Liveness Verification in Audio-Video Speaker Authentication

Girija Chetty and Michael Wagner

Human Computer Communication Laboratory
School of Information Sciences and Engineering
University of Canberra,
Australia
g.chetty@student.canberra.edu.au

Abstract

In this paper we propose liveness verification for audio-video speaker authentication systems to guard against possible replay attacks that employ pre-recorded audio and/or still face samples from a photo. The technique uses a fusion of acoustic features and visual features, automatically extracted from the lip region, to differentiate synchronous audio-video presentations from asynchronous replay attacks. Experiments with the multi-modal VidTIMIT database achieve liveness verification equal-error rates of less than 1% using lip-voice feature vectors comprising mel-cepstral coefficients, eigenlips and lip region measurements.

1. Introduction

Speech and face recognition systems are two well–researched domains in biometric person authentication (Broun, Zhang, Mersereau and Clements, 1999). The use of speech to identify a person has the advantages of requiring little custom hardware and being non-intrusive. However, single-mode speaker verification systems often perform poorly in noisy environments and when testing and enrolment conditions are less than perfectly matched. When incorporating a visual modality, namely a facial image, such systems perform generally better and are particularly more robust to varying environmental conditions. Single-mode face recognition is also a very topical research area because of many potential applications in areas such as access control, automated crowd surveillance, law enforcement, information safety, multimedia communication and human-machine interface. Such systems often perform poorly in conditions of varying illumination and when testing and enrolment conditions are not well matched.

Audiovisual authentication, based simultaneously on the voice and face of a person, offers a number of advantages over both single-mode speaker verification and single-mode face verification (Stork and Hennecke, 1996), including enhanced robustness against variable environmental conditions. One of the most significant advantages of combined face-voice recognition is the decreased vulnerability against replay attacks, which could take the form of presenting either a voice recording to a single-mode speaker verification system or a still photograph to a single-mode face verification system. Without doubt, replay attacks on combined face-voice recognition systems would be more difficult. However, since most proposed combined audiovisual verification systems verify a person's face statically, these systems remain vulnerable to replay attacks that present pre-recorded audio together with a still photograph. To resist such attacks, audio-video authentication should include verification of the "liveness" of the audio-video data presented to the system. Until now, although there has been much published research on the liveness, for example, of fingerprints, research on liveness verification in face-voice authentication has been very limited (Frischholz and Werner, 2003).

Lip-motion is intrinsic to speech production and the inclusion of lip dynamics features along with acoustic speech features not only reflects static and dynamic acoustic and visual features, but also provides proof of liveness. It was shown by Matsui and Furui (1993) that there is a significant correlation between the acoustics of speech and the corresponding facial movements and that the acoustic dynamic features of a speaking face are inextricably linked to the visible articulatory

Proceedings of the 10th Australian International Conference on Speech Science & Technology
Macquarie University, Sydney, December 8 to 10, 2004. Copyright, Australian Speech Science & Technology Association Inc.

Accepted after full paper review

movements of the teeth, tongue and lips. Therefore, using dynamic visible features of the lip region for liveness verification in audiovisual person authentication seems to be a natural, simple and effective technique to address replay attack issues as compared with more complex methods based on 3D depth and pose estimation, or challenge-response techniques (Choudhury, Clarkson, Jebara and Pentland, 1999; Frischholz and Werner, 2003).

The dynamic lip features are determined by using a lip tracking system, which locates the lips in the video sequence and then extracts the dynamic lip parameters. Subsequently, these features are fused synchronously with the acoustic features extracted from the corresponding acoustic data in order to ascertain liveness and to establish the person’s identity.

The detection and tracking of facial visual features is a fundamental and challenging problem in computer vision. As in most vision tasks, two problems need to be solved for lip detection: Spatial segmentation, that is finding and tracking the lips, and recognition, that is estimating the relevant features about the “lip configuration” for classification.

The method used to find the lip region with the facial images is described in detail in Chetty and Wagner (2004b). Essentially, it uses a face detector that is based on finding a candidate area of a typical facial colour, i.e. chrominance value, and of a defined shape. The algorithm then analyses the region of the face where the mouth is normally expected to be and determines the lip area on the basis of hue saturation.

The main objective of most of the methods proposed for lip feature extraction to date is to assist in a speech recognition task by including the visual modality. Though some of the conclusions could be extended to a speaker verification context, not much can be said about the “liveness” aspects of the results of previous experiments. However, by appropriate adaptation of some of the above-mentioned techniques, liveness verification could be made more powerful. One such simple and effective adaptation is to implement the multimodal fusion with synchronous and early integration of acoustic and moving lip region feature vectors before classification. This will require building client and impostor models based on synchronised audio and dynamic visual features, as well as appropriate interpolation of the feature vectors acquired at different rates. With this, the liveness of an audio-video sample would be ascertained more accurately in the verification phase by differentiating a “live” client sample and a “fake” sample consisting of an audio signal in conjunction with a still photograph.

There are not many methods reported in the literature on the conduct of such “liveness” experiments and on the error rates achieved by others. A preliminary work carrying out liveness verification experiments with limited data has been reported in Chetty and Wagner (2004a). In the current paper, experiments carried out to verify liveness in face-voice data with an extensive multi-modal person authentication database (Sanderson and Paliwal, 2003) are reported. The results of liveness experiments conducted allow an improvement in error rates, achieved from 2-7% in previously reported results (Chetty and Wagner, 2004a), to less than 1%. The next four sections describe the details of “liveness” experiments carried out on the VidTIMIT database.

2. Audiovisual Data

The multimodal person authentication database VidTIMIT (Sanderson and Paliwal, 2003) was used for conducting all three experiments in this study. The VidTIMIT database consists of video and corresponding audio recordings of 43 people (19 female and 24 male), reciting short sentences selected from the NTIMIT corpus. The data were recorded in 3 sessions, with a mean delay of 7 days between Session 1 and 2, and of 6 days between Session 2 and 3. The mean duration of each sentence is around 4 seconds, or approximately 100 video frames. A broadcast-quality digital video camera in a noisy office environment was used to record the data. The video of each person is stored as a sequence of JPEG images with a resolution of 512 × 384 pixels (columns × rows), with corresponding audio provided as a monophonic, 16 bit, 32 kHz WAV file.

3. Lip Region and Key Points Extraction

The visual feature extraction technique we have used is based on automatic face detection and lip localisation using skin-colour analysis and morphological segmentation. The detection is done in the first frame of the video sequence and tracking of the lip region in subsequent frames is done by projecting the marker from the first frame. This is followed by measurements on lip region boundaries based on pseudo-hue edge detection and tracking technique. The advantage of this approach is that facial feature extraction is based on exploiting alternate colour spaces in a simple and very powerful way as compared to methods based on deformable templates, snakes, and pyramid images. Moreover, the method can be easily extended to detection/tracking of facial features with multiple faces, and/or faces with natural/complex backgrounds, though the experiments reported here were conducted on the VidTIMIT database, containing only single faces with a uniform and structured background.

The details of the complete face detection, lip localisation and lip feature extraction stages are found in Chetty and Wagner (2004b). Here, some of the main results obtained in each of the stages are included. The face detection sub-system consists of three image processing stages. The first stage is to classify each pixel in the given image as a skin or non-skin pixel. The second stage is to identify different skin regions in the image by using connectivity analysis. The last stage is to determine whether any of the skin regions represents a face based on template matching with an average face and on the aspect ratio constraints of a face.

Figure 1 shows the image processing steps for the face detection stage based on skin colour segmentation and thresholding.

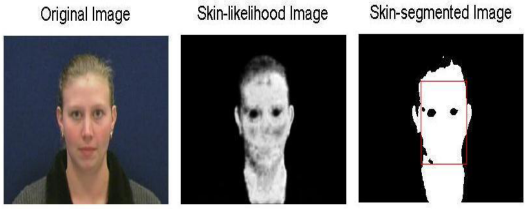

Figure 1: Skin likelihood/skin segmented image.

Once the face region is localised as shown in Figure 1, the lip region is detected using hue-saturation thresholding. Figure 2 shows an example of lip region localisation based on Hue-Saturation thresholding.

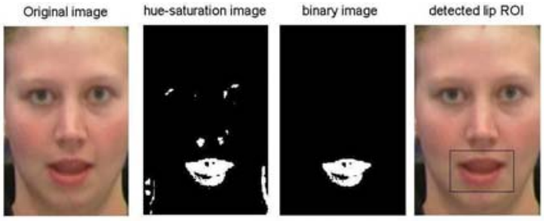

Figure 2: Lip region localisation using hue/saturation thresholding.

To derive the lip dimensions and key point extraction within the lip region, the algorithm detects “pseudo-edges” based on combining pseudo hue-colour and intensity information. From the key points extracted, the major dimensions of the lips can be derived, and subsequently used for the audiovisual fusion vector. An illustration of the geometry of the extracted features is shown in Figure 3. The major dimensions extracted are the inner lip width (*w_1*), outer lip width (*w_2*), inner lip height (*h_1*), outer lip height (*h_2*), upper and lower lip widths (*h_3 and h_4*) and the distance *h_5 =h2/2* between the mid-horizontal line and the upper lip. The technique could be extended to include the visibility of tongue and teeth. However, that extension was not included in the current experiments.

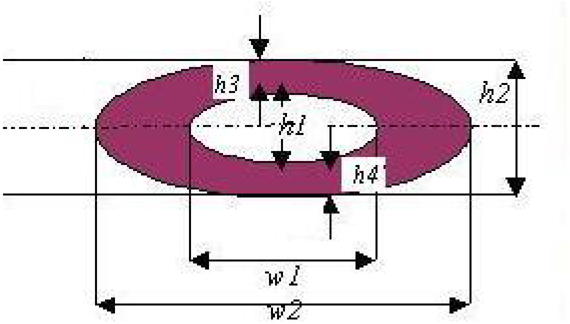

Figure 3: Lip-region and lip-parameters.

4. Audiovisual Feature Vector

The experiments were undertaken with different types of visual feature vectors. The structure of the acoustic feature vector was kept constant for all the experiments and comprises 8 mel-frequency cepstral coefficients (MFCC). The MFCC acoustic vectors were obtained by pre-emphasising the audio signal and processing it using a 30ms Hamming window with one-third overlap, yielding a frame rate of 50 Hz. An acoustic feature vector was determined for each frame by warping 512 spectral bands into 30 mel-spaced bands, and computing the 8 MFCCs. Cepstral mean normalisation (Atal, 1974) was performed in order to compensate for the varying environmental noise and channel conditions. The spectrograms of the audio signals and the corresponding lip motions for a female and a male subject in VidTIMIT are shown in Figure 4, while Figure 5 shows the same spectrograms together with the variations of the inner lip ratio h_1/w_1 and the outer lip ratio h_2/w_2 for the same two subjects. It can be observed from Figures 4 and 5, that the audio signatures for male and female subjects for the same utterance are distinct. Figure 5 also shows that the inner and outer lip ratio variations can be very distinct.

The three experiments were conducted with different types of visual feature vectors. In the first experiment, the visual feature vector comprised the 6 lip parameters *h_1, h_2, h_3, h_4, w_1*, and *w_2* as described in the previous section. In the second experiment, an eigenlip representation of the lip region, based on principal component analysis (Turk and Pentland, 1991; Bregler and Konig, 1994), was used. Before performing principal component analysis of the lip region images, each image was normalised by translation, rotation and scaling operations (Chetty and Wagner, 2004a).

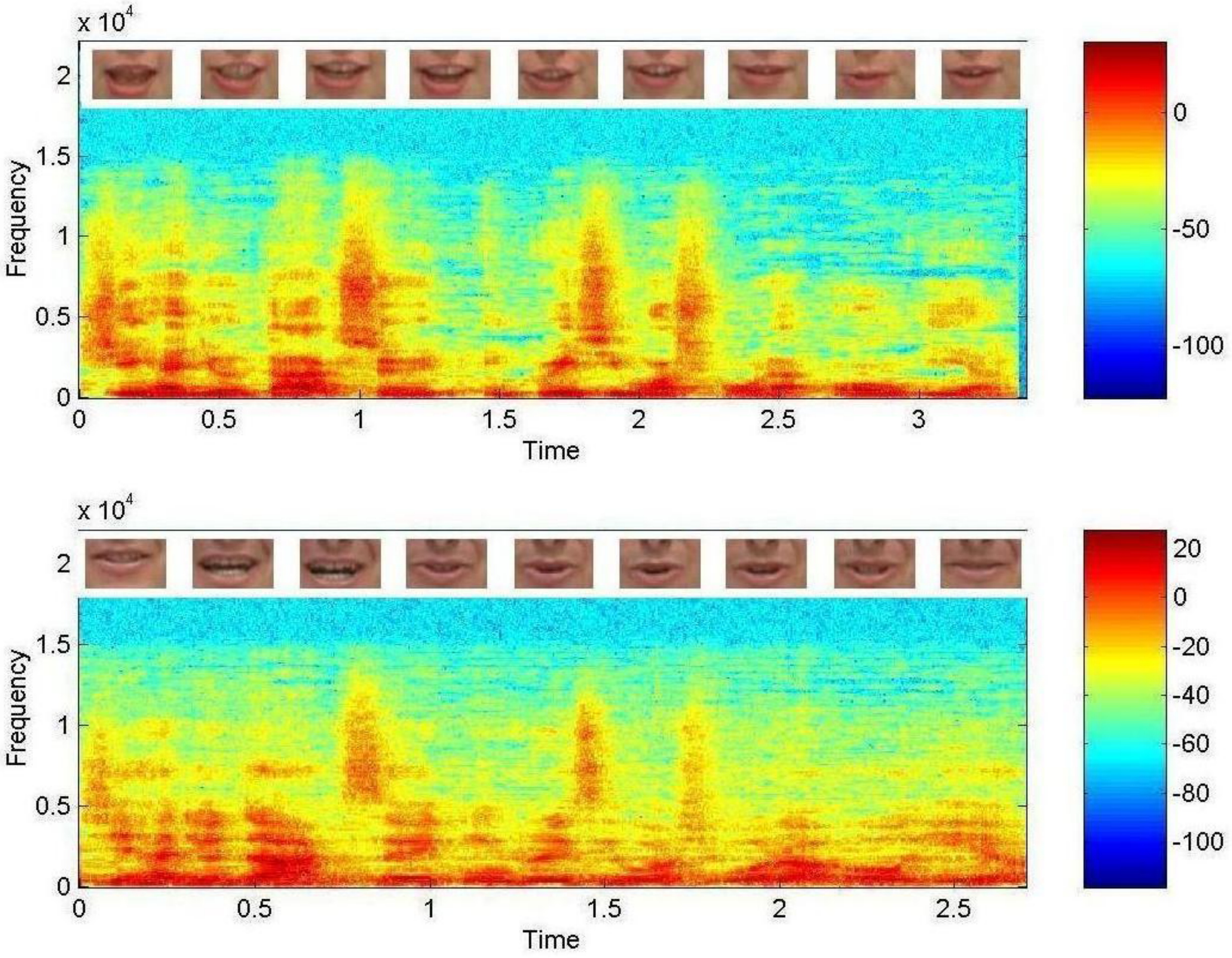

Figure 4: Spectrograms and corresponding lip variations of the utterance *She had your dark suit in greasy wash water all year*, spoken (a) by a male subject, and (b) by a female subject.

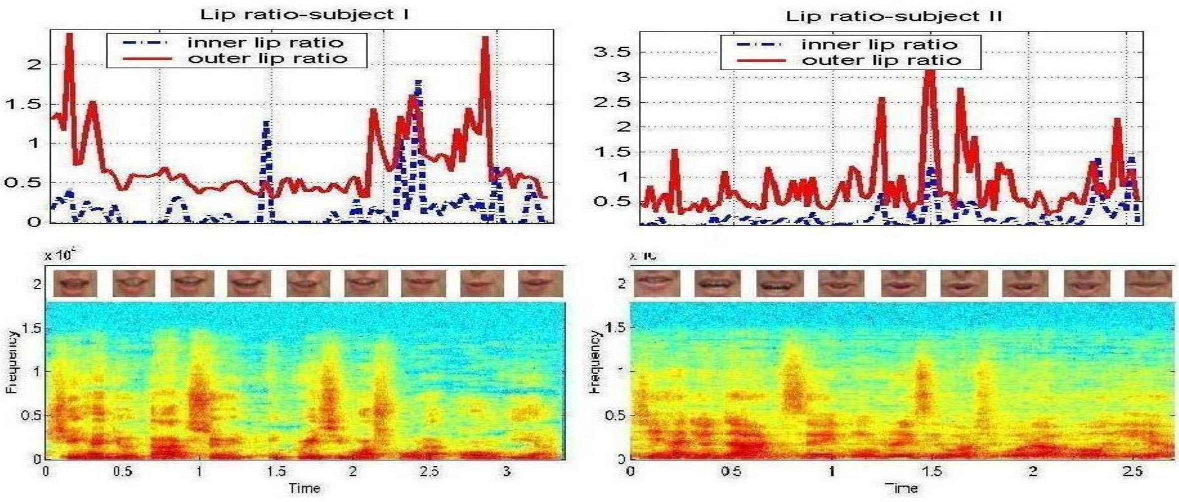

Figure 5: Spectrograms and corresponding lip-ratio functions for the utterance *She had your dark suit in greasy wash water all year*, (a) for the male subject, and (b) for the female subject.

In some preliminary experimentation, it was found that good classification accuracy could be achieved with 8-10 eigenlip projections. The third experiment comprised a combination of both eigenlips and the geometric lip parameters. Figure 6 shows the audio-visual feature vector structure for the three experiments. Since audio features occur at twice the rate (50Hz) of the corresponding video features (25Hz), each audiovisual feature vector comprises 2 audio feature vectors (2x8 components) and one video feature vector (6, 10 or 16 components). For Experiment I, the video feature vector comprises the 6 elements of lip geometry. For Experiment II, the video feature vector comprises 10 eigenlip projection coefficients. And for Experiment III, the geometric lip parameters and the eigenlip coefficients are combined to form a 16-element video feature vector.

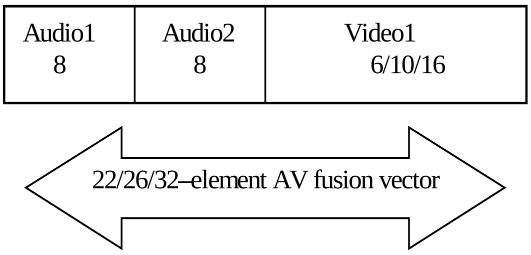

Figure 6: Audiovisual feature vector structure.

5. Liveness Verification

A Gaussian mixture model is trained for each client using the audio-visual feature vectors extracted from the face-voice data of the VidTIMIT database. Sessions 1 and 2 of the database were used for training and Session 3 was used for testing. For each feature extraction method, client models of 10 Gaussian mixtures were generated, reflecting the probability densities for the combined phonemes and visemes in the audiovisual feature space.

For testing purposes, clients' live test recordings were evaluated against the client's model λ by determining the log likelihoods $\log p(X|\lambda)$ of the time sequences X of audiovisual feature vectors under the usual assumption of statistical independence of subsequent feature vectors.

For testing replay attacks, 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 $\log p(X'|\lambda)$ 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 trade-off (DET) curves and equal error rates (EER) were determined. The testing process of a client's live recordings and simulated replay-attacks were repeated with the different visual feature vectors for experiments I – III. The DET curves showing EERs achieved for liveness verification experiments are shown in Figures 7, 8 and 9.

Figure 7 shows the DET curves for Experiment I with the visual vector consisting of 6 lip parameters. The three curves show the error rates for the best speaker, for all speakers and for the worst speaker. The resulting EER for this experiment is about 5%.

Figure 8 shows the DET curves for Experiment II with the visual vector consisting of 10 eigenlip projections. Here an EER of 2% is achieved for all speakers, with the best speaker's EER being 1% and the worst speaker's EER being 2.5%.

The DET curves for Experiment III are shown in figure 9. In this experiment, the 6 lip parameters (Experiment I) were combined with the 10 eigenlip projections (Experiment II), and for the combined visual feature vector, the EER for all speakers was 0.8% with a best-speaker EER of 0.5% and a worst-speaker EER of 1%. Further experiments showed that error rates based on 8 Gaussian mixtures were slightly higher than those based on 10 Gaussian mixtures, but that more than 10 mixtures did not result in a significant error rate reduction.

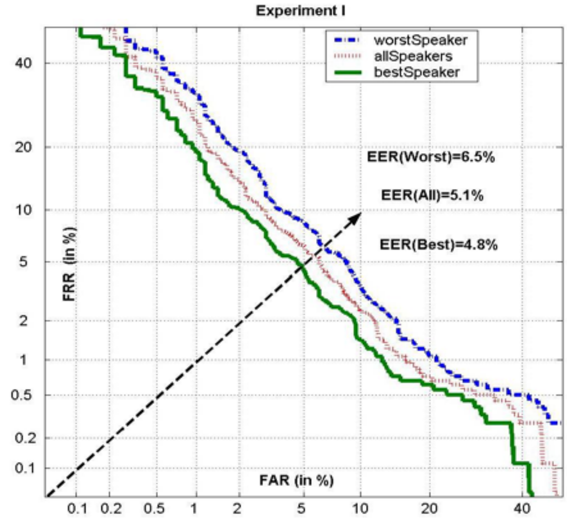

Figure 7: DET curves for Experiment I using a 6-element video vector of geometric lip parameters.

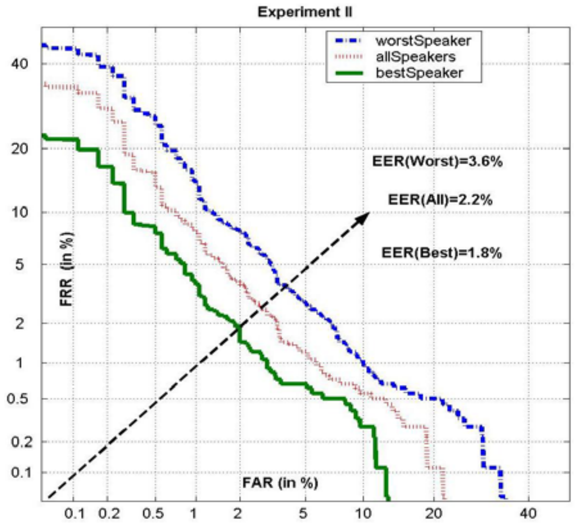

Figure 8: DET curves for Experiment II using a 10-element video vector of eigenlip projections.

6. Conclusions

The results of the liveness verification experiments have shown that with an audiovisual fusion vector combining eigenlip projections and geometric lip parameters with MFCC acoustic vectors, liveness verification equal-error rates of less than 1% can be achieved. The technique demonstrates a simple and powerful method of verifying liveness and thwarting replay attacks. Further experiments will include investigations of more elaborate methods for liveness checks on more extensive databases.

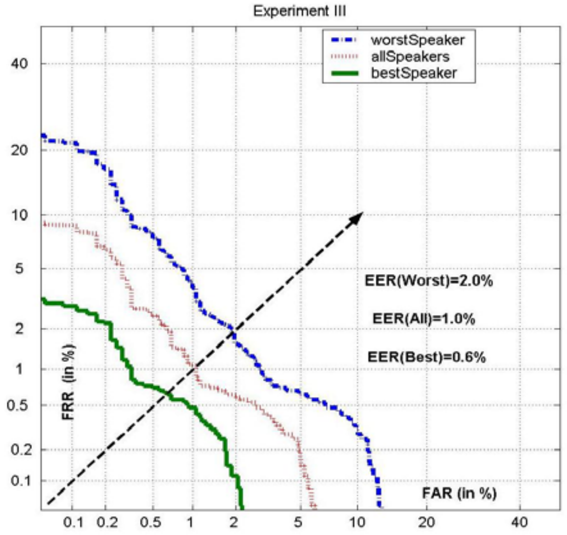

Figure 9: DET curves for Experiment III using a 16-element combined video vector.

7. References

Atal, B. (1974). Effectiveness of linear prediction characteristics of the speech waveform for automatic speaker identification and verification. J Acoustical Society of America 55, 1304-1312.

Bregler, C. and Y. Konig (1994) “Eigenlips” for robust speech recognition. Proc. Int. Conf. On Acoustics, Speech and Signal Processing, ICASSP-1994.

Broun, C.C., X. Zhang, R.M. Mersereau and M. Clements (2002) Automated speechreading with application to speaker verification. Proc. Int. Conf. On Acoustics, Speech and Signal Processing, ICASSP-2002.

Chetty, G. and M. Wagner (2004a) “Liveness” verification in audio-video authentication. Accepted for Int Conf on Spoken Language Processing, ICSLP-2004.

Chetty, G. and M. Wagner (2004b) Automated lip feature extraction for liveness verification in audio-video authentication. Submitted for publication, 2004b.

Choudhury, T., B. Clarkson, T. Jebara and A. Pentland (1999) Multimodal person recognition using unconstrained audio and video, in audio- and video-based biometric person authentication. Int Conf on Audio and Video-Based Biometric Person Authentication, AVBPA-1999.

Frischholz, R. and A. Werner (2003) Avoiding replay attacks in a face-recognition system using head pose estimation. Proc IEEE Int Workshop on Analysis and Modeling of Faces and Gestures, AMFG’03.

Matsui, T and S. Furui (1993) Concatenated phoneme models for text-variable speaker recognition. Proc. Int. Conf. on Acoustics, Speech and Signal Processing, ICSLP-1993, 391-394.

Sanderson, C. and K.K. Paliwal (2003) Fast features for face authentication under illumination direction changes, Pattern Recognition Letters 24, 2409-2419.

Stork, D. and M. Hennecke (1996) Speechreading by man and machine: data, models and systems. Springer-Verlag, Berlin.

Turk, M and A. Pentland (1991) Eigenfaces for recognition. J. Cognitive Neuroscience 3, 71-86.