

## Audio-visual forgery in identity verification

G. Chollet, B. Abboud, G. Aversano

Verification of identity is commonly achieved by looking at the face of a person and listening to his (her) speech.

Automatic means of achieving this verification has been studied for several decades. A speaking face offers many features to achieve a robust verification of identity.

The current deployment of videophony offers new opportunities for a secured access to remote servers (banking, certification, ...)

The synchrony of the speech signal and lip movements is a necessary condition to check that the observed speaking face has not been manipulated and/or synthesized.

This paper will review face, speaker and speaking face verification, face morphing and voice transformation techniques. It is demonstrated that a dedicated impostor needs limited information from a client to fool state of the art audio-visual identity verification systems.

A visual forgery system has been designed allowing to artificially animate a static image in such a way that it reproduces the facial movements (and in particular the lip movements) of a talking face shown in a video. This allows elaborating a forgery scenario where an impostor utters a number of words and uses his own lip movements to artificially animate the lips of a static image of the client in such a way that it reproduces the exact lip movements of the impostor.

Moreover, to modify the voice characteristics of the impostor in such a way that it would resemble to the voice of the client, two different voice transformation techniques have been developed and tested. Forgery attacks against a speaker verification system have been simulated on widely known databases and protocols such as BANCA and NIST.

The choice of the transformation technique is made according to the available quantity of client voice data.

If only a limited amount of client data is available for training (as in the case of BANCA protocol), a spectral conversion technique should be adopted. Considering a sequence of spectral vectors pronounced by the impostor,  $X = [x_1, x_2, \dots, x_n]$ , and a sequence composed by the same words, pronounced by the client,  $Y = [y_1, y_2, \dots, y_n]$ , a spectral transformation can be performed by finding the conversion function  $F$  that minimizes the mean square error:

$$\varepsilon_{\text{mse}} = E[\|y - F(x)\|^2],$$

where  $E$  is the expectation.

This conversion method requires a preliminary word-level segmentation of the training sentences.

If more client data are available (e.g. approximately 1 hour of client's speech), we could use a voice encoder based on the "Automatic Language Independent Speech Processing" (ALISP) approach. The principle of this system is to encode speech by recognition and synthesis in terms of basic acoustic units that can be derived by an automatic analysis of the signal. Such analysis is not based on a priori linguistic knowledge.

In the context of speaking face cloning a system has been implemented that allows automatic lip movement tracking for a speaking face. The same algorithm is also used to automatically place a certain number of MPEG-4 compatible fiducial points around the lip contour of the static image we wish to animate. An image warping technique is then used to move these fiducial points on the static image in such a way that their new position matches the position of the tracked feature points on each frame of the driving video sequence.

Current limitations of the described transformation techniques are the difficulties to achieve real-time performance.