \mathrm{P}\left(q^j_{t-1} = r^j\right)$
3) Stop condition:	$\left\lvert \frac{\mathrm{P}\left(q^i_t=s^i\right)-\mathrm{P}\left(q^i_{t-1}=s^i\right)}{\mathrm{P}\left(q^i_t=s^i\right)} \right\rvert < 10^{-6}$

5 Experimental Framework

The experiments conducted in this paper have been performed on the English part of the BANCA Database [9]. A new protocol focused on detecting audio-visual asynchrony has been designed. Asynchronous recordings are artificially built using audio and video from two different recordings from the same subject. Only client accesses recordings, in which true identity is claimed, were used. All the sessions are used in this protocol, including controlled, degraded and adverse condition recordings. Finally, 622 synchronized videos and 6820 desynchronized videos are used for testing purposes. The *World Model* part of the database, a total of 60 video sequences, 20 from each environment in the database, is used to train the models.

The BANCA database is divided into two disjoint groups, namely group 1 and group 2. Performance in one group is calculated using the thresholds that fix the working point in the other group to the equal error rate (EER), the so-called a priori EER threshold. Half Total Error Rate (HTER) and the Detection Error Tradeoff (DET) curves [10] are provided for performance comparison. HTER will be calculated taking into account both group 1 and group 2 using the thresholding approach described previously:

$$HTER = \frac{1}{2}\left(\frac{FA_1 + FA_2}{NI_1 + NI_2} + \frac{FR_1 + FR_2}{NC_1 + NC_2}\right) \tag{8}$$

where FA_i is the number of not synchronized videos classified as synchronized in group i, NI_i is the number of not synchronized videos in group i, FR_i is the number of synchronized videos classified as not synchronized in group i, and NC_i is the number of synchronyzed videos in group i.

6 Experimental Results

Several CHMM configurations have been tried, and finally all the experiments used CHMMs with 8 Gaussians in every GMM output distribution and output dimension 8 for both acoustic and visual streams (parameter K in the CoIA formulation). A number of Gaussians too big leads to poor performance due to the lack of training examples, whilst a too small number of Gaussians in the GMMs leads to a poor modeling, and therefore poor performance. Performance remains similar and very close to the optimal for a number of Gaussians around 8, and hence this value has been used as a reference for performance comparison of the different methods. The curse of dimensionality makes dimension 8 a suitable value: much bigger dimensions make the training fail, whilst too small dimensions do not provide enough information.

Table 1 shows the HTER related to every asynchrony detection method. Besides, the asynchrony detection performances of the methods described in this paper can be visually compared in figure 3. Mismatch between output distributions drives the first approach to the worst performance. Second approach gets better performances, although it is far from the performances of both third

Table 1. HTER for the different hypothesis tests

Hypothesis test	HTER (%)
First approach	37.58
Second approach	27.30
Third approach	13.06
Fourth approach	17.65

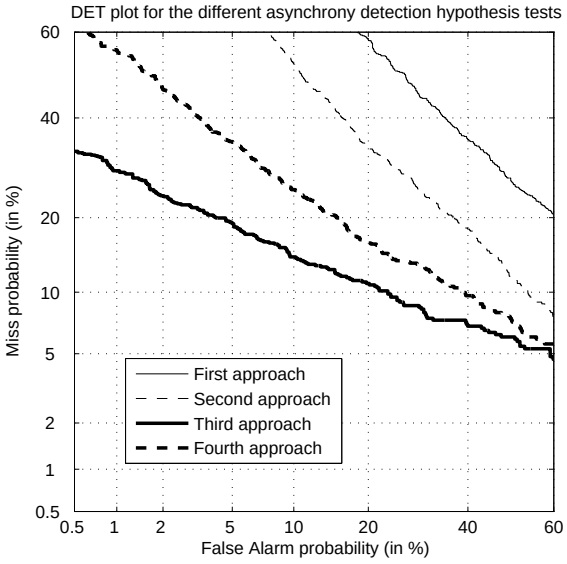

Fig. 3. DET curves for the different hypothesis tests

and fourth approaches. The uncoupling procedure used in third and fourth approaches enhances the synchrony information and gets the best results. Fourth approach is eliminating information that should not be removed, since in that approach the output distributions are the same for both $\boldsymbol{\lambda}$ and $\boldsymbol{\lambda}'_u$. The third and fourth approaches have an additional advantage: they avoid the training of a two stream HMM, since model $\boldsymbol{\lambda}'_u$ is built from the already trained CHMM by means of a simple and inexpensive uncoupling procedure.

7 Conclusions

Different asynchrony detection methods have been derived from the CHMM theory and checked in an audio-visual liveness detection task. Results show that the synchrony detection can become an effective anti-spoofing technique. However these asynchrony detection tasks have application wherever the CHMM can be

used in order to determine the degree of coupling between two or more different streams.

The relative structural simplicity of the CHMM proposed and used here is a contrast to the structural complexity of the CHMM used for audio-visual speech recognition, where many different models are trained, one for each phonetic or visual unit. The lack of training sequences did not allow us to train such a family of CHMM, which could result in a much more accurate synchrony detection performance. Future works come up to use these hypothesis contrasts using more complex CHMM structures, where more accurate states are defined more strongly related to the analyzed signal information.

Possible applications of the principles shown here can emerge in different fields not directly related to biometrics, such as automatic video and soundtrack alignment in a movie postproduction or dubbing evaluation.

References

1. Claude C. Chibelushi, Farzin Deravi, and John S.D. Mason. A Review of Speech-Based Bimodal Recognition. *IEEE Trans. Multimedia*, 4(1):23–37, 2002.
2. Hervé Bredin and Gérard Chollet. Measuring Audio and Visual Speech Synchrony: Methods and Applications. In *IET International Conference on Visual Information Engineering 2006 (VIE 2006)*, pages 255 – 260, Bangalore, India, September 2006.
3. X. Liu, L. Liang, Y. Zhaa, X. Pi, and A. V. Nefian. Audio-visual Continuous Speech Recognition using a Coupled Hidden Markov Model. In *Proceedings of the International Conference on Spoken Language Processing*, 2002.
4. Xiaozheng Zhang, Russell M. Mersereau, and Mark Clements. Bimodal Fusion in Audio-Visual Speech Recognition. In *IEEE 2002 International Conference on Image Processing*, volume 1, pages 964–967, September 2002.
5. Ara V. Nefian, Luhong Liang, Xiaobo Pi, Liu Xiaoxiang, Crusoe Mao, and Kevin Murphy. A Coupled HMM for Audio-Visual Speech Recognition. In *Proceedings of the International Conference on Acoustics Speech and Signal Processing (ICASSP02)*, May 2002.
6. Hervé Bredin, Guido Aversano, Chafic Mokbel, and Gérard Chollet. The Biosecure Talking-Face Reference System. In *2nd Workshop on Multimodal User Authentication*, May 2006.
7. Sylvain Dolédec and Daniel Chessel. Co-Inertia Analysis: an Alternative Method for Studying Species-Environment Relationships. *Freshwater Biology*, 31:277–294, 1994.
8. Sabri Gurbuz, Zekeriya Tufekci, Tufekci Patterson, and John N. Gowdy. Multi-Stream Product Modal Audio-Visual Integration Strategy for Robust Adaptive Speech Recognition. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Orlando, 2002.
9. Enrique Bailly-Bailliére et al. The BANCA Database and Evaluation Protocol. In *Lecture Notes in Computer Science*, volume 2688, pages 625 – 638, January 2003.
10. A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki. The DET Curve in Assessment of Detection Task Performance. In *European Conference on Speech Communication and Technology*, pages 1895 – 1898, 1997.