INTERSPEECH 2004 – ICSLP
8th International Conference on Spoken
Language Processing
ICC Jeju, Jeju Island, Korea
October 4-8, 2004

“Liveness” Verification in Audio-Video Authentication

Girija Chetty and Michael Wagner

Human-Computer Communication Laboratory
University of Canberra, Australia
g.chetty@student.canberra.edu.au

Abstract

This paper proposes to use combined acoustic and visual feature vectors to distinguish live synchronous audio-video recordings from replay attacks that use audio with a still photo. Equal error rates below 2 % are achieved using a multidimensional eigenlip representation and EERs of 7% are achieved with a one-dimensional lip-opening ratio.

1. Introduction

Depending on the type of application and the verification protocol, speaker recognition can be vulnerable to replay attack, i.e. the replaying of a prerecorded sample of the speech of the person that is to be defrauded. A possible scenario for such an attack on a text-dependent speaker verification system is a situation where an attacker records a client speaking the correct passphrase and later replays the passphrase to the system in order to defraud the client. Even if an attacker obtains the correct passphrase by eavesdropping, but without recording it, an attempt to defraud a text-dependent speaker verification system with a known passphrase is more likely to succeed than an attempt without knowledge of the passphrase.

Different algorithms and protocols have been devised to protect speaker verification systems against replay attack. In text-independent speaker verification [1], the system acquires training data for the whole range of speech sounds from each client in order to construct a robust client model that is representative of that client’s speech sounds in all possible contexts. During the verification phase, the system is then able to verify the identity of a client irrespective of the text spoken by the claimant.

In text-prompted speaker verification [2], the system acquires training data for several possible passphrases from each client. During the verification phase, the system chooses one of the trained passphrases and asks (“prompts”) the client to speak that particular passphrase, thereby reducing the chance of success for a replay attack, because it is clearly more difficult for an attacker to collect recordings of all of the client’s passphrases.

Despite recent efforts to facilitate the training of client models by adapting universal speaker models with relatively small amounts of client-specific training data [3], the problem remains that with equivalent amounts of training data, text-independent and text-prompted speaker verification systems are more error-prone than text-dependent speaker verification systems. Or, to put it differently, in order to perform with equally low error rates, text-independent and text-prompted speaker verification systems require larger training times for each client than text-dependent speaker verification systems.

More recently, the audio-video recording of the speaking face has been proposed as an authentication mechanism with the potential to deliver low error rates due to the presence of largely complementary signals in the audio and video channels [4]. Audio-video recording of the speaking face has proved particularly successful in situations where single-mode speaker recognition is undermined by high levels of environmental or channel noise and where single-mode face recognition is unreliable in conditions of low or varying illumination [5].

It was shown by Yehia et al. [6] that there is a significant correlation between the acoustics of speech and the corresponding facial movements. Therefore, authentication by audio-video recording of the speaking face also provides an important new defence against replay attack because it enables the system to verify whether the audio-video recording (a) corresponds to the client’s voice, (b) corresponds to the client’s face, and (c) was recorded synchronously, i.e. with the correct correspondence of speech sounds (phonemes) and facial constellations (visemes). Thus a bimodal audio-video authentication system is able, in principle, to provide text-dependent face-voice authentication and at the same time thwart a replay attack where the attacker replays a prerecorded passphrase in conjunction either with the impostor’s face or with a still photograph of the client’s face – a vulnerability which is clearly more likely than that of a full audio-video recording of a client speaking the correct passphrase.

One approach to liveness verification, which was proposed by Choudhury et al [7], uses a 3-dimensional model of the speaker’s head in order to track the depth coordinates of key facial “landmarks” like pupils, nose tip etc. A replay attack using a still photograph of the face would be detected because all key features would be moving forward or backward together whereas a live video of a rotating head would have some landmarks move forward while others move back etc. The system would still be vulnerable, however, to being attacked with a non-synchronous video or a video unrelated to the audio signal.

This paper presents an algorithm, which allows the verification of the “liveness” of an audio-video recording by measuring the correspondence between the audio and video channels. Acoustic parameters of the speech signal are combined with visual parameters of the lip region of the speaking face and a Gaussian mixture model (GMM) is built for the client’s combined audiovisual feature vectors. During the verification phase, the “liveness” of an audio-video recording is ascertained by distinguishing between a properly synchronous recording and a “fake” recording that consists of an audio channel in conjunction with a still photograph of the client.

The following sections of the paper describe the speaking-face data, the audiovisual feature vectors, the audiovisual GMMs and the results of experiments distinguishing “true” lip-synchronous audio-video recordings of the speaking face from “fake” recordings of client speech in conjunction with photographs of the client’s face.

2. Speaking-Face Data

A database of 8 repetitions of a 4-digit spoken passphrase was recorded for two male speakers and one female speaker. The recordings took place in a typical office environment with air-conditioning and computer-fan noise and with varying illumination conditions through an outside window. The whole face was framed by the camera.

The recordings were made with a DV-PAL camcorder at 25 video frames per second. Frames were resampled to 180x120 pixels per frame and converted to a grey scale. The audio was resampled to 16,000 samples per second.

Each passphrase was completed in approximately 4 seconds resulting in about 100 video frames. Six repetitions were used for training the system and two repetitions were used for testing.

3. Audiovisual Feature Vectors

The audio signal was pre-emphasised and Hamming-windowed for a sequence of one-third overlapping 30ms frames yielding a frame rate of *50Hz*. An acoustic feature vector was determined for each frame by warping 512 spectral bands into 30 mel-spaced bands and computing the first 8 mel-frequency cepstral coefficients (MFCC). Since the goal of the experiment is *not* to provide a high-resolution representation of the speaker’s acoustic space, but solely to verify the correspondence between speech sound and facial configuration, the first 8 MFCC were deemed sufficient and indeed preferable to the construction of a larger-size acoustic feature vector. For the same reason, no derivatives were included in the acoustic feature vector. Cepstral mean normalization was performed in order to compensate for the varying environmental conditions.

A lip “region of interest” (ROI) was determined for each video frame and lip motion features were extracted from each lip ROI, i.e. at a rate of 25Hz. The ROI forms a 31x21-pixel block, which is normalised by translation, rotation and scaling, and contains an area defined by the width of the mouth and a constant ratio of vertical-to-horizontal extent.

The experiment was undertaken with 2 different visual features. Firstly, each lip ROI was histogram-normalised and principal component analysis (PCA) was undertaken on the 31x21=651 pixels of the ROI similarly to the eigenface approach proposed by Turk and Pentland [8] or to the eigenlip approach proposed by Bregler and Konig [9]. The lip ROI from each video frame was projected onto the n-dimensional subspaces generated by the first n eigenvectors with n=1…40, which represent the directions of maximum variance of the data in R^{31x21}. Therefore, each lip ROI yielded a visual feature vector with between 1 and 40 components. Figure 1 shows the normalisation of a video frame and extraction of the lip ROI for a male subject and Figure 2 shows six original visemes and their projections onto n-dimensional subspaces generated by the first n eigenvectors with n = 1, 2, 4, 8, and 10.

A second visual feature was extracted for each video frame by determining the ratio of lip height and lip width from the lip ROI, yielding a visual feature “vector” of dimension 1. Both kinds of visual feature vectors were used in the subsequent experiments.

Audio and visual feature vectors were then combined at the feature level (“feature fusion” or “early fusion”). Since the audio frame rate is 50Hz and the video frame rate is 25Hz, two audio feature vectors (2x8 components) are combined with one visual feature vector (between 1 and 40 components).

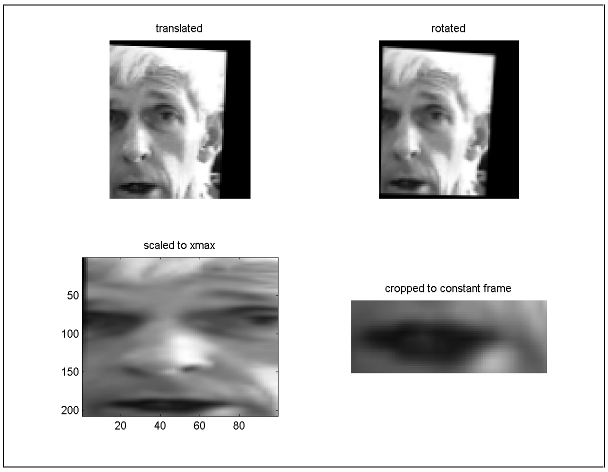

Figure 1: Image normalisation

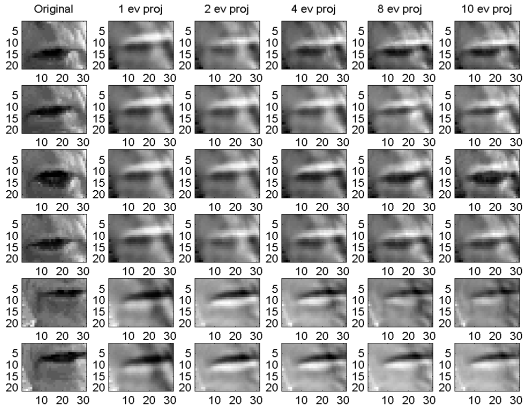

Figure 2: Eigenlip projections for 6 different visemes onto n-dimensional eigenlip subspaces (n = 1, 2, 4, 8, 10)

Given that the number of the audio components of the audiovisual feature vector is constant at 2x8=16, the total size of the audiovisual feature vector depends on the chosen representation of the lip ROI, which can vary between 1 and 40 components. Therefore the audiovisual feature vector is varied in size between 2x8+1=17 and 2x8+40=56. Figure 3 shows an example of a 36-dimensional audiovisual feature vector.

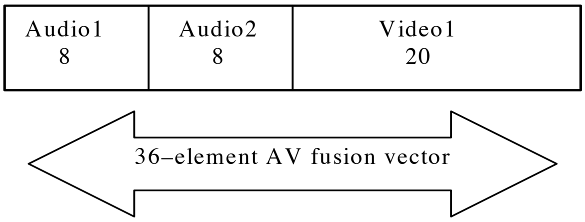

Figure 3: Example 36-D audiovisual feature vector

4. Liveness Verification Using Eigenlips

From the training data for each client, a Gaussian mixture model of the client's audiovisual feature vectors was built, reflecting the probability densities for the combined phonemes and visemes in the audiovisual feature space. The number of mixtures was varied between 5 and 10.

In the first part of the test phase, the client's live test recordings were evaluated against the client's model λ by determining the log likelihoods log p(X|λ) of the time sequences X of audiovisual feature vectors under the usual assumption of statistical independence of subsequent feature vectors.

In the second part of the test phase, a number of "fake" recordings were constructed by combining the sequence of audio feature vectors from each test utterance with ONE visual feature vector chosen from the sequence of visual feature vectors and keeping that visual feature vector constant throughout the utterance. Such a fake sequence represents an attack on the authentication system, which is carried out by replaying an audio recording of the client's utterance while presenting a still photograph to the camera. Four such fake audiovisual sequences were constructed from different still frames of each client test recording. Log-likelihoods log p(X'|λ) were computed for the fake sequences X' of audiovisual feature vectors against the client model λ. In order to obtain suitable thresholds to distinguish live recordings from fake recordings, detection error tradeoff (DET) curves and equal error rates were determined for the different eigenlip projections and for the lip-opening ratio.

The results for liveness verification using eigenlip projections of between 1 and 40 dimensions are shown for the 3 speakers in Table 1.

Table 1 shows the resulting equal error rate for the 3 subjects for 1- to 40-dimensional eigenlip projections of the lip ROI and for 5 Gaussian mixtures. Similarly, Table 2 shows the results for 10 Gaussian mixtures.

n	1	2	4	8	10	15	20	30	35	40
S1	23.7	19.9	17.6	10.5	8.0	6.5	5.9	4.4	3.5	2.9
S2	21.2	17.1	15.4	9.7	6.9	6.0	5.0	3.5	2.6	2.1
S3	22.3	18.6	16.4	10.0	7.0	6.3	5.0	4.0	3.1	2.6

Table 1. Liveness verification equal-error rates for the 3 speakers with 5 Gaussian mixtures and eigenlip projections of between n=1 and n=40 dimensions.

n	1	2	4	8	10	15	20	30	35	40
S1	22.6	18.8	16.7	9.6	7.1	4.6	5.0	3.5	2.6	2.0
S2	20.3	16.1	14.5	8.8	6.0	5.1	4.1	2.6	1.7	1.2
S3	21.4	17.5	15.2	9.1	6.2	5.4	4.2	3.1	2.2	1.7

Table 2. Liveness verification equal-error rates for the 3 speakers with 10 Gaussian mixtures and eigenlip projections of between n=1 and n=40 dimensions.

The results show that a reliable verification of phoneme-viseme correspondence is achieved with a 40-dimensional eigenlip projection of the lip ROI.

The DET curves corresponding to Table 1 are shown in Figure 4 for 10-dimensional eigenlip projections, and in Figure 5 for 40-dimensional eigenlip projections.

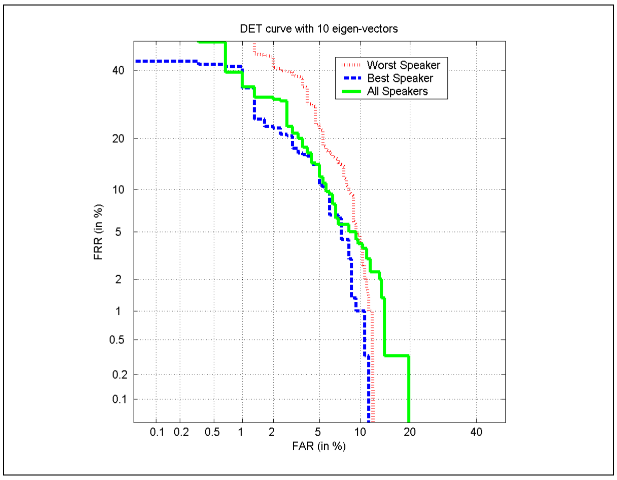

Figure 4: Liveness verification DET curve for 5 Gaussian mixtures and 10-dimensional eigenlip projection

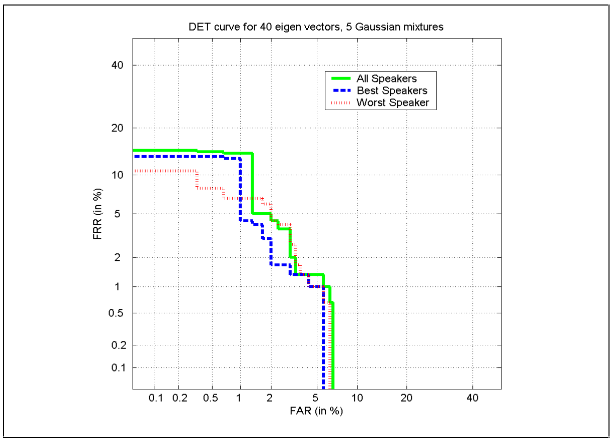

Figure 5: Liveness verification DET curve for 5 Gaussian mixtures and 40-dimensional eigenlip projection.

The DET curves corresponding to Table 2 are shown in Figure 6 for 10-dimensional eigenlip projections, and in Figure 7 for 40-dimensional eigenlip projections.

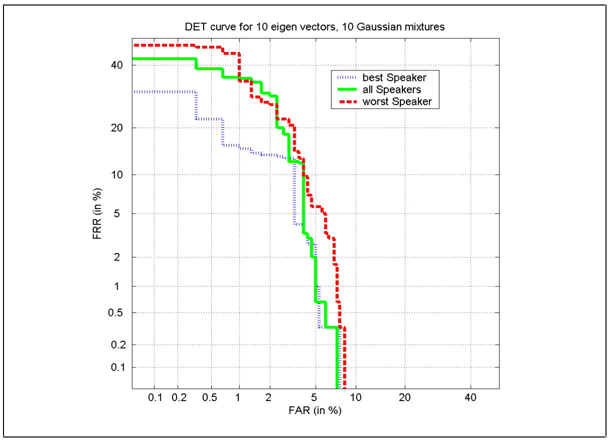

Figure 6: Liveness verification DET curve for 10 Gaussian mixtures and 10-dimensional eigenlip projection.

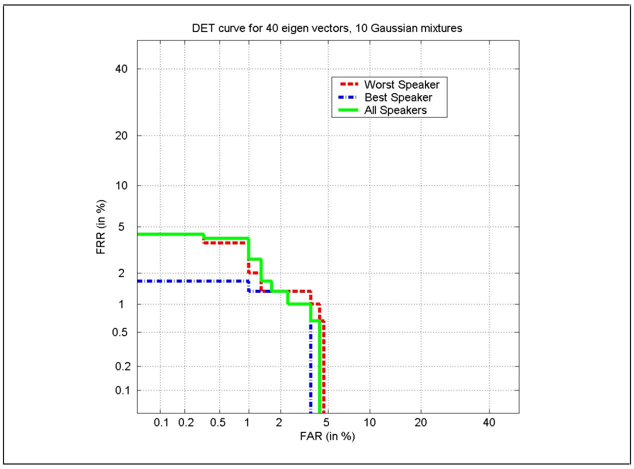

Figure 7: Liveness verification DET curve for 10 Gaussian mixtures and 40-dimensional eigenlip projection.

The results show that with audiovisual data collected in a difficult environment of acoustic noise and illumination changes, synchronous audio-video recordings can be distinguished from still-frame replay attacks with equal error rates of better than 2 per cent. In each graph the "All Speakers" curve indicates the possibility of a speaker-independent threshold to guard against replay attack with a still frame.

5. Liveness Verification Using Lip Opening

Instead of using a visual feature vector consisting of eigenlip projections, the second experiment was carried out with the 1-dimensional visual feature "vector" of the measured ratio between vertical and horizontal lip extent for each video frame. Again, a GMM was constructed from the live audio-video recordings of each client's training data and verification was carried out with the live recordings of each client's test sessions and with fake recordings constructed from live audio and constant still frames.

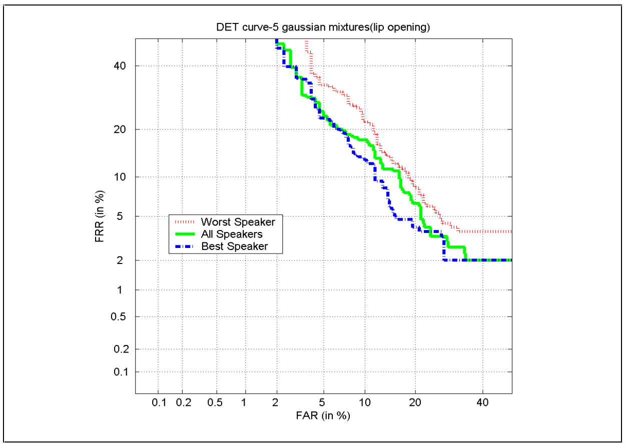

Figure 8: Liveness verification DET curve for the lip-opening ratio with 5 Gaussian mixtures.

The DET curves for the verification results with 5 and 10 mixtures are shown in Figure 8 and 9 respectively.

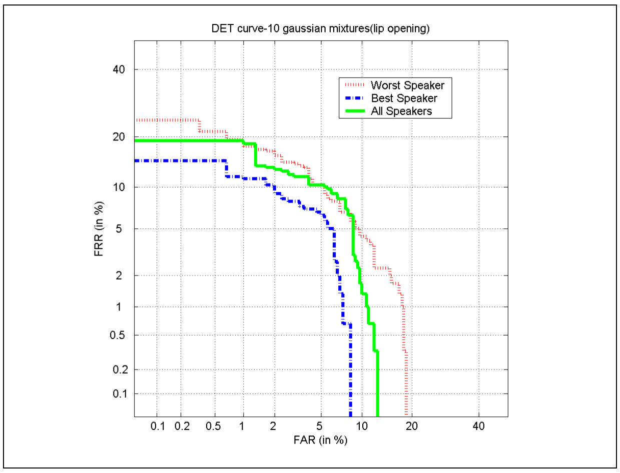

Figure 9: Liveness verification DET curve for the lip-opening ratio with 10 Gaussian mixtures.

With equal error rate of about 12% for 5 mixtures and 7% for 10 mixtures, it is clear that the lip-opening ratio as measured from the video frames is inferior to eigenlip representations of the lip ROI for capturing the correspondence between phonemes and visemes of an audio-video recording.

6. Conclusions

It has been shown that a combined audiovisual feature vector consisting of 8 MFCCs per audio frame and either an eigenlip projection of the lip region of interest or a lip-opening ratio can distinguish synchronous audio-video recordings from replay attacks, which use the audio together with a still photo.

7. References

[1] Furui, S., ``Recent Advances in Speaker Recognition", *Pattern Recognition Letters*, Vol. 18, 1997, pp. 859-872.

[2] Matsui, T and Furui, S., *"Concatenated Phoneme Models for Text-variable Speaker Recognition", in Proc. Int. Conf. On Acoustics, Speech and Signal Processing*, 1993, pp. II 391-394.

[3] Reynolds, D.A., Quatieri, T.F.and Dunn, R.B., "Speaker Verification Using Adapted Gaussian Mixture Models", *Digital Signal Processing*, Vol. 10, 2000, pp. 19-41.

[4] Chibelushi, C.C., Deravi, F. and Mason, J.S., "Voice and Facial-Image Integration for Speaker Recognition, IEEE Int. Symp. on Multimedia Technologies and Future Applications, Southampton, UK, 1983.

[5] Sanderson, C and Paliwal, K.K., "Fast Features for Face Authentication under Illumination Direction Changes", in Pattern Recognition Letters, Vol.24, 2003, pp 2409-2419.

[6] Choudhury, T., Clarkson, B., Jebara, T. and Pentland, A., Multimodal Person Recognition Using Unconstrained Audio and Video, in Audio- and Video-Based Biometric Person Authentication, 1999.

[7] Yehia, H., Rubin, P, Vatikiotis-Bateson, E., "Quantitative Association of Vocal-Tract and Facial Behavior" in *Speech Communication*, Vol. 26 (1998), pp. 23-43.

[8] Turk, M and Pentland, A., Eigenfaces for Recognition, J. Cognitive Neuroscience, Vol. 3 (1991), pp. 71-86.

[9] Bregler, C.and Konig, Y., "Eigenlips" for Robust Speech Recognition, Proc. Int. Conf. On Acoustics, Speech and Signal Processing, ICASSP-1994.