International Journal of Semantic Computing
Vol. 4, No. 2 (2010) 155–179
© World Scientific Publishing Company
DOI: 10.1142/S1793351X10000985

ROBUST MULTIMODAL PERSON RECOGNITION USING LOW-COMPLEXITY AUDIO-VISUAL FEATURE FUSION APPROACHES

DHAVAL SHAH*, KYU J. HAN† and SHRIKANTH S. NARAYANAN‡

Signal Analysis and Interpretation Laboratory (SAIL)
Ming Hsieh Department of Electrical Engineering
Viterbi School of Engineering
University of Southern California
Los Angeles, CA 90089, USA
*dhavalys@usc.edu
† kyuhan@usc.edu
‡shri@sipi.usc.edu
http://sail.usc.edu

In this paper,[a] we first show the importance of face-voice correlation for audio-visual person recognition. We propose a simple multimodal fusion technique which preserves the correlation between audio-visual features during speech and evaluate the performance of such a system against audio-only, video-only, and audio-visual systems which use audio and visual features neglecting the interdependency of a person's spoken utterance and the associated facial movements. Experiments performed on the VidTIMIT dataset show that the proposed multimodal fusion scheme has a lower error rate than all other comparison conditions and is more robust against replay attacks. The simplicity of the fusion technique allows for low-complexity designs for a simple low-cost real-time DSP implementation. We then discuss some problems associated with the previously proposed design and, as a solution to those problems, propose two novel classifier designs which provide more flexibility and a convenient way to represent multimodal data where each modality has different characteristics. We also show that these novel classifier designs offer superior performance in terms of both accuracy and robustness.

Keywords: Audio-visual biometrics; feature-level fusion; nested GMM.

1. Introduction

Biometric recognition holds tremendous promise for security applications. Biometrics can cover a wide range of modalities, including fingerprint, face, hand geometry, iris, retina, signature, voice, keystroke dynamics, gait, ear, physiological signals such as electrocardiograms (ECGs), and so on [2–6]. Each modality has its own advantages and limitations in terms of accuracy, robustness, and usability/user

[a] The work reported here is an expanded version of the paper [1] presented at ISM 2009 by the authors. (Specifically, Chapter 4 is newly added compared to the previous work.)

acceptance. For instance, using iris information offers very high accuracy and robustness but it requires cooperative subjects and expensive equipment. On the other hand, modalities such as the human voice and face (of interest in this paper) that can be accessed in an unobtrusive way and have higher user acceptance have restricted use due to accuracy and robustness issues. These performance challenges need to be addressed if real-life systems incorporating these modalities are to become more prevalent. Availability of robust solutions, however, promises practical applications such as personal computer login and location access.

One of the promising venues for improving biometric technology performance is to consider combining individual modalities [7–11] under the premise that both redundancy and complementarity in information can be advantageously utilized. Among a variety of possible combinations, the voice and face modalities have been broadly used for person verification applications in the past decade [12–19]. Beyond their obvious advantages in terms of usability/user acceptance, these two modalities are considered important cues for personal identity because human communication patterns embody unique personal characteristics. Furthermore, both the spoken and visual modalities are tied tightly to one another when people communicate.

Speaker recognition research using voice and face modalities started with the assumption that these modalities are independent of each other and resulted in simple techniques like score-level fusion which allow each modality to be processed independently. However, this assumption does not hold true as a person's face dynamically and systematically changes as he speaks and there is a strong correlation between these facial changes and the spoken utterance. Consequently, the resulting research tends to ignore the correlation between face and voice and miss out on the benefits offered by this correlation. A strong proof of the existence and benefits of this correlation comes from the field of speech recognition. The studies in [20–22] show that integrating voice and lipreading enhances speech intelligibility in humans and improves automatic speech recognition (ASR) accuracy. Speaker recognition research drew inspiration from speech recognition research, and hence most of the work has focussed on exploiting the correlation between the voice and lip region of the face. This not only results in extra computations for detecting the lip region but also tends to ignore the additional information that can be extracted from other parts of the face (like cheek movements or eye blinks).

In this paper, we present low-complexity approaches[b] that try to capture enhanced talking dynamics from the entire face (instead of just the lip area). For better overall performance, in terms of both accuracy and speed, we use simple modifications on a widely used face detection technique to detect face regions from each frame of the talking face video. We use feature-level fusion as a tool to model the correlated information between voice and face features. Although score-level fusion is widely utilized in practice due to its convenience in terms of handling multiple

[b] The proposed design has been implemented on a DSP processor (TMS320C6713) to work in real-time and it gives online performance comparable to offline evaluations.

information sources compared to feature-level fusion[c] [2, 5, 7, 8], it could possibly miss synchronized characteristics between facial changes and uttered speech, which are important for robust person recognition. To obtain potential benefits from score-level fusion, we also propose two novel approaches (hybrid feature-score level fusion and nested GMMs).

The rest of this paper is organized as follows. In Sec. 2, we propose a simple technique to exploit face-voice correlation. In Sec. 3, we present the experimental results and show that the system proposed in Sec. 2 gives better accuracy in normal use situations compared to audio-only, video-only, and audio-visual systems that use audio and video in no synchronism (i.e. in no correlation) and is very robust to replay attacks.[d] We also show that the proposed technique inherently uses audio and static video features for recognition and dynamic video features for liveness detection (a sub-application domain in person recognition) without adding any extra complexity. In Sec. 4, we investigate possible causes for performance degradation of a feature-level fusion system as compared to a score-level fusion system. As a solution, we propose two novel techniques to improve recognition performance additionally benefitting from score-level fusion. For this, we first introduce a hybrid feature-score level fusion approach that combines the advantages of both feature- and score-level fusion approaches. Then we compare this with a nested GMM approach (also newly proposed in this paper) that uses a two-level nested GMM for classification purposes. We also modify the EM algorithm (originally derived in [23])for the nested GMMs in this section. We conclude the paper in Sec. 5 with a summary of the proposed approaches to low-complexity audio-visual person recognition, aimed at exploiting face-voice correlation, and comments on future research directions in devising reliable multimodal biometrics.

2. Proposed System Description

We first describe the VidTIMIT database used in this research. We then present an overview of feature extraction stages for voice and face recognition and justify the choice of features used. Then we present the proposed multimodal fusion technique. We present several possible ways of fusing the modalities, noting the advantages and disadvantages of each, and then describe the proposed fusion technique and its advantages over the other techniques. Finally, we review some of the possible choices for classifiers and justify the selection of GMMs.

[c] In general, it is believed that feature level-fusion is a better implementation than score-level fusion because feature representation conveys the richest information while scores from multiple classifiers have the least information about decision making [5]. But the ideal feature-level fusion approach would require proper understanding of relationship between information sources being handled, which lacks currently in most application scenarios and score-level fusion systems provide us better performance than feature-level fusion implementations.

[d] Replay attacks refer to impostor attacks where the impostor records client data (audio or video or both) and uses the recorded information to breach security.

2.1. *VidTIMIT database*

The VidTIMIT database [18] is an audio-visual database comprised of audio-visual recordings of 43 people reciting sentences from the test section of the TIMIT corpus [24]. This database has been utilized popularly for audio-visual person recognition research and referred to in the literature, including [15, 16, 25–28]. It was recorded in 3 sessions with a mean delay of 7 days between sessions 1 and 2 and 6 days between sessions 2 and 3. Due to the delay between sessions, the possibility of mood and appearance changes is expected that introduces some real-life aspects in the dataset. There are 10 sentences per person, 6 of them belonging to session 1 and two each to sessions 2 and 3. Two sentences are common to all speakers while the other eight sentences are generally different for each speaker, facilitating text-independent speaker recognition research. The availability of just 10 sentences per person underscores the issue of training data sparsity (although reflective of what is typically feasible in creating practical systems). The recordings were done in an office environment using a broadcast-quality digital video camera. The audio has some background noise (mostly AC and computer fan noise). Thus we expect that any audio-only recognition system would suffer from some performance degradation on this data. The video is relatively clean. Though it is captured using a broadcast-quality camera and compressed (lossy compression) into JPEG images with a quality factor of 90%, the background is fairly plain and constant with only the frontal face of each speaker in the picture. This relieves us of complicated tasks such as face detection from a clustered image or view-angle normalization. This situation is indeed realistic under certain application scenarios, such as personal security systems, where we expect a co-operative user and a fairly controlled data acquisition set up. Nevertheless, the zoom factor of the camera is randomly perturbed while collecting the video, and the face in the video is not at constant positions. Thus some pre-processing is still needed to extract the face from the image and compensate for different zoom factors, but this task is relatively simpler. The audio and video capture rates are also different and some processing needs to be done to compensate for this.

2.2. *Feature extraction*

Feature extraction is the first and the most important stage of any classification system. The quality of the extracted features greatly affects the performance of the complete system. Audio and visual data, though correlated, are in completely different forms and are sensed differently by humans. Thus the features used for both are also different. The fields of voice and face recognition are highly developed and many different ways of capturing features are available in the literature. Our approaches to feature extraction are based on such developed strategies and described in the subsequent sub-sections.

2.2.1. *Voice feature extraction*

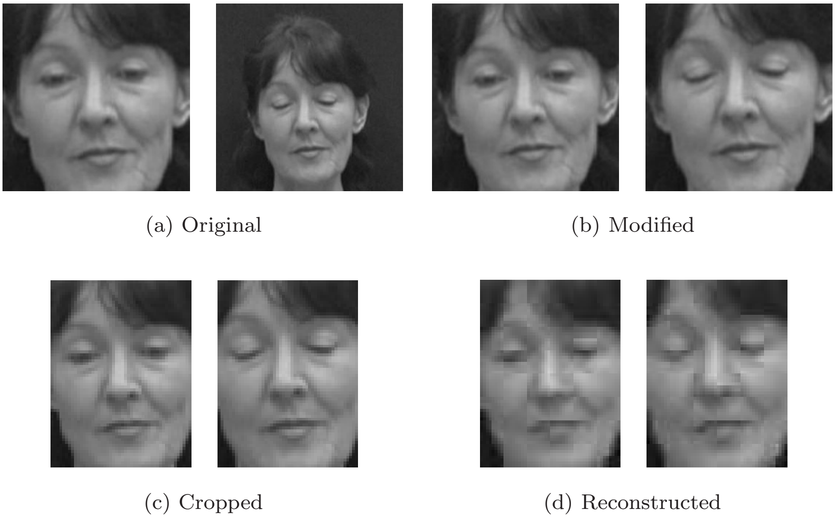
As a pre-processing step on the audio data, we perform pre-emphasis to compensate for the high frequency fall-off in the data. We then use the well-known short-term analysis technique using a 50 ms window with 50% overlap between adjacent windows. We apply the Hamming window to each segment to minimize spectral leakage. The above-mentioned steps are the most widely used and form a part of most (if not all) speech and speaker feature extraction systems. We select mel-frequency cepstral coefficient (MFCC) features for our research due to their demonstrated superior performance [29]. We use the first 36 MFCC features (35 + energy) as the 36-dimensional audio feature vector for each frame.

2.2.2. *Face feature extraction*

Each frame of video in the VidTIMIT database has just the person of interest in it with a frontal view of the face. We first detect the face and discard the background information (box the face) using the Viola-Jones face detection algorithm described in [30]. For speed up and improved accuracy in this face detection stage, we add two modifications to the original algorithm. Specifically we use a minimum size of window for the Haar detector (so that it does not expend time looking for faces which are smaller than a specific size). For the first frame in a video, we set this parameter like a threshold (which is small so it does not give much speedup), but for subsequent frames, we use the past frame's detected face size multiplied by some constant (usually around 0.8–0.9 which specifies the maximum expected decrease in face size as compared to the previous frame) as the minimum window size. This gives a considerable amount of speedup while keeping the face detection performance the same. When multiple faces are detected, if it is the first frame, we use the largest face; otherwise, we use the face which has the size closest to the previous frame's face size. The implicit assumption is that the speaker does not move his face back and forth with very high speed (if it is moved with a low speed, the 0.8–0.9 factor tackles it well). In Figs. 1(a) and 1(b), we show two sets of face detector outputs for consecutive image frames, one generated by the original face detection algorithm and the other by the modified approach, for better understanding of the benefits of this modification. We see that in addition to the speed up in face detection, we get boxed face images with a consistent amount of background. This reduces mismatch in the face modality of the data and consequently leads to an improved classification rate. We then resize this image to a standardized boxed image size of 32 × 32 pixels.[e] We then consider cropping face images since this face detector generally gives a square image while tightly boxed face images, in general, have greater height than width. So we crop the left and right parts of the downsampled image to give the standard face image. In our case, we use a 32 × 32 boxed image and then crop it to 24 × 32, which is shown in Fig. 1(c).

Many different kinds of features can be used for face recognition. The most widely used ones include eigenfaces, discrete cosine transform (DCT), and Gabor wavelets. Eigenfaces are well suited for face recognition. In this technique, the

[e] For this normalization we use OpenCV, which also provides various convenient image processing tools such as bilinear interpolation.



(a) Original

(b) Modified

(c) Cropped

(d) Reconstructed

Fig. 1. Two consecutive frame images after face detection. (a) By the original face detection algorithm (32 × 32 pixels). (b) By the modified version considering the face size of the previous image frame (32 × 32 pixels). (c) By the more refined version considering the cropped faces with a general ratio of width and height (24 × 32 pixels). (d) By the reconstructed procedures from the face features (24 × 32 pixels). Note that we have zoomed all the images by 2× for better display.

features independent of the person's facial expression (principal components) are preserved while dynamically changing features are discarded. Also, the technique needs a group of images to extract features. Our application requires that we extract static as well as dynamically changing features of a person's face per frame instead of averaging out the information contained in neighboring frames. Gabor wavelets and DCT better suit our requirement in this regard as they can be used to extract information based on a single image, and static as well as dynamic features can be captured and preserved. Gabor wavelets are, however, computationally expensive which challenges their use in real-life applications. DCT gives a performance comparable to Gabor wavelets but is simpler to implement and computationally less expensive (desirable for real-time implementation). For these reasons, we use DCT to extract visual features in this research. To extract features, we segment the cropped face images into blocks of size 4×4 pixels and calculate the DCT coefficients of each block separately. From video compression theory, we know that lower-order DCT coefficients contain most of the structural information of given data, and even after throwing away higher-order coefficients, a reasonable re-construction of the original image can be achieved. Thus we use information of the first AC coefficient in either direction as well as the DC coefficient and end up with a 144-dimensional feature vector for each face image (3 features per block for 48 blocks).

The choice of the standardized face image size, block size, and number of features per block was made empirically. A small image size is desired to minimize redundant information and reduce calculations to facilitate real-time implementation. A smaller block size reduces computations for DCT calculation and ensures that the short-term stationarity assumption is satisfied. On the other hand, a larger block size is desired to reduce the number of blocks and hence the number of feature vectors per image. Also, the block size dictates the sampling resolution in the frequency domain. Oversampling leads to larger number of redundant (or even potentially detrimental) features while undersampling may lead to loss of useful information. Considering these, a block size of 4×4[f] seemed reasonable for our experiments. The choice of just 3 features per block can be seen to be reasonable in Fig. 1(d), which shows the two consecutive frame images re-constructed from the facial features we consider as visual features in this paper. As shown in the figure, a reconstructed 24×32 image still contains enough information for a human to recognize a person and thus is deemed to contain enough person-dependent information in it. It should be noted that this may not be an ideal choice of parameters for optimal face recognition (which is, of course, data dependent). Our aim here is not to build an ideal face recognizer; rather it is to show the importance of voice-face correlation for person recognition and thus we work with these parameters as they are primarily designed to reduce computational complexity and memory requirement while giving a reasonably good performance (as will be seen in Sec. 3).

2.3. *Multimodal fusion*

Multimodal fusion is at the heart of any system which uses more than one modality. The choice of a fusion strategy is highly dependent on the modalities being used. In this section, we review some of the possible audio-visual fusion strategies, discuss their advantages and disadvantages, and justify our choice of the feature-level fusion strategy in terms of audio-visual feature correlation.

Fusion techniques can be broadly divided into 3 categories: early integration, intermediate integration and late integration [10, 16]. Late integration techniques use different classifiers for both modalities and combine their decisions. This combination can be decision level fusion (AND, OR, etc.) or opinion (score-level) fusion (weighted summation, weighted product, etc.). The inherent assumption in using such techniques is that the modalities used are independent of each other. This is not the case when audio-visual modalities of speech communication are used. A person's face deforms differently depending on what is being spoken and the underlying speaking style variations. Intermediate integration techniques use multi-stream HMMs. The inherent drawback in this technique is that it again assumes independence between the modalities used. This assumption enables it to handle audio and video streams asynchronously but some useful information correlating the two modalities is lost.

[f] Besides, 4 is a power of 2 and can possibly give some computational or memory advantages.

Early integration offers a natural way of integration for our problem. Feature level fusion is a type of early integration technique. Here, we process the different modalities separately and extract appropriate features and merge them by either concatenating or weighted summation, etc. This enables the use of a single classifier which simplifies system design. It also takes into account correlation between the two modalities inherently. A drawback of this technique is that it needs data in time synchronism. In our application, we desire the data to be in time synchronism irrespective of fusion techniques and thus this drawback is not pertinent. We calculate features for the individual modalities separately and just concatenate them. This effectively ties a spoken utterance and the corresponding face appearance. This correlation is preserved by the classification stage. We will show that this correlation acts as a hidden liveness detector to differentiate between true claims and replay attacks and increases robustness. It should be noted that audio and video are captured at different rates. This poses a problem to synchronism and needs to be addressed. This can be done in two ways. We can either upsample video data or use a hybrid scheme in which we use only audio data when video data is not available and use both when video data is available. The first scheme just adds redundant data, which may not be of use for the classification task while it helps smooth out discontinuities between adjacent frames. It also adds extra amount of processing. On the other hand, the hybrid technique is more suitable for the recognition tasks which have to be done in real-time and all possible redundancies need to be removed. In our work, the first technique has been used for offline training as well as testing while the hybrid technique has been used for online DSP implementation.

Audio and video modalities have complementary as well as redundant information. The complementary information in these modalities (for example, static features of a person's face) is usually independent and provides extra information which helps to increase the accuracy of the system. The complementary information also helps to increase the robustness of the system to some extent (only against simple replay attacks like RP1 described in Sec. 3). The redundant information in these modalities (for example, dynamically-changing utterance-dependent features of the face like lips) is usually correlated and does not provide any extra information for recognition. Thus this information cannot be used to increase the accuracy of the system. However, this redundancy can be advantageously utilized to give a high degree of robustness against many different kinds of replay attacks (as will be seen in Sec. 3). We show in Sec. 3 that the proposed fusion technique preserves both the complementary and redundant information and uses them effectively to provide increased accuracy and robustness.

2.4. *Classification*

Many different classifiers have been used for audio and visual recognition over the years, including dynamic time warping (DTW), Gaussian mixture model (GMM), hidden Markov model (HMM), support vector machine (SVM), and neural network

(NN). HMMs are widely used for speech recognition and they give high accuracy, flexibility, and robustness. They can be used for speaker recognition with the same efficacy. Since our task is text-independent, we do not need to capture/retain phone specific information. GMMs (single state HMMs) exploit this. They give a similar performance as compared to HMMs and are computationally more efficient than HMMs. Other advantages in GMMs include low memory requirement (only means and variances need to be stored), flexibility (well suited for text-dependent as well as text-independent applications), high accuracy, and robustness. Due to these reasons, we use GMMs for our classification task.

3. Experiments and Outcomes

In this section, we first describe the different experiments performed on the VidTIMIT database using the technique proposed in Sec. 2. We then show the results of the experiments followed by a discussion of the results which highlights the importance of exploiting the correlation between audio and video in terms of accuracy and robustness. Finally, we show that the proposed system is capable of operating in real-time with similar performance.

3.1. *Experimental details*

As described in Sec. 2.1, the VidTIMIT database consists of 43 speakers with audio-visual recordings of 10 sentences per speaker. We conduct verification experiments using this data. We use 8 of the 10 sentences (sessions 1 and 2) for training the model for the speaker and the remaining 2 sentences (session 3) for testing. For training the impostor model, we use the universal background model (UBM) technique [31]. The impostor model is supposed to represent as many speakers as possible (ideally other than the clients) and thus should be trained using all possible data collected from people other than the clients. However, due to lack of data, we train the UBM using all the data in the database (including all the training/testing utterances). A test utterance is deemed to be of a speaker if the probability of the speaker model given the utterance is greater than the probability of the UBM given the utterance, otherwise it is deemed to be of an impostor.

We first perform experiments to demonstrate that the proposed system is more accurate than audio-only, video-only, and the audio-visual system[g] in which audio and video are considered uncorrelated to each other. For this purpose, we use the 2 testing utterances for each speaker on the same speaker's model as a true claim and on each of the remaining 42 speaker's models as an impostor attack (or a false claim). This gives us 86 true claims and 3612 (= 2 × 43 × 42) impostor trials (or false claims).

We then move on to demonstrate the robustness of the proposed design to replay attacks. We design three types of replay attacks. The first and the simplest replay

[g] We simulate this by randomizing video frames.

attack (RP1) consists of just the audio from the speaker's test utterance combined with video from another speaker's test utterance. Care has been taken that the gender of the other speaker (whose video is used) is the same as the gender of the original speaker (whose audio is used). It represents an attack where client audio is recorded by an impostor and used to breach security. These kinds of attacks are fairly easy to detect and most audio-visual systems should be able to detect these. The second replay attack (RP2) is more difficult to detect than the first. It consists of pure audio from the client trials and a single still image from the same client trial. It represents a replay attack where along with the recorded audio of the client, his photo is used to breach security. Not all audio-visual systems would be able to detect these. Only those which employ liveness detection would be robust against such attacks. The third replay attack (RP3) is the most difficult to detect. For this replay attack, we just swap the videos of the two client trials from the same client. It represents the video of the client speaking something and the audio of the same client speaking something else. Even audio-visual systems employing liveness detection can be easily fooled by such attacks. Most systems employing liveness detection just concentrate on the lip region of the face to conclude whether a person is actually speaking something or not. They do not take into account what the person is speaking. The only way to be robust against such attacks is to capture and exploit correlation between audio and video. For all three kinds of replay attacks, we have 86 impostor trials and use the same 86 true claims as described above (for the normal situation case) to assess performance.

3.2. *Results and observations*

Figure 2 shows a performance comparison of audio-only, video-only, audio-visual system (with fusion of audio and video features in an uncorrelated fashion), and audio-visual system with the proposed feature-level fusion. In this experiment, we set the number of gaussians in each speaker model (GMM) for the video-only system to 8 and for the audio-only and the two audio-visual systems to 32, with which each system achieves its best performance.

From the figure, we see that the audio-visual system with the feature-level fusion approach in an uncorrelated fashion (i.e., in no time synchronism) has the worst (highest) equal error rate (EER) value. This demonstrates that the assumption that audio and video are uncorrelated does not hold and such assumptions can prove detrimental to performance. One thing that needs to be mentioned is that the audio data in this VidTIMIT database is noisy, which caused the audio-only system to perform poorly as well. The video-only system gives the second best EER. This is because the video data in the database is comparatively clean. The best EER value is given by the proposed system. This indicates that exploiting correlation between audio and visual data can lead to significant improvement in accuracy.

From Fig. 3, we can see that the proposed design is robust against replay attacks. The robustness against RP1 is due to the mere fact that this is a multimodal system.

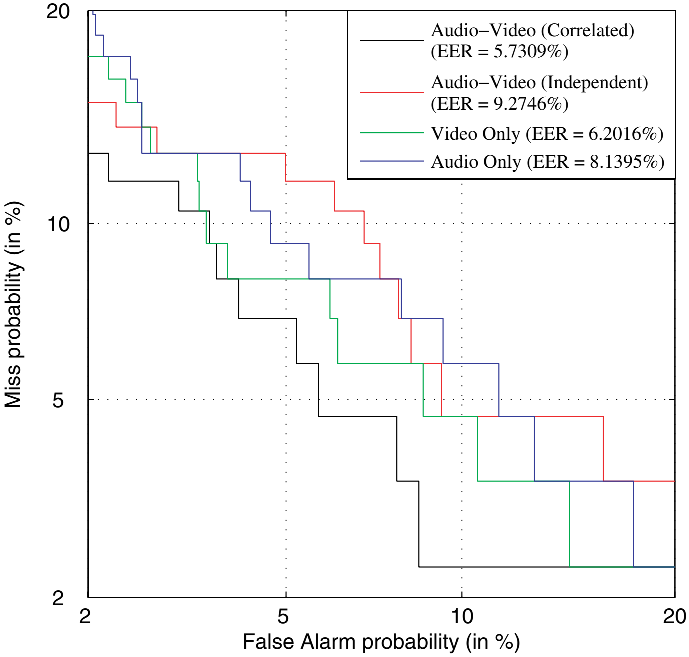


Fig. 2. Performance comparison of audio-only, video-only, audio-visual system (with fusion of audio and video features in an uncorrelated fashion), and audio-visual system with the proposed feature-level fusion.

The video data in RP1 is of an impostor and thus the video modality is responsible for this robustness.[h] RP2 has both audio and video (still image) of the client and still the system is robust. This shows that the system has an inherent liveness detector[i] (though we have not explicitly designed one). The correlation between audio and visual data, which we preserved during training, acts as a hidden liveness detector in our design, which provides robustness against RP2. RP3 has both audio and video of the client speaking different sentences. Most audio-visual systems would fail against such attacks. Even those employing liveness detection are vulnerable to such attacks as they detect liveness using lip movement information and RP3 has a live video. The only way to be robust against such attacks is to make sure that a person is speaking the same sentence in audio as well as video. One possible way would be to perform speech recognition on both audio and visual data. This technique has two problems. One is that speech recognition using visual data only is inherently limited and not reliable. Secondly this adds complexity to the system design which

[h] For RP1, audio-only systems are not secure at all while video-only systems are. To understand difference between the three replay attacks handled in this paper, please review Sec. 3.1 again.

[i] Like [28], one can try to devise a liveness detector for the purpose of making video-only systems robust to this kind of replay attack. Our proposed system provides such a detector inherently as a by-product.

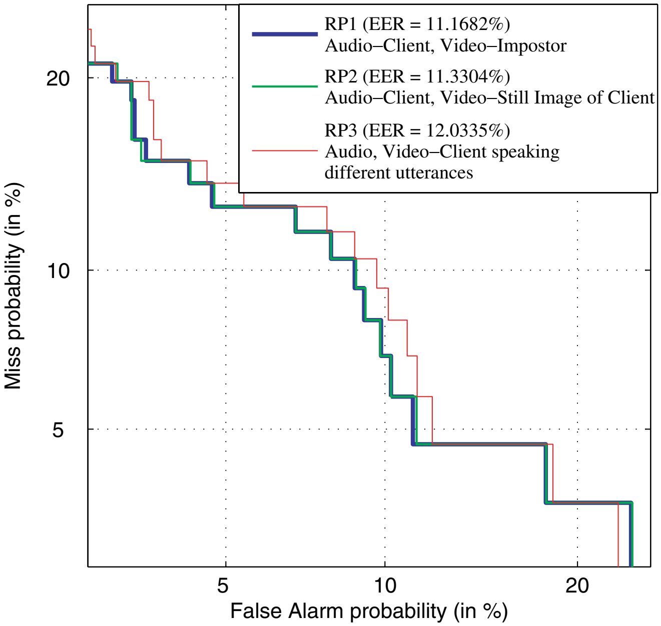


Fig. 3. Performance of the proposed system against the three replay attacks considered in this paper.

makes this technique less desirable for real-time applications. The proposed design inherently does this task without adding any complexity into the system design. It does this at the frame level. For every frame, it implicitly checks if audio and video correspond to the same sound and assigns probabilities accordingly. This figure shows that for RP3, EER as low as around 10% is possible using the proposed technique where other audio-visual systems would break down. To conclude, we see that the proposed technique is robust to a variety of replay attacks and can assure reasonable reliability for high security applications.

A simpler version of the proposed technique has been implemented on a DSP processor (TMS320C6713) using 5 gaussians for client models.[j] The system demands about 100 kb of program memory and 512 kb of data memory (excluding memory required for storing interface messages). The system is able to achieve an average latency of less than 1.5 seconds (ranging from less than a second for fast speakers to about 3 seconds for slow speakers). An additional latency of 2 seconds is introduced

[j] It should be noted that the number of gaussians used for the DSP implementation is far lesser than the number of gaussians used for offline evaluations on the VidTIMIT dataset. The major reason for this discrepancy is that we had around 20–25 seconds of training data for the offline evaluations and only 3–5 seconds of training data for the DSP implementation. Under such circumstances, 5 gaussians for the DSP implementation gave a near-optimal performance while using the same number of gaussians as for the offline evaluations would overfit the models due to lack of training data.

by the voice activity detector (2 seconds of silence is required to conclude the end of speech). The system is able to achieve an online accuracy close to 90% under semi-controlled testing conditions (distance of the person from the microphone, view angle for the person's face, etc are controlled but background noise, lighting conditions, etc. are not controlled).

4. New Approaches to Improving Classification for Feature-Level Fusion

In audio-visual processing, score-level fusion is generally known to perform better than feature-level fusion in terms of classification rate. (For example, refer to [22].) In this section, we try to identify the reason for this discrepancy in performance and propose solutions which aim at mitigating such effects and boost the performance of the feature-level fusion system additionally benefitting from score-level fusion.

In the previous section, we saw that the video-only system achieves its best performance when the number of gaussians in the GMM is set to 8 while the audio-only system achieves its best performance when this number is set to 32. These numbers do make sense as the number of visemes is less than the number of phonemes (and there is a many to one mapping from phonemes to visemes). So, now when we use traditional feature level fusion with the GMM classifier, how many gaussians do we use? We know that a small number of gaussians in a GMM means that data is not well represented and leads to degraded performance. On the other hand, a large number of gaussians may lead to overfitting which again leads to degraded performance. For optimal performance, there is an optimal number of gaussians but this optimal number is different for video-only and audio-only systems (this number is around 8 for the video-only system and around 32 for the audio-only system). Figure 4 shows that the performance of the audio-only system degrades when the number of gaussians is set to 8 (which is an optimal number for the video-only system) because the audio data is now under represented. On the other hand, Fig. 5 shows that the performance of the video-only system degrades when the number of gaussians is set to 32 (which is an optimal number for the audio-only system) because the video data is now severely overfitted. So for the system proposed in Sec. 2, if we use 8 gaussians in the GMM, voice features will cause performance degradation as they are under represented, and if we set this number to 32, the face recognizer will be severely overfitted, again causing performance degradation. We may use a number between 8 and 32 which will balance the degradation in the two modalities; however, the degradation still exists and causes an overall performance decrease. (Empirical results indicate that a better performance is obtained when the number of gaussians is set to 32 as compared to 8 or any other number in between 8 and 32.)

The inherent cause of the above problem is the lack of flexibility that the traditional GMMs can offer. We need a more flexible classifier that can account for data with different characteristics by allowing a different number of gaussians for optimal

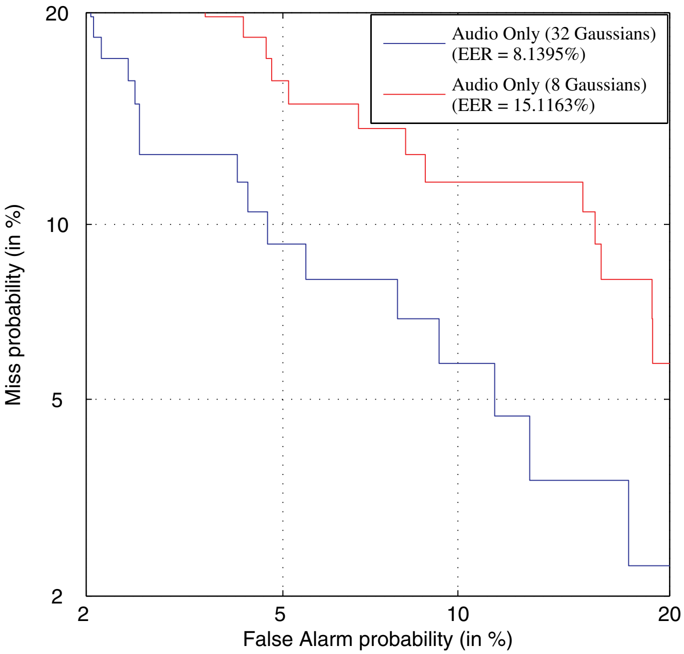


Fig. 4. Performance comparison of the audio-only systems using the different numbers of mixture components (in GMMs) which are optimized for the audio- and video-only systems, respectively.

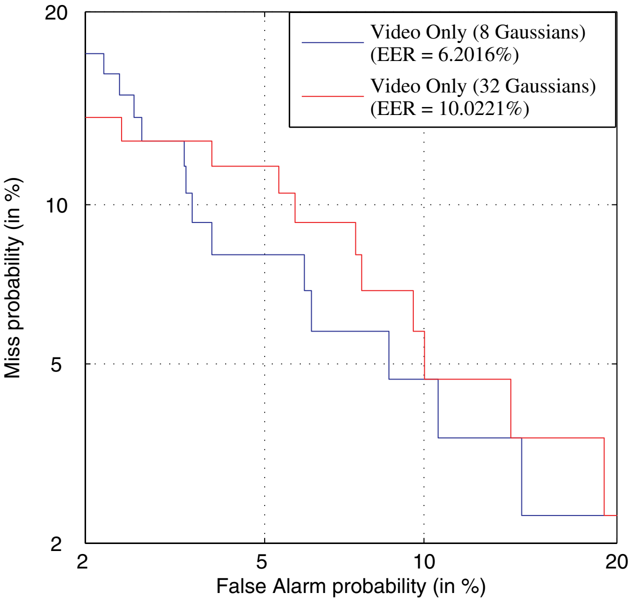


Fig. 5. Performance comparison of the video-only systems using the different numbers of mixture components (in GMMs) which are optimized for the video- and audio-only systems, respectively.

representation. To mitigate this problem, we now propose two novel and powerful, though completely different, classifier designs that give us the flexibility needed to represent different modalities which have different characteristics and which require different numbers of gaussians.

4.1. *Hybrid GMM classifier*

A brute force but powerful solution to the problem described above is a hybrid GMM classifier. This classifier uses dual-trained models with hybrid feature-score level fusion. The primary aim here is to exploit the benefits of both feature-level fusion and score-level fusion. The idea is simple. We try to do justice to both the modalities by having one model each with an optimum number of gaussians for each modality. This ensures that each modality has a model which can represent that modality in an optimal sense (so, in our case, we have two models with 8 and 32 gaussians, respectively, based on previous experimental results). To make sure that the presence of features of the other modality does not adversely affect the training of the model optimal for a particular modality, for training, we use just the features of the modality for which that model is intended. To preserve the correlation between the modalities, we also append the means and variances of the other modality. In other words, while using the EM algorithm, in the E-step, we only use the features of the modality for which that model is optimum, and in the M-step, we calculate means and variances for features of both modalities. Thus training the part of the model which represents the modality for which it is supposed to be optimal is exactly the same as training a model completely using only that modality and, as an add-on, we append trained features for the other modality. This is possible with an assumption (which we proved in the previous section) that the two modalities are strongly correlated. We hypothesize that, because of this strong correlation, the probabilities calculated using features of only one modality also hold for features of the other modality and thus training models in this way is justified.

In our work, we have two models, one with 8 gaussians and another with 32. The model with 8 gaussians is trained using the probabilities calculated on only face features (in the E-step), and the means and variances for voice features are just appended (in the M-step) using these probabilities. Similarly, the model with 32 gaussians is trained using the probabilities calculated on only voice features, and the means and variances for face features are just appended using these probabilities.

During testing, we first calculate probabilities for each of the two models. This step resembles two separate feature-level fusion systems. We then combine the probabilities for the two models at the score-level by weighted summation of the log probabilities. Thus we see that the design has two feature-level fusion systems followed by a score-level fusion system. Each feature-level fusion system is optimized for one modality, and the score-level fusion system can be used to reflect the relative reliability of the two feature-level fusion systems. This solution works and gives better accuracy, which is partly shown in Figs. 6–9, and we will discuss it in more detail in Sec. 4.3 along with the performance of the nested GMM approach that will be presented in the next sub-section.

4.2. *Nested GMM classifier*

A brute force technique like the one mentioned in the previous sub-section might not be suitable for low-cost, real-time applications due to the redundancies that are inherent in the design. In this sub-section we present a nested GMM classifier which is a natural and elegant alternative to addressing the flexibility problem we mentioned earlier. This approach does not introduce any redundancy into the design and might prove to be a better choice for low-cost, real-time applications. Because this design provides a natural way of representing the data, as will be seen in Sec. 4.3, this design outperforms the previous two designs (in Secs. 2 and 4.1, respectively).

First, consider a conventional GMM classifier. The probability of a feature vector x given a GMM whose probability density function (PDF) consists of π_k, μ_k, Σ_k, i.e., weight, mean vector, and covariance matrix for the kth Gaussian mixture component, respectively, is given by:

$$p(x) = \sum_{k=1}^{K} \pi_k \cdot p(x|\mu_k, \Sigma_k), \tag{1}$$

where K is the number of gaussians in the GMM. Assuming diagonal covariance matrices, we can split the multivariate gaussian into a product of univariate gaussians, as follows:

$$p(x) = \sum_{k=1}^{K} \pi_k \prod_{i=1}^{D} p(x_i|\mu_{ki}, \Sigma_{ki}), \tag{2}$$

where $\Sigma_{ki} = \Sigma_{k(ii)}$ and $\Sigma_{k(ij)} = 0$ for $i \neq j$. D is the dimension of the feature vector. For the feature-level concatenation where the first M dimensions are for face and the remaining D–M dimensions for voice,

$$p(x) = \sum_{k=1}^{K} \pi_k \prod_{i=1}^{M} p(x_i|\mu_{ki}, \Sigma_{ki}) \prod_{i=M+1}^{D} p(x_i|\mu_{ki}, \Sigma_{ki}). \tag{3}$$

For face, $K = 8$ gives a near optimal performance but voice is under represented. In the above expression, we replace the single multivariate gaussian for voice by a GMM with P Gaussian components (where $P = 4$), so that the total number of Gaussian components for voice now becomes $K \times P$ $(= 8 \times 4 = 32)$. Then,

$$p(x) = \sum_{k=1}^{K} \pi_k \prod_{i=1}^{M} p(x_i|\mu_{ki}, \Sigma_{ki}) \left\{ \sum_{p=1}^{P} \alpha_{kp} \prod_{i=M+1}^{D} p(x_i|\mu_{kpi}, \Sigma_{kpi}) \right\}, \tag{4}$$

where

$$\sum_{p=1}^{P} \alpha_{kp} = 1, \tag{5}$$

for $k = 1, \ldots, K$. Note that since covariance matrices are assumed diagonal we represent them as scalars for each dimension rather than as matrices, e.g., Σ_{ki} means the ith diagonal element of the kth mixture component while Σ_{kpi} means the ith diagonal element of the pth nested (2nd-level) gaussian in the kth mixture component.

For given data $X = \{x_n\}_{n=1}^{N}$,

$$p(X) = \prod_{n=1}^{N} \left[\sum_{k=1}^{K} \pi_k \prod_{i=1}^{M} p(x_{ni}|\mu_{ki}, \Sigma_{ki}) \left\{ \sum_{p=1}^{P} \alpha_{kp} \prod_{i=M+1}^{D} p(x_{ni}|\mu_{kpi}, \Sigma_{kpi}) \right\} \right]. \quad (6)$$

Taking the logarithm,

$$\log p(X) = \sum_{n=1}^{N} \log \left[\sum_{k=1}^{K} \pi_k \prod_{i=1}^{M} p(x_{ni}|\mu_{ki}, \Sigma_{ki}) \left\{ \sum_{p=1}^{P} \alpha_{kp} \prod_{i=M+1}^{D} p(x_{ni}|\mu_{kpi}, \Sigma_{kpi}) \right\} \right]. \quad (7)$$

In order to find the optimal parameters for this nested GMM, we need to set partial derivatives of log $p(X)$ with respect to each of the parameters to zero:

$$\frac{\partial \log p(X)}{\partial \mu_{lj}} = 0, \quad (8)$$

for $j = 1, \ldots, M$, and $l = 1, \ldots, K$. Solving the above equation gives

$$\mu_{lj} = \frac{\sum_{n=1}^{N} \gamma(z_{nl}) x_{nj}}{\sum_{n=1}^{N} \gamma(z_{nl})}, \quad (9)$$

where

$$\gamma(z_{nl}) = \frac{\pi_l \prod_{i=1}^{M} p(x_{ni}|\mu_{li}, \Sigma_{li}) \{\sum_{p=1}^{P} \alpha_{lp} \prod_{i=M+1}^{D} p(x_{ni}|\mu_{lpi}, \Sigma_{lpi})\}}{\sum_{k=1}^{K} \pi_k \prod_{i=1}^{M} p(x_{ni}|\mu_{ki}, \Sigma_{ki}) \{\sum_{p=1}^{P} \alpha_{kp} \prod_{i=M+1}^{D} p(x_{ni}|\mu_{kpi}, \Sigma_{kpi})\}}. \quad (10)$$

For $j = M + 1, \ldots, D$, $l = 1, \ldots, K$, and $q = 1, \ldots, P$, we then need to solve the following equation, in order to get μ_{lqj}:

$$\frac{\partial \log p(X)}{\partial \mu_{lqj}} = 0. \quad (11)$$

Solving the above equation gives

$$\mu_{lqj} = \frac{\sum_{n=1}^{N} \beta(z_{nlq}) x_{nj}}{\sum_{n=1}^{N} \beta(z_{nlq})}, \quad (12)$$

where

$$\beta(z_{nlq}) = \frac{\pi_l \prod_{i=1}^{M} p(x_{ni}|\mu_{li}, \Sigma_{li}) \{\alpha_{lq} \prod_{i=M+1}^{D} p(x_{ni}|\mu_{lqi}, \Sigma_{lqi})\}}{\sum_{k=1}^{K} \pi_k \prod_{i=1}^{M} p(x_{ni}|\mu_{ki}, \Sigma_{ki}) \{\sum_{p=1}^{P} \alpha_{kp} \prod_{i=M+1}^{D} p(x_{ni}|\mu_{kpi}, \Sigma_{kpi})\}}. \quad (13)$$

From Eqs. (10) and (13), we get

$$\gamma(z_{nl}) = \sum_{q=1}^{P} \beta(z_{nlq}) \tag{14}$$

and

$$\sum_{k=1}^{K} \gamma(z_{nl}) = 1. \tag{15}$$

For $j = 1, \ldots, M$, and $l = 1, \ldots, K$, we can also obtain Σ_{lj} by solving the following equation:

$$\frac{\partial \log p(X)}{\partial \Sigma_{lj}} = 0. \tag{16}$$

Solving the above equation gives

$$\Sigma_{lj} = \frac{\sum_{n=1}^{N} \gamma(z_{nl})(x_{nj} - \mu_{lj})^2}{\sum_{n=1}^{N} \gamma(z_{nl})}. \tag{17}$$

For $j = M + 1, \ldots, D$, $l = 1, \ldots, K$, and $q = 1, \ldots, P$, we have the following equation:

$$\frac{\partial \log p(X)}{\partial \Sigma_{lqj}} = 0. \tag{18}$$

Solving the above equation gives

$$\Sigma_{lqj} = \frac{\sum_{n=1}^{N} \beta(z_{nlq})(x_{nj} - \mu_{lqj})^2}{\sum_{n=1}^{N} \beta(z_{nlq})}. \tag{19}$$

Note that

$$\sum_{k=1}^{K} \pi_k = 1. \tag{20}$$

Using the Lagrange optimization to maximize $\log p(X)$ with respect to π_k and α_{kp},

$$\pi_l = \frac{\sum_{n=1}^{N} \gamma(z_{nl})}{N} \tag{21}$$

and

$$\alpha_{lq} = \frac{\sum_{n=1}^{N} \beta(z_{nlq})}{\sum_{n=1}^{N} \gamma(z_{nl})}. \tag{22}$$

Now that we have derived the EM algorithm to train a nested GMM, let us try to delve into intuitive insights and see how this approach can provide the flexibility needed to improve performance of feature-level fusion systems. We can see from Eq. (3) that a conventional GMM classifier needs the audio and video modalities to have the same number of gaussians ($= K$). This leads to degraded performance as discussed above in Sec. 4. From Eq. (4), we can see that while using a nested

GMM, audio and video modalities can have different number of gaussians (though one should be an integer multiple of the other). In this research, we have used 8 (= K) gaussians for the video modality and 32 (= $K \times P$) gaussians for the audio modality, where the number of gaussians for the audio modality is an integer multiple ($P = 4$) of the number of gaussians for the video modality. Thus, we see that using this approach, we can have different number of gaussians for the two modalities and the design provides the required flexibility to optimally represent the two modalities together. Equation (4) also shows that P (= 4) gaussians for audio are mapped to one gaussian of video. Thus, there is a many-to-one mapping from audio gaussians to video gaussians. We discussed in Sec. 4 that there is a many-to-one mapping from phonemes to visemes in real-life data. We see that the nested GMM allows us to capture this many-to-one mapping and thus, represents a very natural way of representing the audio-visual data.

The conventional GMM classifier is a special case of the nested GMM classifier. By setting $P = 1$ in the nested GMM classifier, we obtain the conventional GMM classifier. The nested GMM classifier, thus, has all the good properties of the conventional GMM classifier with added flexibility which makes it more powerful as compared to the conventional GMM classifier. The nested GMM classifier is a superset of the conventional GMM classifier and removes the restriction that the two modalities should have the same number of gaussians. Instead, now the restriction is that the number of gaussians for one modality should be an integer multiple of the number of gaussians for the other modality. For our application, the effect of the restriction is negligible as the exact number of gaussians required is varying and data-dependent and thus, adjusting the number a little to make it an integer multiple does not result in a big difference in performance. It should also be noted that the nested GMM classifier can never perform worse than the conventional GMM classifier (provided the parameters are set for optimal performance). In the worst case, we can set $P = 1$ in the nested GMM classifier and revert to a conventional GMM classifier.

4.3. *Results and observations*

Figure 6 shows the performance comparison (under normal conditions) for the three systems proposed in this paper: raw feature-level fusion system, hybrid feature-score level fusion system, and nested GMM approach presented in Sec. 2, Sec. 4.1, and Sec. 4.2, respectively. We can see that the increased flexibility offered by the hybrid GMM classifier in the hybrid feature-score level fusion system results in an improvement in performance of about 0.5% EER compared to the raw feature-level fusion system with a conventional GMM classifier. This improvement is mainly due to the incorporation of general score-level fusion approaches which enable weighing the probabilities (scores) flexibly, taking into account the reliability of its inputs (in this case, scores obtained from the two feature-level fusion systems which are fed into the score-level fusion system). The improvement in performance can also be

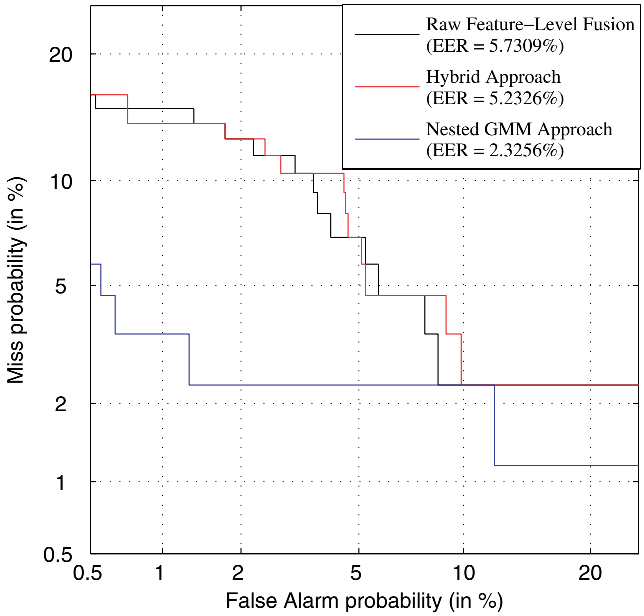


Fig. 6. Performance comparison under normal conditions for the three systems proposed in this paper.

attributed to the modified E-step in the EM algorithm (in which the probabilities are calculated using features from only one of the two modalities). However, the problems discussed in Sec. 4, namely the overfitting of the face features and the under-representation of the voice features, do exist in this system. We mentioned in Sec. 4.1 that this system consists of two feature-level fusion systems followed by a score-level fusion system. In one of the feature-level fusion systems, the face features are overfitted (though voice features are optimally represented) and, in the other feature-level fusion system, the voice features are under-represented (though face features are optimally represented). As a result, the improvement in performance is not substantial.

On the other hand, we can see that the nested GMM approach results in a performance improvement of about 3.5% as compared to the raw feature-level fusion approach with a conventional GMM classifier and about 3% as compared to the hybrid feature-score level fusion approach. The reason for the improvement is that this approach is a natural way to handle the multimodal data which is of interest in our research. It has a single model which represents both the modalities with just the right number of gaussians. So both the modalities are represented in a near-optimal fashion (unlike the other two approaches in which at least one modality is under represented or overfitted).

Figures 7–9 show the performance comparison against the three kinds of replay attacks (described in Sec. 3.1) for the three systems proposed in this paper. We

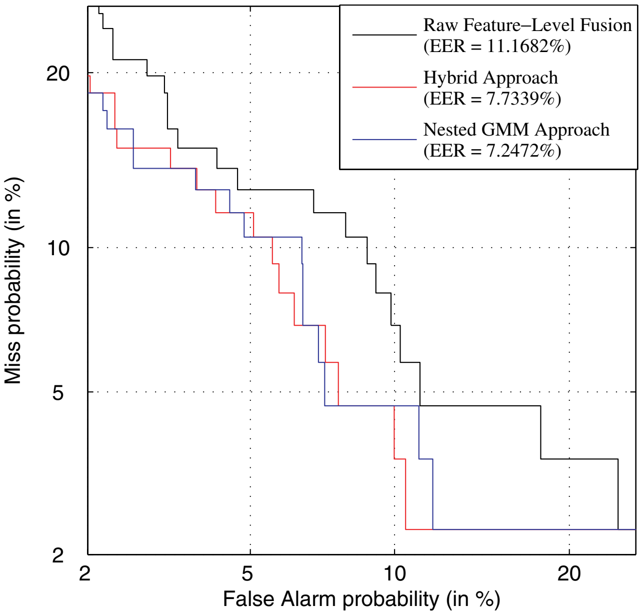


Fig. 7. Performance comparison against RP1 for the three systems proposed in the paper.

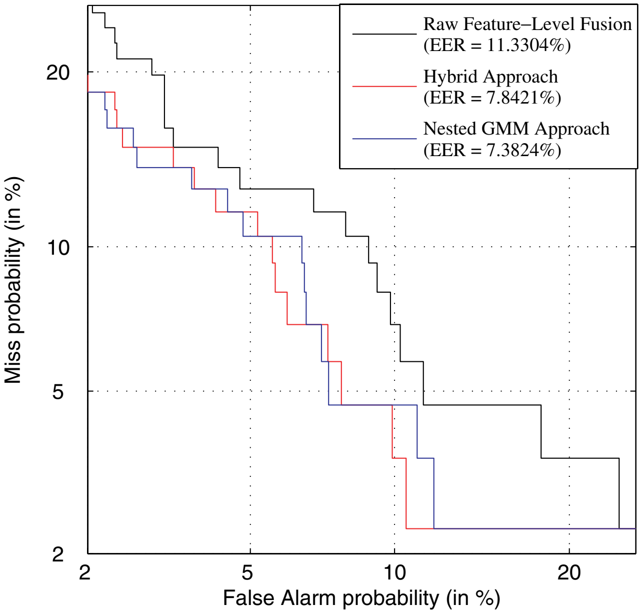


Fig. 8. Performance comparison against RP2 for the three systems proposed in the paper.

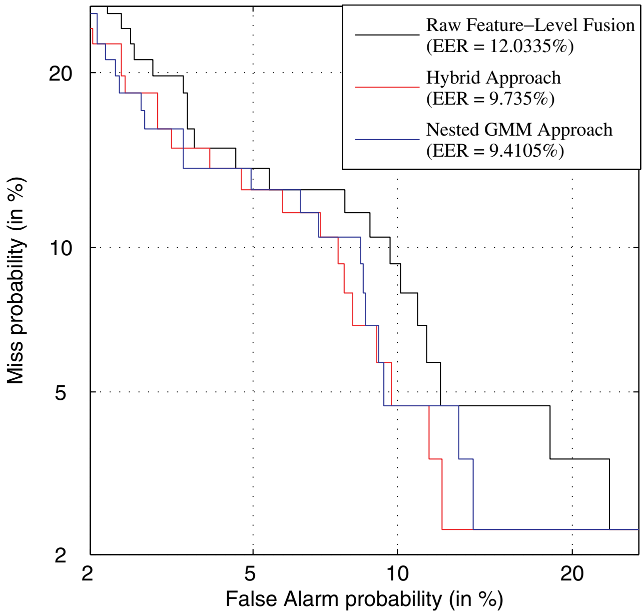


Fig. 9. Performance comparison against RP3 for the three systems proposed in the paper.

can see that the hybrid feature-score level fusion approach and the nested GMM approach both perform better than the raw feature-level fusion approach. The performance improvements reach up to about 4% EER, particularly for RP1 and RP2. Also note that the difference in the performances of the hybrid GMM classifier and the nested GMM classifier is not significant.

The better performance of the hybrid feature-score level fusion approach (as compared to the raw feature-level fusion approach) can be attributed to the multi-scale nature of the approach. Such a design consists of two feature level fusion systems at different scales (different number of gaussians) and thus can better capture the correlation between voice and face to give better robustness against replay attacks. The better performance of the nested GMM approach (as compared to the raw feature-level fusion approach) is attributed to the natural way of representing data which provides a convenient way to represent the many to one mapping from phonemes to visemes, thus allowing efficient representation and better preservation of the correlation between face and voice to give better robustness against replay attacks.

It should also be noted that, among the three systems proposed in the paper, the hybrid feature-score level fusion system has the highest computational complexity while the nested GMM approach has the lowest computational expenses (under the assumption that the feature extraction stage is the same for all three systems and the number of gaussians used is such that it gives the best performance for that system). For example, in this research, the raw feature-level fusion system has 32 gaussians for both modalities while the hybrid feature-score level fusion system has two models, one with 32 gaussians and another with 8 gaussians for both modalities (overall 40 gaussians). This indicates that the hybrid feature-score level fusion approach has higher computational complexity as compared to the raw feature-level fusion approach. On the other hand, the nested GMM approach has just one model with 32 gaussians for voice and 8 gaussians for face. Clearly, computational complexity of this approach is lower than that of the raw feature-level fusion approach.

5. Conclusions

In this paper we have shown that correlation between audio and visual data during spoken utterances offers useful information for person recognition. Assuming these modalities to be uncorrelated can result in degraded performance. Better accuracy in recognition (compared to audio-only, video-only and audio-visual systems which assume the two modalities to be uncorrelated) and a high degree of robustness against a variety of replay attacks can be obtained by exploiting this correlation between audio and visual data. In fact, robustness against certain kinds of replay attacks (RP3) can only be provided by considering this correlation.

We first proposed a simple system design which uses a conventional GMM classifier and feature-level concatenation as a means to exploit the correlation between audio and visual data. This design offers superior performance as compared to audio-only, video-only, and audio-visual systems which assume audio and video data to be uncorrelated. It was shown that the proposed fusion technique effectively captures the correlation between audio and visual data and uses it to give better performance. We demonstrated that this design demands a low amount of memory and fewer computations which makes it suitable for a low-cost real-time DSP implementation. We also showed that this design is capable of operating in real-time, and it gives a reasonably good performance in real-time as well.

We then identified the lack of flexibility offered by the conventional GMM classifier as the main reason for performance degradation of a feature-level fusion technique (as compared to a score-level fusion technique). We proposed two novel classifier designs which provide more flexibility and a better way to represent multimodal data where each modality has different characteristics. We showed that both of the classifier designs offer superior performance as compared to a conventional GMM classifier and thus help in boosting the performance of a feature-level fusion system additionally while also improving the robustness against replay attacks (as compared to a raw feature-level fusion system with a conventional GMM classifier). We also showed that the nested GMM approach, being a natural and elegant way to represent the multimodal audio-visual data, offers better performance as compared to the hybrid feature-score level fusion approach which tends to be more of a brute force technique. We also showed that the hybrid feature-score level fusion approach requires more computations as compared to a raw feature-level fusion approach, indicating redundancies inherent in the design, while the nested GMM approach requires the least computations. Overall, we showed that the nested GMM approach outperforms the other two approaches in terms of accuracy, robustness, and computational complexity.

References

[1] D. Shah, K. J. Han and S. S. Nayaranan, A low-complexity dynamic face-voice feature fusion approach to multimodal person recognition, *Proc. IEEE International Symposium on Multimedia (ISM)*, San Diego, CA, Dec. 2009, pp. 24–31.

[2] A. K. Jain, R. Bolle and S. Pankanti, *Biometrics: Personal Identification in Networked Society* (Kluwer Academic Publishers, 1999).

[3] L. Biel, O. Pettersson, L. Philipson and P. Wide, ECG analysis: A new approach in human identification, *IEEE Trans. Instrum. Meas.* **50**(3) (2001) 808–812.

[4] L. Wang, T. Tan, H. Ning and W. Hu, Silhouette analysis-based gait recognition for human identification, *IEEE Trans. Pattern Anal. Mach. Intell.* **25**(12) (2003) 1505–1518.

[5] A. K. Jain, A. Ross and S. Prabhakar, An introduction to biometric recognition, *IEEE Trans. Circuits Syst. Video Technol.* **14**(1) (2004) 4–20.

[6] D. Maltoni, D. Maio, A. K. Jain and S. Prabhakar, *Handbook of Fingerprint Recognition*, 2nd edn. (Springer, 2009).

[7] R. Brunelli and D. Falavigna, Person identification using multiple cues, *IEEE Trans. Pattern Anal. Mach. Intell.* **12**(10) (1995) 955–966.

[8] J. Kittler, M. Hatef, R. P. W. Duin and J. Matas, On combining classifiers, *IEEE Trans. Pattern Anal. Mach. Intell.* **20**(3) (1998) 226–239.

[9] R. W. Frischholz and U. Dieckmann, BioID: A multimodal biometric identification system, *Computer* **33**(2) (2000) 64–68.

[10] A. Ross and A. K. Jain, Multimodal biometrics: An overview, *Proc. European Signal Processing Conference (EUSIPCO)*, Vienna, Austria, Sept. 2004, pp. 1221–1224.

[11] M. Faundez-Zanuy, J. Fierrez-Aguilar, J. Ortega-Garcia and J. Gonzalez-Rodriguez, Multimodal biometric databases: An overview, *IEEE Aerosp. Electron. Syst. Mag.* **21**(8) (2006) 29–37.

[12] C. C. Chibelushi, J. S. D. Mason and F. Deravi, Integration of acoustic and visual speech for speaker recognition, *Proc. European Conference on Speech Communication and Technology (Eurospeech)*, Berlin, Germany, Sept. 1993, pp. 157–160.

[13] J. Luettin, N. A. Thacker and S. W. Beet, Speaker identification by lipreading, *Proc. International Conference on Spoken Language Processing (ICSLP)*, Philadelphia, PA, USA, Oct. 1996, pp. 62–65.

[14] S. Ben-Yacoub, Y. Abdeljaoued and E. Mayoraz, Fusion of face and speech data for person identity verification, *IEEE Trans. Neural Netw.* **10**(5) (1999) 1065–1074.

[15] C. Sanderson and K. K. Paliwal, Identity verification using speech and face information, *Digit. Signal Process.* **14**(5) (2004) 449–480.

[16] P. S. Aleksic and A. K. Katsaggelos, Audio-visual biometrics, *Proc. IEEE* **94**(11) (2006) 2025–2044.

[17] H. Bredin and G. Chollet, Audiovisual speech synchrony measure: Application to biometrics, *EURASIP J. Appl. Signal Process.* **2007**(1) (2007) 179–189.

[18] C. Sanderson, *Biometric Person Recognition: Face, Speech, and Fusion* (VDM Verlag, 2008).

[19] W. Karam, H. Bredin, H. Greige, G. Chollet and C. Mokbel, Talking-face identity verification, audiovisual forgery, and robustness issues, *EURASIP J. Adv. Signal Process.* **2009**(4) (2009) 1–15.

[20] W. H. Sumby and I. Pollack, Visual contribution to speech intelligibility in noise, *J. Acoust. Soc. Am.* **26**(2) (1954) 212–215.

[21] D. W. Massaro, *Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry* (Lawrence Erlbaum Associates, 1987).

[22] G. Potamios, C. Neti, J. Luettin and I. Matthews, Audio-visual automatic speech recognition: An overview, Chapter 10 in: *Issues in Visual and Audio-Visual Speech Processing* by G. Bailly, E. Vatikiotis-Bateson and P. Perrier (MIT Press, 2004).

[23] A. P. Dempster, N. M. Laird and D. B. Rubin, Maximum likelihood from incomplete data via the em algorithm, *Journal of the Royal Statistical Society* **39**(1) (1977) 1–38.

[24] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett and N. L. Dahlgren, *The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus.* CDROM: NIST order number PB91-100354, 1993.

[25] K. M. Kryszczuk and A. Drygajlo, Color correction for face detection based on human visual perception metaphor, *Proc. Workshop on Multmodal User Authentication (MMUA)*, Santa Barbara, CA, USA, Dec. 2003, pp. 138–143.

[26] T. Lehn-Schioler, L. K. Hansen and J. Larse, Mapping from speech to images using continuous state space models, *Proc. International Workshop on Machine Learning for Multimodal Interaction (MLMI)*, Martigny, Switzerland, June 2004, pp. 136–145.

[27] R. Gocke, Current trends in joint audio-video signal processing: A review, *Proc. International Symposium on Signal Processing and Its Applications (ISSPA)*, Sydney, Australia, Aug. 2005, pp. 70–73.

[28] G. Chetty and M. Wagner, Liveness detection using cross-modal correlations in face-voice person authentication, *Proc. Interspeech — European Conference on Speech Communication and Technology (Eurospeech)*, Lisboa, Portugal, Sept. 2005, pp. 2181–2184.

[29] J. Campbell, Speaker recognition: A tutorial, *Proc. IEEE* **85**(9) (1997) 1437–1462.

[30] P. Viola and M. J. Jones, Robust real-time face detection, *International Journal of Computer Vision* **57**(2) (2004) 137–154.

[31] D. A. Reynolds, Speaker identification and verification using Gaussian mixture speaker models, *Speech Communication* **17**(1–2) (1995) 91–108.