A Robust Speaking Face Modelling Approach Based On Multilevel Fusion

Girija Chetty and Michael Wagner
School of Information Sciences and Engineering
University of Canberra, Australia
Tel: +61-2-62012512, Fax: +61-2-6201 5231
E-mail: girija.chetty@canberra.edu.au

Abstract— **In this paper, we propose a robust face modelling approach based on multilevel fusion of 3D face biometric information with audio and visual speech information for biometric identity verification applications. The proposed approach combines the information from three audio-video based modules, namely: audio, visual speech, and 3D face and performs tri-module fusion in an automatic, unsupervised and adaptive manner, by adapting to the local performance of each module. This is done by taking the output-score based reliability estimates (confidence measures) of each of the module into account. The module weightings are determined automatically such that the reliability measure of the combined scores is maximised. To test the robustness of the proposed approach, the audio and visual speech (mouth) modalities are degraded to emulate various levels of train/test mismatch; employing additive white Gaussian noise for the audio and JPEG compression for the video signals. The results show improved fusion performance for a range of tested levels of audio and video degradation, compared to the individual module performances. Experiments on a 3D stereovision database AVOZES show that, at severe levels of audio and video mismatch, the audio, mouth, 3D face, and tri-module (audio+mouth+3D face) fusion EERs were 42.9%, 32%, 15%, and 7.3% respectively for biometric speaker identity verification application.**

I. INTRODUCTION

.Biometrics is a field of security technology devoted to verification or identification of individuals using physiological or behavioral traits. Verification, a binary classification problem, involves the validation of a claimed identity whereas identification, a multi-class problem, involves identifying a user from a set of enrolled subjects; and becomes more difficult as the number of enrollees increases. In audiovideo processing, the video modality lends itself to two modules, the face module and the visual speech module (referred to as the mouth module here).

Most of the speaker recognition systems currently deployed are based on modelling a speaker based on unimodal information, i.e. either audio or visual features. Audio-based identification achieves high performance when the signal-to-noise ratio (SNR) is high. Yet, the performance degrades quickly as the test SNR decreases (referred to as a train/test mismatch), as shown in [1] and elsewhere. Using visual modality in addition to voice information, such as 3D face or 2D region around mouth can make the system robust against SNR degradation, typical of mismatch between and training and test operating environment. However, visual modality based speaker modelling approaches are also susceptible to pose/illumination variation, occlusion, and poor image quality [2], [3]. Further, use of 2D visual speech features extracted from mouth region on its own cannot model a speaking face in its entirety, and normally used along with other biometric modalities. However, mouth region contains important liveness related information, which can be used to detect fraudulent replay attacks involving a still photo of the speaker and replay of audio, or artificially synthesized speaking face.

To combat these limitations of unimodal modules, a modelling approach based on multilevel multimodal fusion approach can be adopted. This can both improve robustness and overall system performance against impostor attacks and fraudulent replay attacks for speaker identity verification application . The audio, face, and mouth modalities contain non-redundant, complementary information about speaker identity.

Further, using 3D face dynamics in addition, allows better modelling of a speaker [4, 5], as we can better quantify the differences between two persons' facial feature variations in 3D as compared to 2D face images. The subtle nuances related to facial expressions and gestures during speaking act that can best discriminate individuals can also be modeled better with 3D face dynamics. From a biometrics point of view, the concept of recognizing a person based on 3D facial motion during speech is attractive; since facial movements comprise a complex sequence of muscle activations, and it is almost impossible to imitate another person's facial speech and expressions, as these characteristics are unique to an individual [6,7]. In experimental psychology, determining the precise role of 3D facial motion in ascertaining identity is still largely unknown, and is being actively pursued [6]. However, some recent findings described in the next section provide considerable motivation for using 3D face models during speech production for identity verification tasks.

This paper is organised as follows. The next section describes the motivation for multilevel speaking face modelling. Section III and IV describe the proposed multilevel fusion approach for speaking face modelling. In Section V, the stereovision audio visual corpus AVOZES used for evaluation is described. In Section VI, the experimental results of extensive evaluations examining the

0-7695-3067-2/07 $25.00 © 2007 IEEE
DOI 10.1109/DICTA.2007.88

individual module performance and multilevel fusion performance for a SIV(Speaker Identity Verification) application scenario are presented. The results are discussed in Section VII and finally in SectionVIII, conclusions from the results are drawn.

II. MULTILEVEL SPEAKING FACE MODELLING

This section discusses the motivation for using 3D information for robust speaking face modelling based on some recent findings in cognitive psychology [6] and psychophysical analysis of visual speech [7]. As with the other forms of biological motion, humans are known to be very sensitive to the realism in the ways the lips move. One of the most significant finding by Yehia, Kuratate, Munhall, and Bateson [8,9] suggest that in order to determine the elements that come to play during analysis of visual speech, it is important to capture the detailed 3D deformations of faces when talking [9]. Yehia, Bateson and Kuratate [8] suggest that a speaking face is a kinematic-acoustic system in motion, and the shape, texture and acoustic features during speech production are correlated in a complex way, and a single neuromotor source controlling the vocal tract behavior is responsible for both the acoustic and the visible attributes of speech production. Hence, for a speaking face not only the facial motion and speech acoustics are correlated, but the head motion and fundamental frequency (F0) produced during speech are also related. Though there is no clear and distinct neuromotor coupling between head motion and speech acoustics, there is an indirect anatomical coupling created by the complex of strap muscles running between the floor of the mouth, through the thyoid bone, attaching to the outer edge of the cricothyroid cartilage, as shown in Figure 1. Due to this indirect coupling, speakers tend to raise the pitch when head goes up while talking. These spatio-temporal correlations can be modeled better with 3D face models instead of just using 2D dimensional face or lip region images.

It has also been shown by several other linguistic and psychophysical researchers [6,7,8,9], that the facial movements play an important role in interpreting spoken conversations and emotions. They occur continuously during social interactions and conversations. They include lip movements when talking, conversational signals, emotion displays and manipulators to satisfy biological needs. Unfortunately when and how a movement appears and disappears, and how co-occurrent movements are integrated (co-articulation effects, for instance) are difficult to quantify.

In addition, the problem of overlaying and blending facial movements in time, and the way felt emotions are expressed in facial activity during speech, have not received much attention. This suggests that during speech production other regions of the face in addition to the lip region are active, and the activities of human facial muscles for this act is far from simply additive.

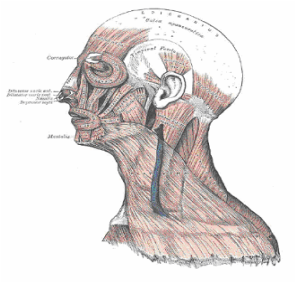

Fig. 1. The facial muscles (from [2])

A typical example would be smiling while speaking. The Zygomatic Major and Minor muscles contract to pull the corner of lip outward, resulting in a smile. The viseme corresponding to the diphthong /oU/ in the word “Hello” requires the contraction of the lip funneler Orbicularis Oris, which drives the lips into a tight, pursed shape. However, the activation of the Zygomatic Major and Minor muscles together with the lip funneler Orbicularis Oris would create an unnatural movement. The activation of a muscle may require the deactivation of other muscles in the jaw and chin region.

These findings from face speech anatomy provide clues that facial movements during speech involve highly complex biomechanics with depth, motion and correlation interactions. Capturing these interactions can truly enhance the performance of face modelling approaches for complex application such as speaker identity verification systems. We propose a novel multilevel modelling approach using features which capture the multiple channels of spatio-temporal facial movements during speech involving 3D, 2D and 1D dynamics and correlations between acoustic-labial articulators as well as other areas of face and head such as jaw, chin, forehead and eyebrows.

III. MULTILEVEL AUDIO VISUAL FUSION

The three individual audio visual modules used for multilevel fusion are described in this section

A. Audio Module

The MFCC features (mel frequency cepstral coefficients) of dimension 16 were extracted from each frame. The energy of each frame was also calculated and used as a 17th static feature. Seventeen first order derivatives or delta features were calculated using WD adjacent static frames, where WD is the delta window size. The delta frames were appended to the static audio features to give an audio feature vector of dimension 34. Cepstral mean normalization [10] was performed on the audio feature vectors (of each audio utterance).

Each speaker is represented by a GMM(Gaussian Mixture Model) model λ. The speaker utterance that is to be classified (the unknown pattern) is a sentence, which is represented by a sequence, O_A, of speech feature vectors or observations denoted by,

$$O_A = \{o_1, o_2, \cdots o_t \cdots, o_{T_A}\} \quad (1)$$

where o_t is the speech observation (frame) at time t and T_A denotes the number of observation vectors in the sentence.

We obtain N class-conditional joint probabilities

$$p(O_A|\lambda) = p(o_1, o_2, \cdots, o_t, \cdots o_{T_A}|\lambda) \quad (2)$$

that the observation sequence O_A was produced by the client speaker model λ.. $p(O_A|\lambda)$ is referred to as the likelihood that O_A was caused by λ. For GMM classifiers, the output scores are in log-likelihood form, denoted by $ll(O_A|\lambda)$.

B. Visual Speech (Mouth) Features module

The visual sentences were modeled using the same GMM methodology described for the audio sentences. Three types of features used are DCT features f_{DCT}, the explicit grid based lip motion features f_{GRD} and the contour based lip motion features f_{CTR} were extracted. The dimension of the visual lip feature vector is 24 with 8 f_{GRD}, 8 f_{CTR} and 8 f_{SHP} features.

For the normal visual mouth DCT features, the mouth ROI .consists of a 49×49 colour pixel block. To account for varying illumination conditions across sessions, the grey scale ROI was histogram equalised and the mean pixel value was subtracted. The two dimensional DCT was applied to the preprocessed gray scale pixel blocks.

For lip motion features, the explicit lip motion feature extraction technique involves the stages of face detection, normalisation and lip region extraction from 2D face images. Grid based motion features were extracted by estimating dense motion over a uniform grid of size $G_x \times G_y$ on the extracted lip region image. We use hierarchical block matching to estimate the lip motion with subpixel accuracy (quarterpel) by interpolating the original lip image using the 6 tap Wiener and bilinear filters specified in H.264/MPEG4 AVC [11]. The motion estimation procedure yields two 2D matrices, which contain the G_x and G_y components of the motion vectors at grid points, respectively. The first M DCT coefficients along the zigzag scan order, both for x and y directions, are combined to form a feature vector f of dimension 2M as depicted in Fig. 2. This feature vector representing the dense grid motion will be denoted by f_{GRD} in rest of the paper.

For lip contour extraction, we employ the lip geometric key points and fit polynomials on the outer lip contour based on a technique proposed in [12,13]. The technique is based on six designated key points detected on the lip contour. The algorithm fits additional points on the outer lip key points by guiding a "jumping snake" onto the upperlip boundary [12]. The additional detected key points serve as the junction points of four cubic polynomials and two line segments to be fitted onto the lip contour via least squares optimization.

The DCT coefficients computed separately for the x and y directions are concatenated to form the feature vector that is denoted by f_{CTR}.

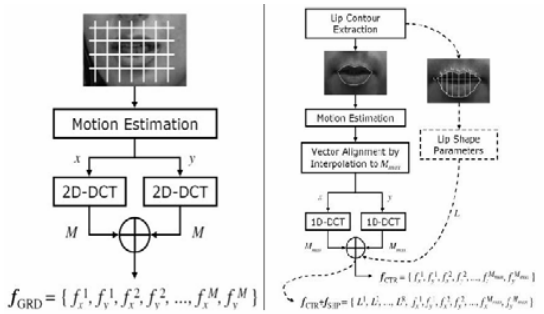

Fig. 2. Grid and Contour Based Lip motion feature extraction

The three types of visual features were concatenated to form a 24 dimensional feature vector. This is shown in Eqn. .(3) where f_{DCT} , f_{GRD} , f_{CTR} represent the DCT, grid, and contour based lip motion features respectively and o_t refers to the observation feature vector for the frame at time t.

$$o_t = \left[o_t^{DCT}, o_t^{GRD}, o_t^{CTR}\right] \quad (3)$$

Similarly to the audio case, we have T_V visual observations ($T_A \approx 2T_V$) and a sequence, O_M, of visual mouth speech feature vectors or observations denoted by:

$$o_M = \{o_1, o_2, \cdots, o_t, \cdots, o_{T_v}\} \quad (4)$$

C. 3D Face Module

The 3D facial feature module is described in detail in [14,15,16]. The dimensionality of the features varies depending on the type of the facial data representation and feature extraction techniques. For 3D face module, each face is modelled with 3D shape and texture features, the TEX-GABOR for texture features [14,15], and CURV-PD for the shape features [15,16].

IV. MULITLEVEL FUSION STRATEGY

For each transaction, the audio-video sentence observation from a speaker is decomposed into its three constituent parts, giving a sequence of audio feature vectors O_A, a sequence of visual speech (mouth) feature vectors, O_M, and a sequence of 3D facial feature vectors O_{3F}. These three observations are processed by the three classifier modules to give three individual sets of likelihoods, $ll(O_A|\lambda)$, $ll(O_M|\lambda)$, and $ll(O_{3F}|\lambda)$. The objective is to discern from these sets of scores, the reliability of each module and hence determine appropriate module weights.

However, we used the following design criteria were taken into account, when designing the proposed multilevel fusion strategy. The multilevel fusion method should easily allow the

addition of other modules. The system must be robust to mild through adverse test levels of both audio and visual speech (mouth) noise. The contribution from each source of information to the final decision must be weighted dynamically by taking the current reliability of each source of information into account. The module score weightings must be determined in an automatic unsupervised manner. The best performing fusion mode and the score weightings from SIV experiments are then to be used for performing LV experiments.

Given these criteria, we decided to use late fusion at the score level, based on the theoretical and empirical evidence from findings in the previous related work and the related literature [1,2,3]. With regards to the type of fusion rule, the sum rule is known to be superior to the product rule [2,3], particularly when the module scores have large errors. Thus the weighted sum rule should be resilient to noise and is a good choice for score level fusion, particularly in this application where either or both of the audio and video (mouth) modalities may be highly degraded.

By taking all of this into consideration, the proposed multilevel fusion strategy is based on weighted sum score fusion with *min-max* normalization. The fusion is implemented as follows. We first perform a fusion of two modules (e.g. audio with 3D face, audio with mouth, or 3D face with mouth). Then this bi-module fusion is extended to include an additional third module, thus yielding tri-module fusion at two different levels which can be applied to the audio, mouth, and 3D face modules.

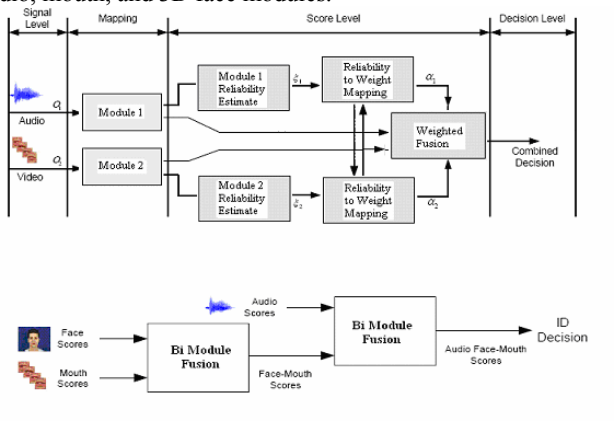

Fig. 3. Multilevel Fusion with cascaded Biomodal Fusion modules

We use $ll(O_m|\lambda)$ to denote the confidence score output from the m^{th} module representing the log-likelihood that the observation O_m was caused by the client model/template λ where $m \in \{A, M, 3F\}$, with A, M, and $3F$ representing the audio (O_A), mouth (O_M), and face (O_{3F}) module observations respectively. Figure 3 illustrates the proposed multilevel fusion strategy with cascaded modules.

This multilevel fusion strategy consisting of combining two bi-module fusion modules in cascade can take account of a noisy audio or video signal and also of any one of the three modules performing poorly, thus weighing the contribution of each module to the final decision appropriately. The advantage of this fusion method is that, being adaptive, the training of the fusion parameters is not required. Importantly, no assumption has been made about the type or level of audio or video noise that may cause a module to perform poorly. This is important for a practical audio-video system because learned noise statistics that are used to map the reliability estimate to the weighting parameter have been previously shown to vary with the type of degradation causing the train/test mismatch [4,5]. This compromises the mapping, as it must perform well for all types of noise (audio or video) and not just for one specific type of noise. Furthermore, the training of fusion parameters requires additional audio-visual data, which poses problems for the testing of existing audio-visual databases and also for practical applications, due to the small amounts of available audio-visual data. The proposed method requires no training data, and the weights are determined solely on the outputs scores from each module. We will now describe the three dimensional audio-video data corpus used, and the fusion experiments that were carried out using the proposed method.

V. 3D AUDIO VISUAL DATA CORPUS

The AVOZES 3D stereovision database [17] was used for all the experiments described in this paper. AVOZES contains video recordings from 20 native speakers (10 male and 10 female) of Australian English. Video recordings were made using a calibrated stereo camera system. Video frames are stored as DV-AVI files in the NTSC format (29.97Hz frame rate, 720x480 pixels resolution). Audio recordings were made using a mono microphone. Audio data are stored both in the DV-AVI files as well as in separate WAV files as 48 kHz 16 bit linear encoded samples. Module 6 of the corpus was used for training, and sentences from Module 4 were used for testing. Module 6 contains application-driven sequences with examples of continuous speech from each speaker. The three sequences are:

1. "Joe took father's green shoe bench out."
2. "Yesterday morning on my tour, I heard wolves here."
3. "Thin hair of azure colour is pointless."

Together with the first sentence, the second and third sentences were designed in such a way that they contain almost all phonemes and visemes of AuE (/æ/ is the only phoneme missing). Module 4 contains several short sentences in CVC/VCV words enclosed by the carrier phrase "You grab /WORD/ beer.".

To test the robustness of the proposed system, both the audio and video test signals were degraded to provide a train/test mismatch. Ten levels of audio and video degradation were

applied. This mild to adverse train/test mismatch noise levels emulates the operating scenarios encountered in a realistic operating environment. The audio models were trained on the "clean" audio speech, which was the original AVOZES audio data. Additive white Gaussian noise was applied to the clean audio at SNR levels ranging from 48 dB to 21 dB in decrements of 3 dB. In order to account for practical video conditions encountered in real operating scenarios, the video frame images were compressed using JPEG compression. Ten levels of JPEG QF were tested, with $QF = \{50,28,18,14,10,8,6,4,3,2\}$, where a QF of 100 represents the original uncompressed image. The variation of the mouth ROI images w.r.t. JPEG QF is shown in Figure 4. JPEG blocking artifacts are evident at the lower QF levels.

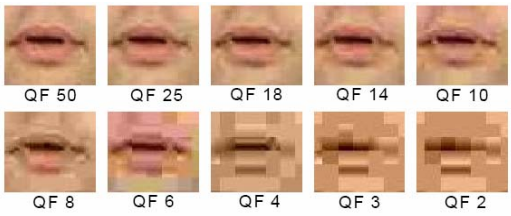

Fig. 4. Ten levels of JPEG compression on mouth ROI images

VI. EXPERIMENTAL RESULTS

For evaluating the performance of the proposed multilevel fusion approach in this paper, we have used the AVOZES 3D stereovision face database [17]. The AVOZES database contains 10 male and 10 female stereo video data speaking phonetically balanced sentences. The AVOZES database consists mostly of frontal faces and does not exhibit significant expression variations. However, some scans have slight in-depth pose variations and different expressions. Although the quality of the data is high, we used median filtering after three-dimensional reconstruction, first to remove the impulse noise, and then mean filtering was applied to smoothe the facial surface. Module 6 of the database was used for training and Module 4 was used for testing. The neutral face image (1st frame of the sequence) was used for building the face template.

First we report the performance of audio only and visual speech only module results, followed by the performance of the fusion of the two modules and then of all three modules.

A. Performance of Audio only Module

For examining audio only performance, we built ten-mixture GMM speaker models trained with 34-dimensional audio MFCC features. Gender specific UBM were used as described. The three sentences from Module 6 in AVOZES were used for training and the 2 sentences from Module 4 were used for testing. The gender specific UBMs were trained using all three sentences from all the speakers from the separate male and female cohorts. All models were trained using the clean speech and tested using the various SNR levels. Figure 5 shows how the audio-only module performs w.r.t. the audio degradation. The numerical EERs are given in Table 9.4. The best EER of 2.4% was achieved at 48dB. At 21dB the EER dropped to worst possible EER of 50%.

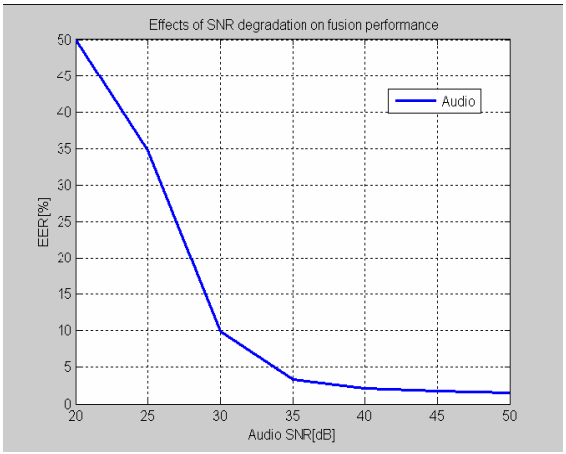

Fig. 5. Effects of audio degradation on fusion performance

B. Performance of Visual Speech only Module

In this set of single mode experiments, the effect of the GMM mixtures on the performance of the four visual speech feature types (f_{DCT}, f_{GRD}, f_{CTR}), and concatenated (f_{DCT}-f_{GRD}-f_{CTR}) was tested initially. These tests were carried out using matched training and testing data sets, i.e., the original "clean" images. To examine whether the dynamic lip motion features, such as the f_{GRD} and f_{CTR}, features, would perform better with a larger number of GMM mixtures, we increased the number of mixtures from one until a performance trend became apparent. For each lip feature type, a trend in the EERs with respect to the number of mixtures can be seen. The number of mixtures that maximised the visual speech features performance for each of the four feature types, are given in Table I.

TABLE I: NUMBER OF GAUSSIAN MIXTURES THAT MAXIMISES THE EER PERFORMANCE FOR EACH OF THE FOUR TYPES OF VISUAL SPEECH FEATURES ACROSS TEN LEVELS OF JPEG Q

Features	GMM mixtures	Clean	QF 50	QF 25	QF 18	QF 14	QF 10	QF 8	QF 6	QF 4	QF 3	QF 2
f_{DCT}	2	13.5	14.3	15.1	15.9	15.9	17.5	19.9	20.7	39.8	49.4	50.0
f_{GRD}	15	23.5	25.9	31.9	35.5	47.4	50.0	50.0	50.0	50.0	50.0	50.0
f_{CTR}	18	20.7	22.7	29.9	32.3	43.8	50.0	50.0	50.0	50.0	50.0	50.0
f_{DCT}-f_{GRD}-f_{CTR}	4	8	9.6	10.8	12.0	12.4	16.7	24.3	32.7	50.0	50.0	50.0

The f_{DCT}, features performed best even with just two mixtures and decreased steadily with increasing number of states. The number of states, that maximised the EERs for the f_{GRD} and f_{CTR} features, were 15 and 18 respectively. The

concatenated f_{DCT} -f_{GRD}- f_{CTR} feature vector was modelled best using four mixtures.

For the video degradation experiments the mouth module GMMs were trained on the "clean" (uncompressed) video images and tested on the degraded video images. This provided for a mismatch between the testing and training video conditions. The tests on the degraded mismatched video data were carried out using different number of mixtures, which maximised the performance for each of the four visual feature types (as above).

This is different to all the experiments conducted for audio only mode, where ten Gaussian mixtures were always used for both training and testing. Table I and Figure 6 show how different visual speech features perform w.r.t. JPEG degradation.

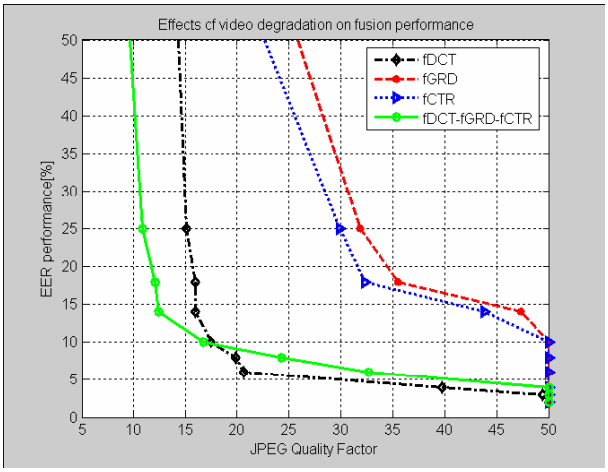

Fig. 6. Effects of video quality degradation on fusion performance

C. *Performance of Fusion of 2D Mouth-3D Face Features*

The face gallery (training) set, comprising three images, was formed by arbitrarily extracting the first image frame from each of the first three training sentences from AVOZES module 6. These were used to form a face template for each of the *N* subjects.

TABLE II: THE MOUTH, FACE AND FACE-MOUTH EERS FOR TEN LEVELS OF JPEG QF

JPEG QF	50	25	18	14	10	8	6	4	3	2
Mouth [%]	14.1	14.9	15.7	15.7	17.3	19.8	20.6	39.5	50.0	50.0
Face [%]	1.2	1.2	0.4	0.4	1.2	1.2	2.0	8.1	14.1	25.0
Mouth-Face[%]	0.0	0.8	0.0	0.0	0.0	0.0	0.0	1.6	7.3	12.5

In all the face experiments, the probe images used for testing was obtained from the module 4 sentences (again, the first frame). The gallery sets consisted of the original uncompressed images and the probe sets consisted of degraded images at the ten levels of JPEG compression. This provided for a gallery/probe mismatch.

The fusion of TEX-GABOR for texture module and CURV-PD for the shape module was used in score-level fusion, and the performance of the 2D Mouth - 3D face fusion module w.r.t. JPEG QF is given in Table II and Figure 7.

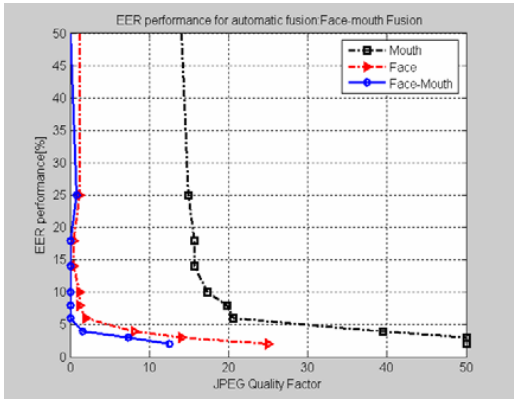

Fig. 7. Fusion Performance for .3D Face- 2D mouth Features

D. *Performance of TriModule Fusion*

For this set of experiments, we examined the performance of all the three modules, involving the fusion of audio, 2D mouth features and 3D face features in a cascaded fusion .strategy shown in Figure 3. The results for this set are shown in Table III and Figure 8.

TABLE III: THE AUDIO – FACE – MOUTH PERFORMANCE FOR DIFFERENT LEVELS OF JPEG QF AND AUDIO SNRS IN TERMS OF EERS

Audio Face-mouth	dB	48	45	42	39	36	33	30	27	24	21
QF	V \ A	2.4	2.4	2.8	2.8	4.4	6.9	10.9	24.2	42.7	50
50		0	0	0	0	0	0	0	0	0	0
25		0.8	0	0	0	0	0	0	0.4	0.4	0.8
18		0	0	0	0	0	0	0	0	0	0
14		0	0	0	0	0	0	0	0	0	0
10		0	0	0	0	0	0	0	0	0	0
8		0	0	0	0	0	0	0	0	0	0
6		0	0	0	0	0	0	0	0	0	0
4		1.6	0	0	0.4	0	0	0.4	0.8	1.2	1.6	1.6
3		7.3	0	0	0	0	0	0.8	1.2	1.6	2.4	3.6
2		12.5	0.8	0.8	0.8	1.2	1.2	2.4	2.8	3.6	4.4	7.3

We define an operating point as the fusion of the audio module with the video module/modules at a particular audio SNR and video JPEG QF levels. For a clearer comparison, ten operating points are shown in Figure 9.

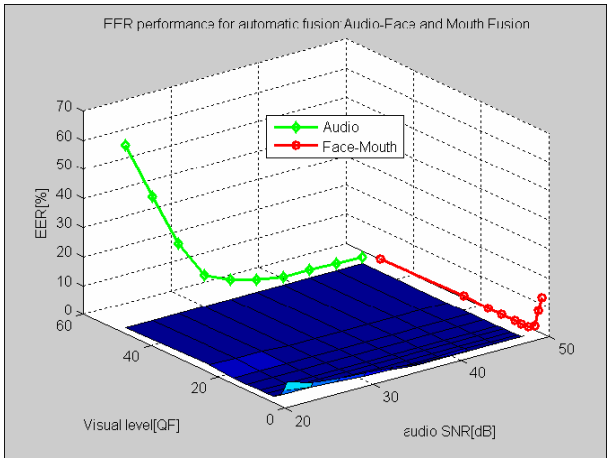

Fig. 8. Tri-module Fusion Performance

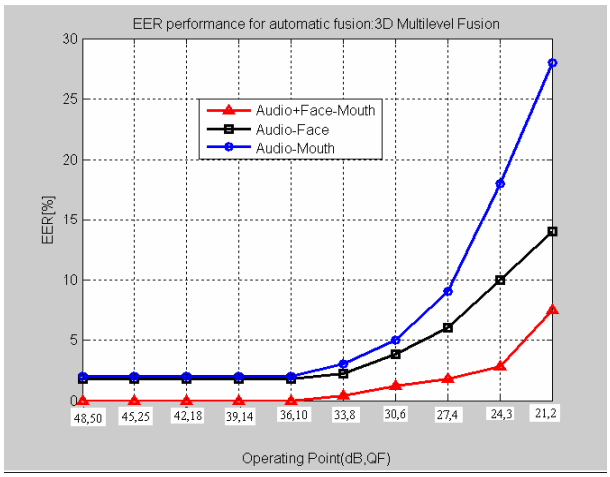

Fig. 9. Ten operating points (dB, QF), comparing multilevel fusion modes

VII. Discussion

The audio module performed very well under near "clean" testing conditions, however the performance roll off w.r.t. SNR is very high, which can be seen in Figure 5. This highlights the vulnerability of a unimodal acoustic based modeling approach to mismatched train and test conditions. For the mouth module experiments, the fact that the f_{DCT} visual features performed best with just two Gaussian mixtures indicates that GMMs can really model the static lip shapes better, and there is not need for exploring other complex type of Gaussian models such as HMMs and embedded HMMs. Other person recognition studies based on the mouth ROI have ignored the temporal mouth information and modelled the statistical distribution of the mouth shape using just static DCT type features using GMMs [79,96]. Here we have used explicit DCT based lip motion features f_{GRD} and f_{CTR} in addition to static f_{DCT} features. The best mouth module performance of 8% is surprisingly high, considering that only mouth information was employed. While the f_{DCT} - f_{GRD} –f_{CTR} concatenated features outperform the static f_{DCT} features for high QFs (f_{DCT} - f_{GRD} -f_{CTR} 9.6% versus f_{DCT} 14.3% at a QF of 50), the performance at a QF of 2 is 50%, which is similar to f_{DCT} performance. The lip motion features perform very poorly for low QF levels, both falling to around 50% at a QF of 2. The results also show that the static f_{DCT} features are more important, and are more robust than the dynamic lip motion features for the identity verification scenario. Also, non-temporal GMM modelling used here may be more suitable than temporal HMM modelling. This validates our observations for viewing the mouth features as a supplement to the existing facial features employed by the face recognition systems.

It was expected that the 3D face module, employing features located throughout the entire face would outperform the visual speech module, employing features extracted from just the mouth ROI. The 3D face module outperformed the mouth module at all levels of train/test mismatch. The highest face module performance was 1.2% EER, which is 15% better (relative) than the highest mouth EER accuracy. The face module also exhibits higher robustness to JPEG compression, when compared to the mouth module, with EERs less than 2%, for all test mismatch levels exceeding a QF of 4. At the highest mismatch QF level of 2, the 3D face module EER was 25%, and the mouth module EER was 50%. The superior performance of 3D face module is more impressive when considering that the 3D face training set consists of only three images, whereas the mouth model has the advantage of "seeing" three sequences of video frames (100 visual frames in one sequence) and hence more variation in the subjects' appearance. Nonetheless, it is still interesting to examine if the combination of the 3D and mouth modules would yield any improvement in performance and robustness.

For the fusion of the 3D face and mouth modules, a perfect face-mouth EER of 0% is achieved at several levels of JPEG QF mismatch. Also, the face-mouth EERs are lower than either of the face or mouth expert EERs for all levels of JPEG QF mismatch, i.e. we have synergistic fusion. The most significant improvements are obtained for the higher levels of mismatch, for example at the lowest QF level of 2, the face-mouth, face and mouth EERs are 12.5%, 25%, and 50% respectively. The performance of the face and mouth modules both roll off suddenly at a QF of 4. This is also the case for the face-mouth EERs, which are approximately 0% until a QF of 4 and then rise up, albeit with a lower roll off compared to either the face or mouth modules. The improved face-mouth performance indicates that the mouth features complement the acial features that the 3D face module uses. The improvement may be due to two factors. Firstly the 3D face module emphasises eye information and hence the mouth module is complementary, and secondly the mouth module can capture the variation of the mouth ROI better over the three training video frame sequences.

The audio-mouth EER performance also represents an improvement over the individual audio and mouth module performance at all tested levels of audio and video train/test

mismatch. At the (21dB, 2QF) operating point, the audio, mouth, and audio-mouth EERs are 50%, 50%, and 28.6%, respectively, representing a relative improvement of 49% over the mouth module.

Further, the audio-face results show an improvement over the individual modules. At the (21dB, 2QF) operating point, the audio, face, and audio-face accuracies EERs are 50%, 25%, and 13.7% respectively.

For the Tri-module experiments, perfect audio-face-mouth EERs of 0% were achieved at the majority of operating points. The tri-module fusion attains a significant increase in robustness to both audio and video degradations. This is evident from the flatness of the audio-video surface in Figure 8.

As can be seen in Figure 9, the improvements in robustness were most significant at the highest levels of train/test mismatch. At 21dB, the audio EER is 50% and at a JPEG QF of 2, the face and mouth EERs are 25% and 50% respectively. At the (21dB, 2QF) operating point, the audio-mouth, audio-face and audio-face-mouth EERs are 28.6%, 13.7%, and 7.3% respectively. Improvements over the face-mouth EERs were also achieved, particularly at the (21dB, 2QF) operating point, where an EER of 7.3% outperforms the face-mouth EER of 12.5% at a QF of 2. This exemplifies the increased robustness of the tri-module fusion over bi-module fusion with audio and video degradation. Importantly, fusion in a highly mismatched scenario (e.g. audio 50% at 21dB) with a "clean" test (e.g. face 25%, mouth 50% at QF2) does not result in catastrophic fusion (audio-face-mouth 7.3%). These results were achieved with the Tri-module fusion block having no prior knowledge of the level or type of audio or video degradation. Hence, we have a generalised fusion methodology, which will not be adversely affected by varying types of audio and video degradations.

VIII. CONCLUSIONS

In this paper, a multilevel fusion approach to 3D face modelling was proposed for biometric speaker identity verification applications. The approach combines information from three modules, namely audio, visual speech, and 3D face information in an automatic unsupervised fusion, adapting to the local performance of each module, and taking into account the output-score based reliability estimates of each of the modules. These results as a whole are important for remote authentication applications, where bandwidth is limited and uncontrolled acoustic noise is probable, such as video telephony and online authentication systems. Experiments on a 3D stereovision database AVOZES show that, at severe levels of audio and video mismatch, the audio, mouth, 3D face, and tri-module (audio+mouth+3D face) fusion EERs were 42.9%, 32%, 15%, and 7.3% respectively for a biometric speaker identity verification application.

IX. REFERENCES

[1] G. G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. W. Senior, "Recent Advances in the Automatic Recognition of Audiovisual Speech," Proceedings of the IEEE, vol. 91, pp.1306-1324, Sept. 2003.

[2] C. C. Chibelushi, F. Deravi, and J. S. D. Mason, "A Review of Speech-Based Bimodal Recognition," IEEE Transactions on Multimedia, vol. 4, pp. 23-35, Mar 2002.

[3] R. Brunelli and D. Falavigna, "Person Identification Using Multiple Cues," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 17, pp. 955-966, Oct. 1995.

[4] S. Pigeon and L. Vandendorpe, "Image-based multimodal face authentication," Signal Processing, vol. 69, pp. 59-79, 1998/8/31 1998.

[5] J. Ortega-Garcia, et al., "MCYT baseline corpus: A bimodal biometric database," IEE Proc.VISP, vol. 150 2003.

[6] Santi, A., Servos, P., Vatikiotis-Bateson, E., Kuratate, T. & Munhall, K. (2003). Perceiving biological motion: Dissociating talking from walking. Journal of Cognitive Neuroscience. 15, 800-809.

[7] Callan, D., Jones, J.A., Munhall, K.G., Kroos, C., Callan, A. & Vatikiotis-Bateson, E. (2003).Neural processes underlying perceptual enhancement by visual speech gestures. Neuroreport, 14 2213-2218.

[8] Hani Camille Yehia, Takaaki Kuratate, and Eric Vatikiotis-Bateson (2002). Linking facial animation, head motion, and speech acoustics, Journal of Phonetics, Vol.30, No.3, pp.555-568.

[9] Christian Kroos, Takaaki Kuratate, and Eric Vatikiotis-Bateson (2002). Video-based face motion measurement, Journal of Phonetics, Vol.30, No.3, pp.569-590.

[10] T. F. Quatieri, "Discrete Time Speech Signal Processing", Signal Processing Series: Prentice Hall, 2002.

[11] H.264/MPEG-4 AVC standard, http://www.itu.int/rec/T-RECH.264-200305-S/en.

[12] Chetty, G. and Wagner, M., "Automated lip feature extraction for liveness verification in audio-video authentication", Proc. Image and Vision Computing 2004, New Zealand, pp 17-22.

[13] Chetty, G. and Wagner, M., "Audio-Visual Speaker Identity Verification using Lip Motion Features, INTERSPEECH 2007 conference.

[14] B. G¨okberk, M.O.Irfano˘glu, and L. Akarun, "3D shape-based face representation and facial feature extraction for face recognition (in press)," Image and Vision Computing, 2006.

[15] H. Duta˘gacı, B. Sankur, and Y. Yemez, "3D face recognition by projection-based features," in Proc. SPIE Conf. on Electronic Imaging: Security, Steganography, and Watermarking of Multimedia Contents, 2006.

[16] K. W. Bowyer, K. Chang, and P. Flynn, "A survey of approaches and challenges in 3D and multimodal 3D + 2D face recognition," Computer Vision and Image Understanding, vol. 101, no. 1, pp. 1–15, 2006.

[17] R. Goecke and J.B. Millar. The Audio-Video Australian English Speech Data Corpus AVOZES. In Proceedings of the 8th International Conference on Spoken Language Processing INTERSPEECH 2004 - ICSLP, Volume III, pages 2525-2528, Jeju, Korea, 4 - 8 October 2004. University Science, 1989.