

A Speaker independent “Liveness” Test for Audio-Visual Biometrics

Nicolas Eveno and Laurent Besacier

Laboratoire CLIPS

Grenoble, France

nicolas.eveno@imag.fr

laurent.besacier@imag.fr

Abstract

In biometrics, it is crucial to detect impostors and thwart replay attacks. However, few researches have focused yet on the “liveness” verification. This test ensures that biometric cues being acquired are actual measurements from a live person who is present at the time of capture. Here, we propose a speaker independent “liveness” verification method for audio-video identification systems. It uses the correlation that exists between the lip movements and the speech produced. Two data analysis methods are considered to model this statistical link. Finally, according to tests carried out on the XM2VTS database, the best liveness verification *EER* achieved is 14.5% .

1. Introduction

Over the last few decades, significant research efforts have been put on biometric systems. This can be explained by the considerable needs for persons identification in many sectors such as law enforcement, communications or transportations. The most usual identification approaches rely on a single cue such as voice, face, fingerprints or retina. As pointed out by Jain *et al.* in [1], one of the fundamental barrier of these monomodal methods is their weakness against fraudulent attacks. The most well known “spoofing” technics use fake fingers (for fingerprint based systems), high resolution still images (for facial or retina based systems) and voice recorder (for speech based systems). A solution to thwart these attacks is to use a “liveness test” ensuring that the biometric cue being acquired is an actual measurement from a live person who is present at the time of capture. In this paper, we focus on the liveness test for speaker recognition systems.

A simple way to thwart replay attacks on a voice recognition system is to use a text-prompted protocol (e.g. [2]). In that case, the system is trained by using several key sentences and asks the client to pronounce one of them to grant access. With that solution, replay attacks are more difficult because they imply that the impostor has recorded all the key sentences. This exhaustive recording is however very easy if the impostor and the authorized client are collusive. Moreover, such a system requires an utterance verification module which can be the source of false rejections.

Another strategy to improve the robustness of identification systems is to use jointly audio and video data. It is well known that visual informations can significantly improve the robustness of speech recognition or speaker identification in adverse audio conditions (e.g. [3]). The first bimodal identification systems were using the speech and a static image of the face (e.g. [4]). They can be very easily fooled with a recorded voice

associated to a high resolution picture of an authorized client. To avoid this, more recent bimodal identification systems use the facial movements (e.g. [5], [6]). As said in [5] or [1], this kind of system can be seen as an implicit solution to the problem of liveness verification because it is extremely difficult for an impostor to reproduce both audio and visual dynamics. However, in the case of a replay attack with a recorded passphrase, the audio fits perfectly with an authorized client voice. Then, the system becomes mono-modal because the only remaining discriminating cue is the video. In general, the video based identifications perform worse than bimodal ones. For example, the accuracy rates reported in [6] are 57% and 87% for visual alone and audio-visual identifications respectively. In these conditions, it is crucial to use a separate liveness test, to preserve the performance of bimodal identification systems.

The liveness test proposed in this paper uses the tight link that exists between the lip movements and the speech produced. This high dependence has been exploited for a long time in audio-visual ASR systems. In [7], Yehia *et al.* have measured a high correlation between the shape of the vocal tract, the positions of several points around the mouth and the speech. Though this phenomenon is now very well known and is one of the theoretical base of many audio-visual systems, the automatic measurement of speech-face synchronism remains an open issue. In [8], the authors propose a very interesting approach based on *Canonical Correlation Analysis (CANCOR)*. This method allows the measure of statistical dependence between two sets of data and is used here to quantify the degree of synchronism between an audio track and the corresponding video sequence. However, their method needs a training step to compute the audio-visual model of the speaker. Obviously, this requirement can not be met in the case of an impostor trying to fool a biometric system. In the framework of identification, the liveness test must therefore be speaker-independent.

Here, we introduce a speaker independent liveness verification method which exploits the synchronism between the lip movements and the speech. In the audio-video data analysis step, we compare two statistical methods : the *CANCOR* (as in [8]) and the *Coinertia Analysis (COIA)*. This latter has been introduced recently to solve statistical problems in ecology [9] and is quite unknown in the signal processing community. Finally, we propose an original *liveness score* that enables a good detection of replay attacks.

The following section describes the audio and visual parameters extraction. In section 3, the *CANCOR* and *COIA* analysis methods and the deriving *liveness scores* are presented. Experimental results and their analysis are presented in section 4. Section 5 concludes this paper and indicates several possible future extensions.

2. Parameters extraction

There are many ways to parameterize the speech signal. However, since our study focuses on the relationships between the voice and the lip movements, we chose the LPC model because it enables a separate modelization of the source and the vocal tract. The visual parameters that we use are derived from the lips, which are the most visible parts of the vocal tract. We can thus expect the LPC filter parameters to correlate well with these visual parameters.

The LPC filter parameters are extracted at the video frame rate (25 fps). The computation window is twice as long as the video frame interval (80 ms) and is centered on the corresponding video frame. Moreover, as our goal is not to make a high resolution voice parameterization, but solely to evaluate the synchronism of audio and video streams, we only compute the first 5 LPC parameters.

The video parameters are derived from the outer lips contour. The method used to achieve the lip segmentation is explained in [10]. This algorithm requires the manual selection of a single point located above the mouth in the first frame of the video sequence. Then, the segmentation is automatically achieved over all the other frames by the fitting of a deformable template composed of several polynomial curves. As explained previously, our purpose is not to get a high resolution audio-video speaker description. So, we derive only 3 basic video parameters from the model : the mouth width, height and area.

Finally, 5 audio parameters and 3 video parameters are associated to each video frame, and are thus acquired at the video frame rate (25 fps). Moreover, each one of these parameters is centered so that its mean is 0. In the following sections, the centered audio and video data sets are notated $\mathbf{A}=(\mathbf{a}_1, \dots, \mathbf{a}_p)$ and $\mathbf{V}=(\mathbf{v}_1, \dots, \mathbf{v}_q)$ respectively. Here, the number p of audio parameters is 5, and the number q of video parameters is 3. The total number of samples (i.e. number of video frames) is n .

3. Liveness evaluation

3.1 CANCOR analysis

The *canonical correlation analysis (CANCOR)* is a statistical method to measure the relationship between two sets of multi-dimensional data [11]. We can use it to find the linear combinations of audio variables and video variables whose correlations are mutually maximized. In other words, it computes two sets of vectors $\{\mathbf{t}_1, \dots, \mathbf{t}_q\}$ and $\{\mathbf{u}_1, \dots, \mathbf{u}_q\}$ on which the projections of the data maximize :

$$\rho_i = \text{corr}(\mathbf{A}\mathbf{t}_i, \mathbf{V}\mathbf{u}_i), \quad i \leq q \quad (1)$$

where *corr* is the Pearson's correlation. Then, $\mathbf{A}\mathbf{t}_i$ and $\mathbf{V}\mathbf{u}_i$ are the projections of the audio and video data on vectors \mathbf{t}_i and \mathbf{u}_i respectively. Computational details and a MATLAB package can be found in [12]. The correlation coefficients ρ_i take values in the interval $[-1, 1]$ and are ordered in descending order. So, the projections of audio and video data on the first axes \mathbf{t}_1 and \mathbf{u}_1 are the most correlated. As explained in section 3.3, the first correlation coefficient ρ_1 can be used to compute a liveness score.

3.2 COIA analysis

The *Coinertia Analysis (COIA)* has been introduced to solve statistical problems in ecology [9]. It has been used recently by Goecke *et al.* in [13] to analyze the dependency of lip and voice parameters. For different phonemes of the Australian English, a coinertia analysis is performed and audio-video statistical dependences are derived. This study does not enable a "liveness" verification, but it shows that COIA is more stable than CANCOR, i.e. the number of samples does not affect significantly the accuracy of the results.

Like CANCOR, COIA aims at providing two sets of axes, one for each data set, on which the projections of the data maximize a statistical criterion. However, COIA analysis does not maximize the correlation, but the covariance of the projections:

$$\text{cov}(\mathbf{A}\mathbf{t}_i, \mathbf{V}\mathbf{u}_i) = \text{corr}(\mathbf{A}\mathbf{t}_i, \mathbf{V}\mathbf{u}_i) \times \sqrt{\text{var}(\mathbf{A}\mathbf{t}_i)}\sqrt{\text{var}(\mathbf{V}\mathbf{u}_i)}, \quad i \leq q \quad (2)$$

This covariance can be decomposed in three terms, showing that COIA is a compromise between the correlation (maximized by CANCOR) and the variance (maximized by *Principal Component Analysis*). In other words, COIA is a compromise between the inter-set and the intra-set dependency modelization methods. The details for deriving the coinertia axes $\{\mathbf{t}_1, \dots, \mathbf{t}_q\}$ and $\{\mathbf{u}_1, \dots, \mathbf{u}_q\}$ are given in [9]. Like in CANCOR, the relationship between audio and video data can be measured by the correlation between the projections on the first axes \mathbf{t}_1 and \mathbf{u}_1 .

3.3 "Liveness" scores

The correlation coefficients provided by CANCOR and COIA (notated ρ_{CANCOR} and ρ_{COIA} respectively) give information about the relationship between audio and video data. So, a simple technic to detect an audio-video asynchronism would be the direct thresholding of these coefficients. However, as it is clearly demonstrated in [13], their value is highly dependent on the pronounced phonemes. We observed that, for a given speaker, the correlation coefficients vary from a sentence to another. Moreover, some people articulate less than others when speaking, which leads to low ρ_{CANCOR} and ρ_{COIA} even in the case of a live recording with the same sentence pronounced (see the plain lines in Fig. 1). Finally, we also measured the correlation coefficients in the case of replay attacks simulated by associating video tracks with wrong audio tracks. As shown in Fig. 1 (dashed lines), the means of ρ_{CANCOR} and ρ_{COIA} are lower for replay attacks than for live recordings. However, the distributions overlap in large areas, which makes difficult the separation of the classes "live recording" and "replay attack" by using only the value of ρ_{CANCOR} and ρ_{COIA} .

Consequently, the liveness score that we propose is not based on a single value of the correlation coefficients, but on their evolution when audio is progressively shifted. In the case of a live recording, ρ_{CANCOR} and ρ_{COIA} are maximum when the delay between the audio and video tracks is very small, and decreases when this delay increases. For a replay attack, audio and video do not fit perfectly. So, a shift of audio track can lead to higher values of ρ_{CANCOR} and ρ_{COIA} .

These observations are summarized in Fig. 2, showing typical evolutions of the correlation coefficients when audio is shifted. On this figure, we associated a video track with three

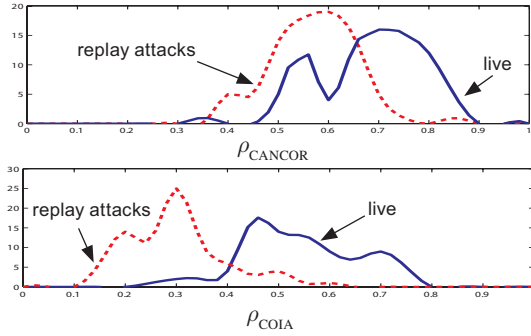


Figure 1: Histograms of ρ_{CANCOR} (up) and ρ_{COIA} (down) for live recordings (plain lines) and replay attacks (dashed lines), both with the same sentence pronounced.

different audio tracks : the original audio track (live recording), the same sentence pronounced by another speaker (first kind of replay attack) and another sentence pronounced by another speaker (second kind of replay attack). As said above, the replay attacks lead to lower correlation coefficients than the live recording and their associated correlation coefficients are not maximum for a null audio delay. Moreover, in the case of the second kind of replay attack, the curves are quite flat and do not have any salient maximum. Contrarily, the curves associated to the live recordings decrease when the delay increases. It is also interesting to note that their maxima are not obtained with a null delay but with a slight negative one, i.e. a small backward shift of audio track. This asynchronism between the lip movements and the speech produced does not come from a recording artifact and is well known in the AV speech processing community. It exists because the articulatory movements occur before the sound during the speech production. In the sequences of our database, this natural delay ranges from -2 to 0 frame interval, i.e. from -80 to 0 ms.

The liveness score is derived from the shape of these correlation coefficient curves. As said previously, the values of ρ_{CANCOR} and ρ_{COIA} vary from a speaker to another. So, the coefficient curves are first normalized so that their maximum is 1 and their minimum is 0. Then, we propose an original liveness score L as follows:

$$L = \frac{\sum_{i=-D}^D f(\tilde{\rho}, i)}{2D+1} \left(\frac{\tilde{\rho}_{\text{ref}}}{\text{mean}(\tilde{\rho})} - 1 \right) \quad (3)$$

with:

$$\begin{cases} f(\tilde{\rho}, i) = \begin{cases} 1, & \text{if } \tilde{\rho}(i) \leq \tilde{\rho}_{\text{ref}} \\ 0, & \text{else} \end{cases} \\ \tilde{\rho}_{\text{ref}} = \text{Max}_{-2 \leq i \leq 0} [\tilde{\rho}(i)] \end{cases}$$

where $\tilde{\rho}(i)$ is the normalized value of ρ_{CANCOR} or ρ_{COIA} computed for an audio track shift of i video frames. D is the maximum shift (we take $D=10$), $\tilde{\rho}_{\text{ref}}$ is the maximum value of the correlation coefficient for small negative audio shifts, and $\text{mean}(\tilde{\rho})$ is the average value of $\tilde{\rho}(i)$. The liveness score L is composed of two terms. The first one ranges from 0 to 1 and is maximum when the highest correlation coefficient is found with small negative audio shifts. The second term increases when the central peak of the curve is narrow and high.

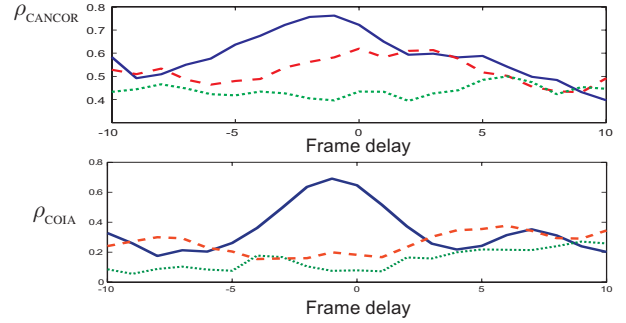


Figure 2: Typical effect of audio shifting on ρ_{CANCOR} (up) and ρ_{COIA} (down) for a live recording (plain) and for a replay attack with the same (dashed) and with a different (dotted) sentence.

4. Results and discussion

To evaluate the performances of our method, we used the XM2VTS database [14]. It consists in audio-video sequences recorded from 295 subjects in four sessions. The pronounced sentence is the same in the four sessions for all the speakers. For this study, we used the first session. Moreover, some audio tracks were containing errors such as recording “clicks” or very low signal to noise ratios. So only 264 sequences were used for our experiments. For each sequence, audio and video parameters have been extracted at the video frame rate, as explained in the section 2. Moreover, replay attacks have been simulated by associating each video track with four audio tracks containing the **same** sentence pronounced by other speakers (first kind of replay attack) and with four audio tracks containing **different** sentences pronounced by other speakers (second kind of replay attack). So, a total of 264 live recording sequences, 1056 replay attacks of type 1 and 1056 replay attacks of type 2 are used for our tests. For all the attacks, the beginning and the end of the fake speech have been aligned with those of the real speech. It is important to note that the replay attacks of type 1 are very challenging because the fake audio associated to the original video contains the same initial sentence with a perfect start and stop timing (only the speaker has changed).

False Acceptance (FA) and *False Rejection (FR)* curves for the liveness verification (i.e. “is it a live recording or an attack?”) are shown in Fig. 3. They have been computed by applying a varying threshold on the liveness scores L_{CANCOR} and L_{COIA} associated to all the sequences (live recordings and replays attacks). The derived *Equal Error Rate (EER)* are presented in Tab. 1. Moreover, we also computed the *EER* obtained with direct thresholdings of the correlation coefficients $\rho_{\text{CANCOR}}(0)$ and $\rho_{\text{COIA}}(0)$ (parenthesized numbers in Tab. 1), as suggested at the beginning of the section 3.3.

It appears clearly that our method provides much more accurate results than the direct thresholding of a single correlation value (see Tab. 1). The lowest *EER* is 14.5% and is obtained with the COIA based liveness score. In comparison, the simple thresholding of $\rho_{\text{COIA}}(0)$ leads to an *EER* of 24.5% with the same kind of attacks. It appears also that COIA performs always better than CANCOR both with our method and with the correlation thresholdings. In the case of a live recording, the central peak of the correlation curves computed by COIA are generally higher and narrower than those computed with CANCOR (as shown in Fig. 2). This leads to higher values of

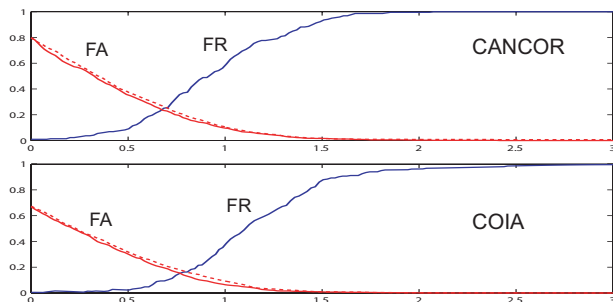


Figure 3: FA and FR error curves of our liveness verification method, obtained with CANCOR (up) and COIA (down). The FA curves associated to the 2nd kind of replay attacks are dotted.

the second term in Eq. 3, and therefore to an easier separation of live and fake recordings. So, the simple maximization of the correlation (achieved by CANCOR) between the lip movements and the audio speech is not sufficient in the framework of liveness verification. Our results show that the audio-video statistical modelization must be a compromise between the separate audio and video modelizations on the one hand, and their statistical dependency on the other hand. As explained in section 3.2, this compromise is realized by COIA.

We expected the two kinds of replay attacks (same or different sentences) to provide very different results. Quite surprisingly, their associated FA curves and EER are very close (see Fig. 1 and Tab. 1). This could come from the liveness score computation which is achieved by using only the shape of the correlation coefficient curve. In particular, the normalization process cancels the difference that exists between the average correlations associated to the two kinds of attacks. Contrarily, the simple thresholding of the correlation coefficients is more dependent on the kind of attacks (see parenthesized numbers in Tab. 1).

Bad segmentations of the lip contour are possible sources of error for our method. We determined that the segmentations were inaccurate in 44 sequences. Without considering these sequences, the computed EER dropped by 1.5% approximately.

Finally, it is also important to note that the EER point may not be the actual operating point of our liveness test since this module has to be followed by a conventional bi-modal biometric system. The liveness test can then be used to filter out the main part of impostor attacks, while keeping a very low FR rate (i.e. live recording rejection). For instance, a threshold of 0.5 for L_{COIA} enables the rejection of 70% of the fake recordings, while keeping a very low FR (approx. 3%).

5. Conclusion

The liveness verification method proposed in this paper uses the tight link that exists between the lip movements and the speech produced. Two data analysis methods are tested to measure the audio-visual parameters dependency. The successive

EER (%)	CANCOR	COIA
Replay attacks 1	25.5 (35.5)	17 (30)
Replay attacks 2	25 (31)	14.5 (24.5)

Tab. 1: EER of our method for CANCOR and COIA. The parenthesized results have been obtained with a direct thresholding of the correlation coefficients $\rho_{CANCOR}(0)$ and $\rho_{COIA}(0)$.

values of this measure when the audio track is progressively shifted are then used to compute a liveness score. It appears that COIA performs better than CANCOR and leads to an EER of 14.5% for the liveness verification. Finally, it is interesting to note that COIA enables a robust modelization of the speech process, i.e. the lip movements, the voice and their statistical relationship. Therefore, a bi-modal biometric system using parameters derived from COIA seems very promising because even the perfect reproduction of one of the cues (e.g. the voice) leads to a completely different global model. Our future research will focus on such a system.

References

- [1] A. K. Jain, S. Pankanti, S. Prabhakar, L. Hong, A. Ross, J.L. Wayman, "Biometrics : A Grand Challenge", *Proc. International Conference on Pattern Recognition (ICPR)*, Cambridge, UK, Vol. 2, pp. 935-942, August 2004.
- [2] T. Matsui, S. Furui, "Concatenated Phoneme Models for Text-Variable Speaker Recognition", *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 391-394, 1993.
- [3] C.C. Chibelushi, F. Deravi, J.S.D. Mason, "A Review of Speech-Based Bimodal Recognition", *IEEE Transaction on Multimedia*, vol. 4, no. 1, pp. 23-36, March 2002.
- [4] R. Brunelli, D. Falavigna, "Person Identification Using Multiple Cues", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, no. 10, pp. 955-966, 1995.
- [5] C.C. Broun, X. Zhang, R.M. Mersereau, M.A. Clements, "Automatic Speechreading with Application to Speaker Verification", *In Proc. ICASSP*, Orlando, May 2002.
- [6] N. Fox, R.B. Reilly, "Audio-Visual Speaker Identification Based on the Use of Dynamic Audio and Visual Features", *Proc. 4th International Conference on Audio and Video Based Biometric Person Authentication*, June 2003.
- [7] H. Yehia, P. Rubin, E. Vatikiotis-Bateson, "Quantitative association of vocal-tract and facial behavior", *Speech Communication*, 26:23-43, 1998.
- [8] M. Slaney, M. Covell, "FaceSync: A Linear Operator for Measuring Synchronization of Video Facial Images and Audio Tracks", *NIPS 2000*, pp. 814-820, 2000.
- [9] S. Doledec, D. Chessel, "Co-inertia analysis: an alternative method for studying species-environment relationships", *Freshwater Biology*, vol. 31, pp. 277-294, 1994
- [10] N. Eveno, A. Caplier, P.-Y. Coulon, "Accurate and quasi-automatic lip tracking". *IEEE Transactions on Circuits and Systems for Video Technology*, 14(5):706-715, May 2004.
- [11] H. Hotelling, "Relations between two sets of variates", *Biometrika*, 28:321-377, 1936.
- [12] Magnus Borga's page [Online], Available : <http://people.imt.liu.se/~magnus/>
- [13] R. Goecke, J.B. Millar, "Statistical Analysis of the Relationship between Audio and Video Speech Parameters for Australian English", *In J.L. Schwartz, F. Berthommier, M.A. Cathiard, and D. Sodoier (eds.), Proc. AVSP 2003*, pp.133-138, France, September 2003.
- [14] K. Messer, J. Matas, J. Kittler, J. Luettin, G. Maitre; "XM2VTSbd: The Extended M2VTS Database", *in Proc. 2nd Conf. on Audio and Video-base Biometric Personal Verification (AVBPA99)*, Springer Verlag, New York, 1999.