Audio-Visual Identity Verification and Robustness to Imposture

Walid Karam[1,2], Chafic Mokbel[1], Hanna Greige[1], and Gérard Chollet[2]

[1] University of Balamand, Deir El-Balamand, Al-Kurah, Lebanon
{Walid.Karam,Chafic.Mokbel,Hanna.Greige}@balamand.edu.lb
[2] CNRS-LTCI, TELECOM ParisTech, 46 rue Barrault, 75634 Paris, France
{karam,chollet}@telecom-paristech.fr

Abstract. The robustness of talking-face identity verification (IV) systems is best evaluated by monitoring their behavior under impostor attacks. We propose a scenario where the impostor uses a still face picture and a sample of speech of the genuine client to transform his/her speech and visual appearance into that of the target client. We propose *MixTrans*, an original text-independent technique for voice transformation in the cepstral domain, which allows a transformed audio signal to be estimated and reconstructed in the temporal domain. We also propose a face transformation technique that allows a frontal face image of a client to be animated, using principal warps to deform defined MPEG-4 facial feature points based on determined facial animation parameters. The robustness of the talking-face IV system is evaluated under these attacks. Results on the BANCA talking-face database clearly show that such attacks represent a serious challenge and a security threat to IV systems.

Keywords: Identity verification, audio-visual forgery, talking-face imposture, voice conversion, face animation, biometric verification robustness.

1 Introduction

Biometric identity verification (IV) systems are starting to appear on the market in various commercial applications. However, these systems are still operating with a certain measurable error rate that prevents them from being used in a full automatic mode, and still require human intervention and further authentication. This is primarily due to the variability of the biometric traits of humans over time because of growth, aging, injury, appearance, physical state, etc. Impostors attempting to be authenticated by an IV system to gain access to privileged resources could take advantage of the non-zero false acceptance rate of the system by imitating, as closely as possible, the biometric features of a genuine client.

The purpose of this paper is threefold. 1– It evaluates the performance of IV systems by monitoring their behavior under impostor attacks. These attacks include the transformation of the face and the voice biometric modalities. 2– It introduces *MixTrans*, a novel mixture-structure bias voice transformation technique in the cepstral domain, which allows a transformed audio signal to be estimated and reconstructed in

M. Tistarelli and M.S. Nixon (Eds.): ICB 2009, LNCS 5558, pp. 796–805, 2009.

the temporal domain. 3– It proposes a face transformation technique that allows a 2D face image of the client to be animated. This technique employs principal warps to deform defined MPEG-4 facial feature points based on determined facial animation parameters (FAP). The talking-face BANCA database is used to test the effects of voice and face transformation on the IV system.

The rest of the paper is organized as follows. Section 2 describes the imposture techniques used, including *MixTrans* and the MPEG-4 face animation with thin-plate spline warping technique. Section 3 discusses the experimental results on the BANCA talking-face database. Section 4 wraps up with a conclusion.

2 Audiovisual Imposture

Audiovisual imposture is the deliberate modification of both speech and face of a person so as to make that person sound and look like someone else. The goal of such an effort is to analyze the robustness of biometric identity verification systems to deliberate forgery attacks. An attempt is made to increase the acceptance rate of an impostor. Talking-face impostor has been reported in [12] with a single modality conversion, i.e. face animation, using a commercial animation software, CrazyTalk [1]. Other types of deliberate impostor evaluations are reported in [6] and [4], all based on face conversion. In this work, techniques for conversion of both the speech and the visual appearances of clients are developed and are treated below.

2.1 Speaker Transformation

Speaker transformation, also referred to as voice transformation, is the process of altering an utterance from a speaker (impostor) to make it sound as if it were articulated by a target speaker (client.) Such transformation can be effectively used to replace the client's voice in a video to impersonate that client and break an IV system.

MixTrans. A linear time-invariant transformation in the temporal domain is equivalent to a bias in the cepstral domain. However, speaker transformation may not be seen as a simple linear time-invariant transformation. It is more accurate to consider it as several linear time-invariant filters, each operating in a part of the acoustic space. This leads to the following form of the transformation:

$$\mathcal{T}_\theta(\mathbf{X}) = \sum_k \Pi_k(\mathbf{X} + \mathbf{b}_k) = \sum_k \Pi_k \mathbf{X} + \sum_k \Pi_k \mathbf{b}_k = \mathbf{X} + \sum_k \Pi_k \mathbf{b}_k \,. \tag{1}$$

where $\mathbf{b}_k$ represents the k^{th} bias and Π_k is the probability of being in the k^{th} part of the acoustic space given the observation vector $\mathbf{X}$. Π_k is calculated using a universal GMM modeling the acoustic space.

Once the transformation is defined, its parameters have to be estimated. Assume that speech samples are available for both the source and the target speakers, but do not correspond to the same text. Let λ be the stochastic model for a target client. λ is a GMM of the client. Let $\mathbf{X}$ represent the sequence of observation vectors for an impostor (a source client.) Our aim is to define a transformation $\mathcal{T}_\theta(\mathbf{X})$ that makes the source client vector resemble the target client. In other words, we would like to have

the source vectors be best represented by the target client model λ through the application of the transformation $\mathcal{T}_\theta(\mathbf{X})$. In this context the Maximum likelihood criterion is used to estimate the transformation parameters.

$$\hat{\theta} = \arg\max_{\theta} \mathcal{L}(\mathcal{T}_\theta(\mathbf{X})|\lambda)\,. \tag{2}$$

Since λ is a GMM, and $\mathcal{T}_\theta(\mathbf{X})$ is a transform of the source vectors $\mathbf{X}$, and $\mathcal{T}_\theta(\mathbf{X})$ depends on another model λ_m, then $\mathcal{L}(\mathcal{T}_\theta(\mathbf{X})|\lambda)$ in (2) can be written

$$\begin{aligned}
\mathcal{L}(\mathcal{T}_\theta(\mathbf{X}|\lambda) &= \prod_{t=1}^{T} \mathcal{L}(\mathcal{T}_\theta(\mathbf{X}_t)|\lambda) \\
&= \prod_{t=1}^{T}\sum_{m=1}^{M} \frac{1}{(2\pi)^{\frac{D}{2}}\left|\sum_m\right|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathcal{T}_\theta(\mathbf{X}_t)-\mu_m)^T \sum_m^{-1}(\mathcal{T}_\theta(\mathbf{X}_t)-\mu_m)} \\
&= \prod_{t=1}^{T}\sum_{m=1}^{M} \frac{1}{(2\pi)^{\frac{D}{2}}\left|\sum_m\right|^{\frac{1}{2}}} e^{-\frac{1}{2}\left(\mathbf{x}_t+\sum_{k=1}^{K}\Pi_{kt}b_k-\mu_m\right)^T \sum_m^{-1}\left(\mathbf{x}_t+\sum_{k=1}^{K}\Pi_{kt}b_k-\mu_m\right)}.
\end{aligned} \tag{3}$$

Finding $\{b_k\}$ such that (3) is maximized is found through the use of the EM algorithm. In the expectation "E" step, the probability α_{mt} of component m is calculated. Then, at the maximization "M" step, the log-likelihood is optimized dimension by dimension for a GMM with a diagonal covariance matrix:

$$ll = \sum_{t=1}^{T}\sum_{m=1}^{M} \alpha_{mt}\left[\log\frac{1}{\sigma_m\sqrt{2\pi}} - \frac{1}{2}\frac{\left(\mathbf{X}_t+\sum_{k=1}^{K}\Pi_{kt}\mathbf{b}_k-\mu_m\right)^2}{\sigma_m^2}\right].$$

Maximizing,

$$\frac{\partial ll}{\partial b_l} = 0 \Rightarrow -\sum_{t=1}^{T}\sum_{m=1}^{M}\alpha_{mt}\frac{\left(\mathbf{X}_t+\sum_{k=1}^{K}\Pi_{kt}\mathbf{b}_k-\mu_m\right)\Pi_{lt}}{\sigma_m^2} = 0, \qquad \text{for } l = 1\ldots K$$

then, in matrix notation,

$$-\left(\sum_m\sum_t \frac{\alpha_{mt}\Pi_{lt}\Pi_{kt}}{\sigma_m^2}\right)\left(\mathbf{b}_k\right) = \left(\sum_m\sum_t \frac{\alpha_{mt}\Pi_{lt}(\mathbf{X}_t-\mu_m)}{\sigma_m^2}\right). \tag{4}$$

This matrix equation is solved at every iteration of the EM algorithm.

Speech Signal Reconstruction. It is known that the cepstral domain is appropriate for classification due to the physical significance of the Euclidean distance in this space. However, the extraction of cepstral coefficients from the temporal signal is a nonlinear process and the inversion of this process is not uniquely defined. Therefore, a solution has to be found in order to take advantage of the good characteristics of the cepstral space while applying the transformation in the temporal domain. Several techniques have been proposed to solve this problem. In [10], harmonic plus noise analysis has been used for this purpose. Instead of trying to find a transformation allowing the passage from the cepstral domain to the temporal domain, we adopt a different strategy. Suppose that an intermediate space existed where transformation could be directly transposed into the temporal domain. Fig. 1 shows the process where

the temporal signal goes through a two-step feature extraction process leading to the cepstral coefficients that may be easily transformed into target speaker-like cepstral coefficients by applying the transformation function $\mathcal{T}_\theta(\mathbf{X})$ as discussed previously.

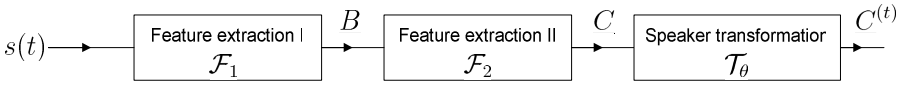


Fig. 1. Signal to transformed cepstral coefficients

The transformation that is trained in the cepstral domain cannot be directly projected on the temporal domain since the feature extraction module ($\mathcal{F}_1 \circ \mathcal{F}_2$) is highly nonlinear. However, a speaker transformation determined in the $\underline{B}$ space may be projected on the signal space, e.g. $\underline{B}$ space may be the spectral domain. But, for physical significance it is better to train the transformation in the cepstral domain. Therefore, we suppose that another transformation $\mathcal{T}'_\theta(\mathbf{X})$ existed in the B space leading to the same transformation in the cepstral domain satisfying thereby the two objectives; transformation of the signal and distance measurement in the cepstral domain. This is shown in Fig. 2.

Fig. 2. Transformation in a signal-equivalent space

The remaining issue is estimating parameters θ of the transformation $\mathcal{T}'_\theta(\mathbf{X})$ in order to get the same transformation result as in the cepstral domain. This is detailed next.

Estimating Signal Transformation is Equivalent to Calculating Cepstral Transformation. The transformation in the cepstral domain is presumably determined; the idea is to establish a transformation in the B space leading to cepstral coefficients similar to the one resulting from the cepstral transformation.

Let $\hat{C}^{(t)}$ represents the cepstral vector obtained after applying the transformation in the B domain; let $C^{(t)}$ represents the cepstral vector obtained when the transformation is applied in the cepstral domain. The difference defines an error vector: $\underline{e} = C^{(t)} - \hat{C}^{(t)}$ and the quadratic error $E = |e|^2 = \underline{e}^T\underline{e}$**.** Starting from a set of parameters for $\mathcal{T}'_\theta$, the gradient algorithm may be applied in order to minimize the quadratic error E. At each iteration of the algorithm the parameter θ is updated using the equation $\theta^{(i+1)} = \theta^{(i)} - \mu\frac{\partial E}{\partial \theta}$**,** where μ is the gradient step.

The gradient of the error with respect to parameter θ is given by:

$$\frac{\partial E}{\partial \theta} = -2\underline{e}^T \frac{\partial \hat{\underline{C}}^{(t)}}{\partial \theta}\,. \tag{5}$$

Finally, the derivative of the estimated transformed cepstral coefficient with respect to θ, can be obtained using a gradient descent:

$$\frac{\partial \hat{C}^{(t)}}{\partial \theta} = \frac{\partial \hat{C}^{(t)^T}}{\partial B^{(t)}} \frac{\partial B^{(t)}}{\partial \theta} . \tag{6}$$

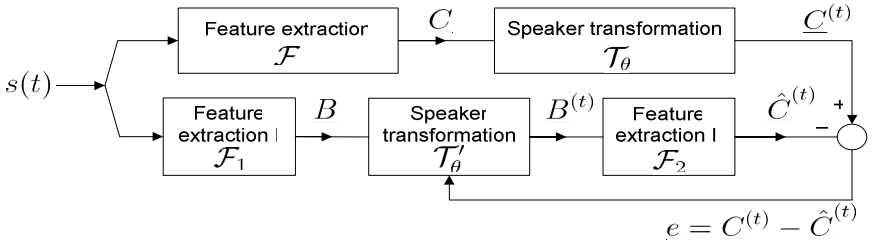


Fig. 3. Signal-level transformation parameters tuned with a gradient descent algorithm

In order to illustrate this principle, let us consider the case of MFCC analysis leading to the cepstral coefficients. In this case, $\mathcal{F}_1$ is just the Fast Fourier Transform (FFT) followed by the power spectral calculation (the phase being kept constant). $\mathcal{F}_2$ is the filterbank integration in the logarithm scale followed by the inverse DCT transform. We can write:

$$\hat{C}_l^{(t)} = \sum_{k=1}^{K} log\left(\sum_{i=1}^{N} a_i^{(k)} B_i^{(t)}\right) cos\left(2\pi l \frac{f_k}{F}\right) \cdot \tag{7}$$

$$B_i^{(t)} = B_i . \theta_i . \tag{8}$$

where $\{a_i\}$ are the filter-bank coefficients, f_k the central frequencies of the filter-bank, and θ_i is the transfer function at frequency bin *i* of the transformation $\mathcal{T}'_\theta(\mathbf{X})$.

Using (7) and (8), it is straightforward to compute the derivatives in (6):

$$\frac{\partial \hat{C}_i^{(t)}}{\partial B_j^{(t)}} = \sum_{k=1}^{K} \frac{a_j^{(k)}}{\sum_{i=1}^{N} a_i^{(k) B_i^{(t)}}} cos\left(2\pi l \frac{f_k}{F}\right) \cdot \tag{9}$$

$$\frac{\partial B_i^{(t)}}{\partial \theta_j} = B_j \delta_{ij} . \tag{10}$$

(5), (6), (9), and (10) allow the implementation of this algorithm in the case of MFCC.

Once $\mathcal{T}'_\theta(\mathbf{X})$ completely defined, the transformed signal may be determined by applying an inverse FFT to $B(t)$ and using the original phase to recompose the signal window. In order to consider the overlapping between adjacent windows, the Overlap and Add (OLA) algorithm is used.

Initializing the Gradient Algorithm. The previous approach is computationally expensive. Actually, for each signal window, i.e. 10ms to 16ms, a gradient algorithm is to be applied. In order to alleviate this high computational algorithm, a solution consists in finding a good initialization of the gradient algorithm. This may be obtained by using as initial value for the transformation $\mathcal{T}'_\theta(\mathbf{X})$, the transformation obtained for the previous signal window.

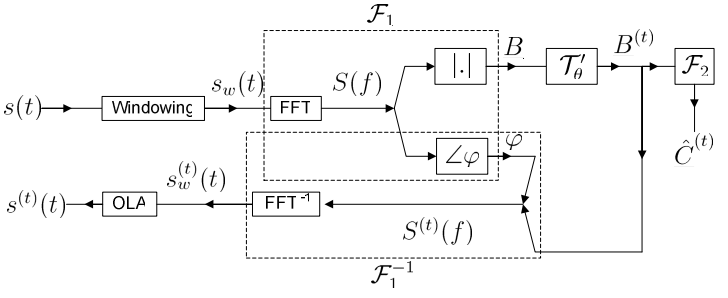


Fig. 4. Speech signal feature extraction, transformation, and reconstruction

2.2 Face Animation

To complete the scenario of audiovisual imposture, speaker transformation is coupled with face transformation. It is meant to produce synthetically an "animated" face of a target person, given a still photo of his face and some animation parameters.

The face animation technique used in this paper is MPEG-4 compliant, which uses a very simple thin-plane spline warping function defined by a set of reference points on the target image, driven by a set of corresponding points on the source image face.

MPEG-4 2-D Face Animation. MPEG-4 is an object-based multimedia compression standard, which defines a standard for face animation [**11**]. It specifies 84 feature points that are used as references for Facial Animation Parameters (FAPs). 68 FAPs allow the representation of facial expressions and actions such as head motion and mouth and eye movements. Two FAP groups are defined, visemes (FAP group 1) and expressions (FAP group 2). Visemes (FAP1) are visually associated with phonemes of speech; expressions (FAP2) are joy, sadness, anger, fear, disgust, and surprise.

An MPEG-4 compliant system decodes a FAP stream and animates a face model that has all feature points properly determined. In this paper, the animation of the feature points is accomplished using a simple thin-plate spline warping technique.

Thin-Plate Spline Warping. The thin-plate spline (TPS), initially introduced by Duchon [5], is a geometric mathematical formulation that can be applied to the problem of 2D coordinate transformation. The name *thin-plate spline* indicates a physical analogy to bending a thin sheet of metal in the vertical z direction, thus displacing x and y coordinates on the horizontal plane.

Given a set of data points $\{w_i, i = 1, 2, ..., K\}$ in a 2D plane – for our case, MPEG-4 facial feature points – a radial basis function is defined as a spatial mapping that

maps a location x in space to a new location $f(x) = \sum_{i=1}^{K} c_i \phi(\|x - w_i\|)$, where $\{c_i\}$ is a set of mapping coefficients and the kernel function $\phi(r) = r^2 \ln r$ is the thin-plate spline [3]. The mapping function $f(x)$ is fit between corresponding sets of points $\{x_i\}$ and $\{y_i\}$ by minimizing the "bending energy" *I*, defined as the sum of squares of the second derivatives

$$I[f(x,y)] = \iint_{\mathbb{R}^2} \left[\left(\frac{\partial^2 f}{\partial x^2} \right)^2 + 2 \left(\frac{\partial^2 f}{\partial xy} \right)^2 + \left(\frac{\partial^2 f}{\partial y^2} \right)^2 \right] dx\, dy \cdot \tag{11}$$

3 Effects of Imposture on Verification – Experimental Results

To test the robustness of IV systems, a state-of-the-art baseline audio-visual IV system is built. This system follows the BANCA [8] "P" protocol and is based on a classical GMM approach for both speech and face modalities. It is completely independent from the voice and face transformations described above.

BANCA defines "random" impostures, where a speaker proclaims in his/her own voice and face to be someone else. This "zero-effort" imposture is unrealistic and IV systems detect easily the forgery by contrasting the impostor model against that of the claimed identity. To make the verification more realistic, deliberate attacks are modeled with the transformation of both the voice and the face of the impostor.

To verify a claimed identity, audio and face feature vectors are matched against the claimed speaker model and against the world model. GMM client training and testing is performed on the open-source speaker verification toolkit BECARS [2].

3.1 Verification Experiments

Speaker Verification. To process the speech signal, a feature extraction module calculates relevant feature vectors from the speech waveform. On a signal "FFT" window shifted at a regular rate, cepstral coefficients are derived from a filter bank analysis with triangular filters. A Hamming weighting window is used to compensate for the truncation of the signal. Then GMM speaker classification is performed with 256 Gaussians. The world model of BANCA is adapted using MAP adaptation, and its parameters estimated using the EM algorithm.

A total of 234 true client tests and 312 "random impostor" tests per group were performed. Fig. 7 (a) shows the DET curves for speaker verification on g1 and g2, with an EER of 4.38% and 4.22% respectively.

Face Verification. Face verification is based on processing a video sequence in four stages: 1– Face detection, localization and segmentation, 2– Normalization, 3– Facial Feature extraction and tracking, and 4– Classification.

The face detection algorithm used in this work is a machine learning approach based on a boosted cascade of simple and rotated haar-like features for visual object detection [7]. Once a face is detected, it is normalized (resized to 48×64, cropped to 36×40, gray-scaled, and histogram equalized) to reduce the variability of different aspects in the face image such as contrast and illumination, scale, translation, and rotation. The face tracking module extracts faces in all frames and retains only 5 per video for training and/or testing.

The next step is face feature extraction. We use DCT-*mod2* proposed in [9].

In a similar way to speaker verification, GMM's are used to model the distribution of face feature vectors for each person.

For the same BANCA "P" protocol, and a total of 234 true clients and 312 "random impostor" tests (per group per frame – 5 frames per video) the DET curves for face verification are shown in Fig. 7 (b) with an EER of 23.5% (g1) and 22.2% (g2).

Score Fusion. A final decision on the claimed identity of a person relies on both the speech-based and the face-based verification systems. To combine both modalities, a fusion scheme is needed. The simple weighted sum rule fusion technique is used in this study. The sum rule computes the audiovisual score s by weight averaging: $s = w_s s_s + w_f s_f$, where w_s and w_f are speech and face score weights computed so as to optimize the equal error rate on the training set. The speech and face scores must be in the same range (e.g. $\mu = 0, \sigma = 1$) for the fusion to be meaningful. This is achieved by normalizing the scores as follows: $s_{norm(s)} = (s_s - \mu_s)/\sigma_s$ and $s_{norm(f)} = (s_f - \mu_f)\sigma_f$.

Fig. 7 (c) shows an improvement of the verification by score fusion of both modalities, with an EER of 4.22% for g1, and 3.47% for g2.

3.2 Transformation Experiments

Voice Conversion Experiments. BANCA has total of 312 impostor attacks per group in which the speaker claims in his own words to be someone else. These attempts are replaced by the transformed voices as described in section 2.1 above. For each attempt, MFCC analysis is performed and transformation coefficients are calculated in the cepstral domain using the EM algorithm. Then the signal transformation parameters are estimated using a gradient descent algorithm. The transformed voice signal is then reconstructed with an inverse FFT and OLA as described in section 02.1. The pitch of the transformed voice had to be adjusted to better match the target speaker's pitch. Verification experiments are repeated with the transformed voices. The result is an increase of the EER from 4.38% to 10.6% on g1, and from 4.22% to 12.1% on g2 (Fig. 7 (a)).

Face Conversion Experiments. Given a picture of the face of a target person, the facial feature points are first annotated as shown in Fig. 5. A total of 61 feature points out of the 83 of MPEG-4 are used, the majority of which belong to the eyes and the mouth regions. Others have less impact on FAP's or do not affect them at all.

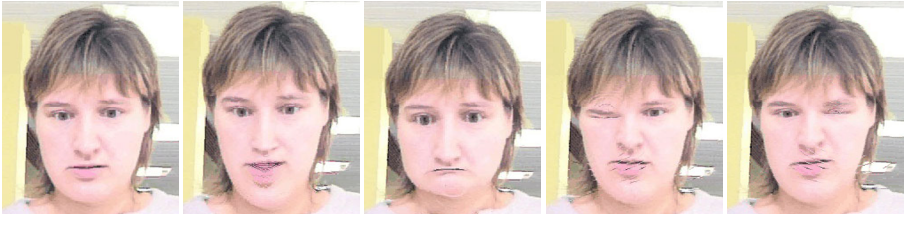
The FAP's used in the experiments correspond to a subset of 33 out of the 68 FAP's defines by MPEG-4. Facial actions related to head movement, tongue, nose, ears, and jaws are not used. The FAP's used correspond to mouth, eye, and eyebrow movements, e.g. horizontal displacement of right outer lip corner (`stretch_r_cornerlip_o`). Fig. 6 shows frames animating the noted expressions.

A synthesized video sequence is generated by deforming a face from its neutral state according to determined FAP values. For the experiments presented in this work, these FAP's are selected so as to produce a realistic talking head. The detection and the measure of the level of audiovisual speech synchrony is not treated in this work, but has been reported in [4] to improve the verification performance.

(a) Original sample video frame (b) Annotated face

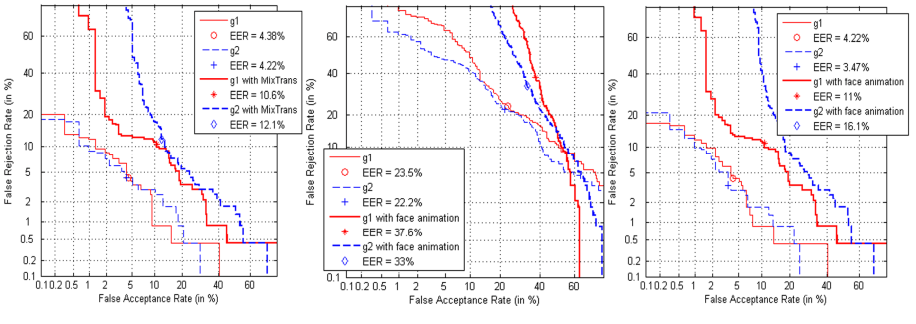
Fig. 5. Feature point annotation on the BANCA database (client number 9055)

BANCA has total of 312 impostor attacks per group in which the speaker claims in his own words and facial expressions to be someone else. These are replaced by the synthetically animated videos with the transformed speech. The experiments have shown a deterioration of the performance from an EER of [23.5%, 22.2%] on [g1, g2] to [37.6%, 33.0%] (Fig. 7 (b)) for face, and from [4.22%, 3.47%] to [11.0%, 16.1%] for the audio-visual system (Fig. 7 (c)).



(a) neutral face (b) joy expression (c) sad expression (d) right eye wink (e) left eye wink

Fig. 6. Selected frames from an animated face with various expressions



(a) Speech (b) Face (c) Audio-visual (fused results)

Fig. 7. Verification and imposture results on BANCA

4 Conclusion

This paper evaluates the robustness of audio-visual imposture on biometric identity verification systems. It proposes *MixTrans*, a mixture-structure bias voice transformation technique in the cepstral domain, which allows a transformed audio signal to be estimated and reconstructed in the temporal domain. It also couples the audio conversion with an MPEG-4 compliant face animation system that warps facial feature points using a simple thin-plate spline. The proposed audiovisual forgery is completely independent from the baseline audiovisual IV system, and can be used to attack any other audiovisual IV system. The Results drawn from the experiments show that state-of-the-art IV systems are vulnerable to forgery attacks, which indicate more impostor acceptance, and, for the same threshold, more genuine client denial.

References

1. Reallusion crazytalk animation studio software, http://www.reallusion.com/crazytalk/
2. Blouet, R., Mokbel, C., Mokbel, H., Soto, E.S., Chollet, G., Greige, H.: Becars: A free software for speaker verification. In: Proc. ODYSSEY 2004, pp. 145–148 (2004)
3. Bookstein, F.: Principal warps: Thin-plate splines and the decomposition of deformations. IEEE Transactions on Pattern Analysis and Machine Intelligence 11(6), 567–585 (1989)
4. Bredin, H., Chollet, G.: Making talking-face authentication robust to deliberate imposture. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2008), pp. 1693–1696 (2008)
5. Duchon, J.: Interpolation des fonctions de deux variables suivant le principe de la flexion des plaques minces. R.A.I.R.O. Analyse numérique 10, 5–12 (1976)
6. Fauve, B., Bredin, H., Karam, W., Verdet, F., Mayoue, A., Chollet, G., Hennebert, J., Lewis, R., Mason, J., Mokbel, C., Petrovska, D.: Some results from the biosecure talking face evaluation campaign. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2008), vol. 1, pp. 4137–4140 (2008)
7. Lienhart, R., Maydt, J.: An extended set of haar-like features for rapid object detection. In: Proceedings of the International Conference on Image Processing, vol. 1, pp. I–900–I–903(2002)
8. Popovici, V., Thiran, J., Bailly-Bailliere, E., Bengio, S., Bimbot, F., Hamouz, M., Kittler, J., Mariethoz, J., Matas, J., Messer, K., Ruiz, B., Poiree, F.: The BANCA database and evaluation protocol. In: Kittler, J., Nixon, M.S. (eds.) AVBPA 2003. LNCS, vol. 2688, pp. 625–638. Springer, Heidelberg (2003)
9. Sanderson, C., Paliwal, K.K.: Fast feature extraction method for robust face verification. IEE Electronics Letters 38(25), 1648–1650 (2002)
10. Stylianou, Y., Cappe, O., Moulines, E.: Continuous probabilistic transform for voice conversion. IEEE Transactions on Speech and Audio Processing 15(6), 131–142 (1998)
11. Tekalp, A., Ostermann, J.: Face and 2-d mesh animation in mpeg-4. Image Communication Journal 15(4-5), 387–421 (2000)
12. Verdet, F., Hennebert, J.: Impostures of talking face systems using automatic face animation. In: Proceedings of the IEEE Conference on Biometrics: Theory, Applications and Systems (BTAS 2008) (2008)