A Low-Complexity Dynamic Face-Voice Feature Fusion Approach to Multimodal Person Recognition

Dhaval Shah, Kyu J. Han, Shrikanth S. Nayaranan
Ming Hsieh Department of Electrical Engineering, Viterbi School of Engineering
University of Southern California, Los Angeles, CA, USA
dhavalys@usc.edu, kyuhan@usc.edu, shri@sipi.usc.edu

Abstract

In this paper, we show the importance of face-voice correlation for audio-visual person recognition. We evaluate the performance of a system which uses the correlation between audio-visual features during speech against audio-only, video-only and audio-visual systems which use audio and visual features independently 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 scheme has lower error rate than all other comparison conditions and is more robust against replay attacks. The simplicity of the fusion technique also allows the use of only one classifier which greatly simplifies system design and allows for a simple real-time DSP implementation.

1. Introduction

Biometric recognition holds tremendous promise for security applications. Biometrics can include a wide range of modalities including voice, face, fingerprint, retina and iris [1]. Each modality has its own advantages and limitations in terms of robustness, accuracy and usability/user acceptance. For instance, using iris information offers very high accuracy and robustness but its usability/user acceptance is limited. On the other hand, modalities like face and voice (the modalities of interest in this paper) that can be accessed in an unobtrusive way and have higher user acceptance have restricted use due to robustness and accuracy issues. We need a way to deal with these issues in order to build real-life systems incorporating these modalities. Availability of robust solutions, however, promise practical applications such as personal computer login and location access.

One of the promising venues for improving overall performance, in terms of both accuracy and robustness, is to consider combining individual modalities under the premise that both redundancy and complementarity in information can be advantageously utilized. Many different ways of combining the face and voice modalities have been presented in the literature [2]-[8]. For simplicity, it is often assumed that the audio and video features are independent. In real-life speech communication scenarios i.e., considering the speaking face, however, this assumption does not hold true. A person's face dynamically and systematically changes as he speaks and there is a strong correlation between this deformation and the spoken utterance. Hence, their joint modeling can be potentially beneficial to improve person recognition performance.

In this paper, we present an audio-visual recognition technique using feature-level fusion which captures the correlation between audio-visual (speech and face) features. We show that such a system gives better accuracy in normal situations and is more robust to replay attacks[1] than audio-only, video-only and audio-visual systems which use audio and video independently. We also show that this technique inherently uses audio and static video features for recognition and dynamic video features for liveness detection (a sub-application in person detection) without adding any extra complexity. Lastly, we show that the low complexity of the proposed design allows for an implementation which works in real-time. The proposed design has been implemented on a DSP processor (TMS320C6713) to work in real-time and it gives on line performance comparable to off line evaluations.

2. Proposed System Description

We first describe the VidTIMIT database used in this research. We then present an overview of the feature extraction stage for speaker recognition and face recognition and

[1]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.

justify the choice of features used. Then, we present the proposed multimodal fusion technique. We present several possible ways of fusing the modalities with 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 the GMM model for this research.

2.1. VidTIMIT Database

The VidTIMIT database [8] is an audio-visual database comprised of audio-visual recordings of 43 people reciting sentences from the test section of the TIMIT corpus [12]. 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, possibility of mood and appearance changes is expected and 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 training data sparsity issue (although reflective of what is 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 quality factor of 90%, the background is fairly plain and constant with only the speaker's frontal face in the picture. This alleviates 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 goodness of the extracted features highly 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 audio and face recognition are highly developed and many different ways of capturing features are available in the literature. Our approaches to feature extraction are among such developed strategies and described in the subsequent sub-sections.

2.2.1 Voice Feature Extraction

As a pre-processing step on the audio signal, we perform pre-emphasis to compensate for the high frequency fall-off. We then use the short-term analysis technique using a 25ms window with 50% overlap between adjacent windows. We apply 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 the widely accepted MFCC features for our research due to their demonstrated superior performance [2]-[3], [5]-[6]. We use the first 13 MFCC features (12 + energy) along with their delta and accelerations to form a 39-dimensional audio feature vector for each frame.

2.2.2 Face Feature Extraction

As discussed in the VidTIMIT database section, we assume that each frame of video has just the person of interest in it with a frontal view of the face (VidTIMIT provides video data in this form). We first detect the face and throw away the background information (box the face) [9]. We get different sizes for the boxed faces because of the random perturbation of the camera zoom factor. We need to compensate for this effect and rescale the face image into a standard size suitable for feature extraction. We use bilinear interpolation and get the standardized face image of 24x24 pixels.

Many different kinds of features can be used for face recognition [2]-[6], [10]. The most widely used ones include Eigenfaces, DCT and Gabor wavelets. Eigenfaces are well suited for face recognition. In this technique, the 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 the person per frame instead of averaging out the information contained in neighboring frames. Gabor wavelets and DCT suit our requirement 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

(a) Original Image (b) Reconstructed Image

Figure 1. (a) Original face image: Input to the face feature extractor. (b) Face image reconstructed using the DC coefficient and the proposed DCT features (first 2 AC coefficients: 1 in either direction.). Note that these figures have been enlarged for better display.

performance comparable to Gabor wavelets and 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 paper. To extract features, we segment the face images into blocks of size 6x6 pixels and calculate the DCT of each block separately. From video compression theory, we know that the lower order DCT coefficients contain most of the structural information and even after throwing away the higher order coefficients, a reasonable re-construction of the original image can be achieved. The DC coefficient contains the average gray level of the block and is most affected by illumination changes. It does not carry any information about variations in the block. Thus, we also discard this coefficient. We use information of the first AC DCT coefficients in either direction as features for the face image. Thus, we end up with a 32-dimensional feature vector for each face image (2 features per block for 16 blocks).

The choice of the standardized face image size, block size and number of features per block was made empirically. A small image size was desired to minimize redundant information and reduce the calculations to facilitate real-time implementation. A smaller block size reduces computations for DCT calculation and ensures that 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 6x6 seemed reasonable for our experiments. The choice of just 2 features per block was made based on human feedback. As shown in Figure 1, a 24x24 image reconstructed using a 6x6 block size with all except for the first 2 AC co-efficients (1 in either direction) set to 0 (except the DC co-efficient) for each block still contains enough information for a human to recognize a person and thus, is deemed to contain enough person dependent information in it. This choice was also verified experimentally (a face recognizer using these features gives a reasonably good performance as reported in Section 3). 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 Section 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 [2]-[3]. 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. Also, such systems require separate classifiers for each modality which may complicate system design. Intermediate integration techniques use multi-stream HMMs. Though better than the late integration techniques, 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.

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 by weighted summation, etc. This enables the use of a single classifier which simplifies system design. It also takes into account the correlation between 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 the fusion technique 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 classifier. 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 the 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 is available. The first scheme just adds redundant data, which while helps smooth out discontinuities between adjacent frames, may not be of use for the recognition task. It also adds extra amount of processing. On the other hand, the hybrid technique is more suitable for the recognition task which has 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 Section 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 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 Section 3). We show below in Section 3 that the proposed fusion technique preserves both the complementary and the 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 DTW, GMM, HMM, SVM and NN) and significant literature is available on them [2]-[3], [10]. 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. The GMMs (single state HMM) exploit this. They give a similar performance as compared to HMMs and computationally are more efficient than the HMMs. Other advantages of 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 proposed technique. 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

The VidTIMIT database consists of 43 speakers, the first 19 are female and the remaining 24 are male (alphabetically arranged). We use the last 4 females and last 4 males as impostors and the remaining 35 speakers as clients. There are 10 sentences per speaker. For the clients, we used 8 of them (sessions 1 and 3) for training the client model and 2 (session 2) for testing. We use all 10 sentences of impostors for impostor trials. In short, we have 70 client trials and 80 impostor trials. For training the impostor model, we use the UBM technique. Ideally, we need to train the UBM 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 impostor trials and client train and test utterances).

We first perform experiments to demonstrate that the proposed system is more accurate than audio-only, video-only and the audio-visual system in which audio and video are considered independent of each other (we simulate this by randomizing the video frames). For this, we use the client trials and the impostor trials mentioned above.

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 attack (RP1) consists of pure audio from the client trials combined with video from one of the impostor trials. Care has been taken that the gender of both client and the impostor used to form this replay attack is the same. It represents an attack where client audio is recorded by an impostor and used to breach the 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 the 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 to conclude whether the 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 the correlation between audio and video.

3.2. Results and Observations

To simplify system design, we use the same number of Gaussians for each client model and an integer multiple of this number as the number of Gaussians for the impostor model. These simplifications might lead to sub-optimal results. Nevertheless, the performance degradation is negligible (for example, for 15 Gaussians, an FR of 8.57% can be reduced to 7.14% for the same FA by fine tuning the number of Gaussians for the impostor model) and it simplifies system design to a great extent and helps us compare the performance of the systems in a better and simpler way. We vary the number of Gaussians for the client model (Ng) from 1 to 50 and calculate the FA and FR values. Instead of plotting the FA vs. FR curve, we plot FA and/or FR vs. Ng. This benefits us in many ways. The foremost advantage is that for the required FA and/or FR values, we directly get the number of Gaussians for the design. This speeds up the design process. For real-time systems, computational complexity (or latency) and memory requirement increases as Ng increases. Thus, FA and/or FR vs. Ng plot allows us to compare FA, FR, latency and memory requirement for the real-time system to be designed. These are as important as accuracy and robustness for real-time systems. It also enables us to study how the system behaves when Ng is varied. For a given Ng, the FA and FR values obtained are always data dependent. For real-life systems, when test data is unknown, we need to ensure stability of the system in the given region of operation (specified by Ng). The FA and/or FR vs. Ng plot enables us to do this. For example, for the audio only plot in Figure 2, when Ng is 1, FR is about 40% while FR increases to 80% when Ng is increased to 2. This shows us that the system is unstable in this region and reliable performance cannot be guaranteed. Such conclusions are impossible to make when FA vs. FR curve is used and we might end up designing an unstable and non-reliable system.

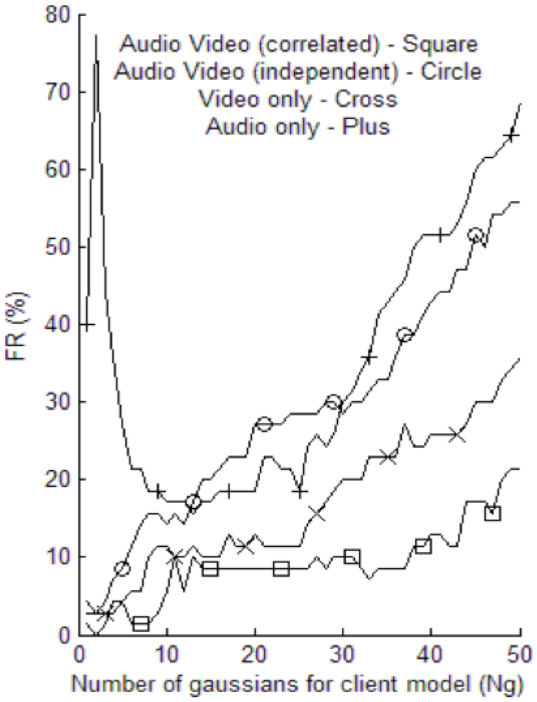

Figure 2. Client Trials (FR vs. Ng).

3.3. Discussion on the Results

From Figure 2, we see that the audio-only system has the highest FR values. One reason for this is that the audio in this database is noisy. The video-only system gives the second best FR. This is because the video in the database is comparatively clean. The FR values for the audio-visual system where audio and video are assumed independent of each other are higher than the video-only system. This proves that the assumption that audio and video are independent does not hold and such assumptions can prove detrimental to performance. The best FR values are given by the proposed system. This indicates that exploiting the correlation between audio and visual data can lead to significant improvement in accuracy.

From Figure 3, we see that in most regions of operation, FA is 0 for the video-only and both types of audio-visual systems. The FA for the audio-only system is again high due to noisy data. It is worthy to note that small values of FR for the proposed system does not come at the cost of increased FA.

From Figures 2 and 3, it can be seen that the region where Ng is very small is unstable for most systems. The performance in this region is unpredictable and data depen-

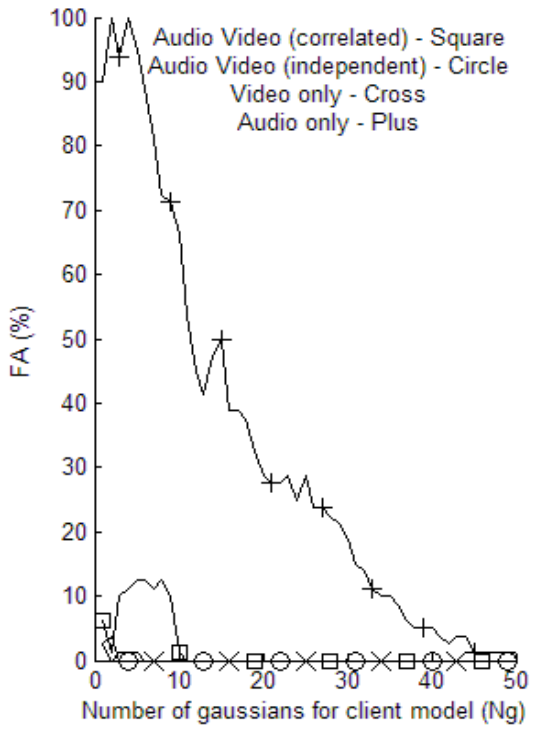

Figure 3. Imposter Trials (FA vs. Ng).

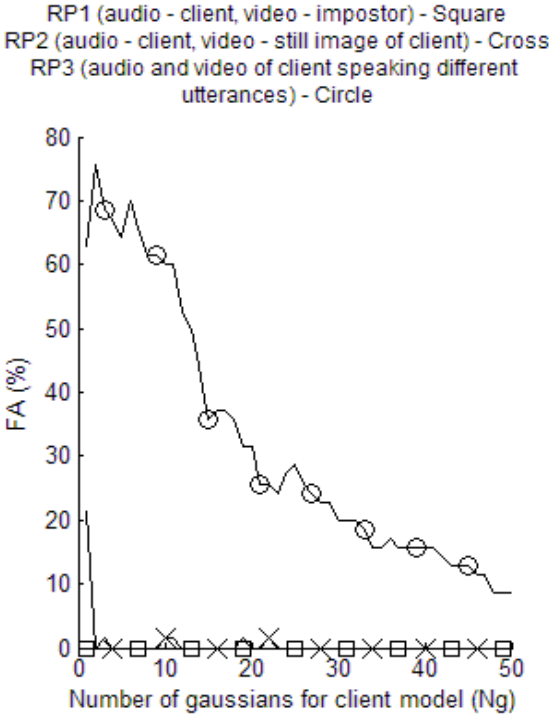

Figure 4. Replay Attacks (FA vs. Ng).

dent. Also, when Ng becomes very large, FR starts increasing for all systems which shows that performance degrades due to overfitting when excessively large number of Gaussians are used. We can see that the stable region of operation is the smallest for audio-only system and is the largest for the proposed design. This shows that the proposed design promises greater stability and reliability.

From Figure 4, we can see that the proposed design is very robust against RP1 and RP2 replay attacks. The FA values in almost all regions of operation are 0. The robustness against RP1 is due to the mere fact that this is a multimodal system. The video in RP1 is of an impostor and thus, the video modality is responsible for this robustness. RP2 has both audio and video (still image) of the client and still the system is very robust. This shows that the system has an inherent liveness detector (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 the 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 using visual data only speech recognition is inherently limited and not reliable. Second is that this adds complexity into the design which makes this technique unsuitable 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. Figure 4 shows that for RP3, FA as low as 10% is possible using the proposed technique where other audio-visual systems would breakdown. We see that for low values of Ng, FA is very high and it drops steadily as Ng increases. This is because at low values of Ng, different sound units and the corresponding facial expressions are averaged out. Thus, the system is not able to decide reliably if the voice and face information correspond to the same utterance and hence FA is high. As Ng is increased, each Gaussian represents fewer number of sound units and thus, the averaging effect is reduced. Now, the system has enough information to decide whether the voice and face information correspond to the same utterance and hence, FA decreases. 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.

Table 1 shows the best-case comparison between the pro-

Table 1. Best-case comparison among systems.

System	Best FR	Best FA
Audio only	15.71%	1.25%
Video only	5.71%	0.00%
Audio-Visual (independent)	11.43%	0.00%
Audio-Visual (correlated)	2.86%	0.00%

Table 2. Some useful design values.

Ng	FR (%)	FA (%)	FA for Replay Attacks (%)		
			RP1	RP2	RP3
6	2.86	10.00	0.00	0.00	70.00
9	5.71	6.25	0.00	0.00	65.71
16	8.57	0.00	0.00	0.00	37.14
31	12.86	0.00	0.00	0.00	20.00
48	30.00	0.00	0.00	0.00	8.57

posed design, audio-only system, video-only system and audio-visual system which uses audio and video independently. Except for the audio-only system, all other systems are capable of achieving an FA value of 0%. Due to the noisy nature of audio data, the audio-only system is unable to achieve this value. Due to the same reason, the audio-only system also gives the highest best-case FR value. The video-only system gives the second lowest best-case FR value. FR value of audio-visual system which uses audio and video independently is higher than that of the video-only system which again proves that performance can degrade if audio and video are assumed independent. The best FR value (as low as 2.86%) is given by the proposed design which shows that the proposed design offers superior performance.

Table 2 lists some useful design values. Accuracy and robustness measures (in terms of FA, FR and FA against replay attacks) are listed against the design factor Ng (number of Gaussians for the client model) to speed up the design process. Though the exact values are data dependent, depending on the accuracy and robustness requirements, a rough estimate of the design factor Ng can be obtained from the table.

A simpler version of the proposed technique has been implemented on a DSP processor (TMS320C6713) using 5 Gaussians for client models. The system demands about 100kb of program memory and 512kb 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 by the End-of-Speech Detector (2 seconds of silence is required to conclude 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 mic, view angle for the person's face, etc are controlled but background noise, lighting conditions, etc. are not controlled).

4. Conclusions

In this paper we have shown that the correlation between audio and visual data during spoken utterances offers useful information for person recognition. Assuming these modalities to be independent can result in degraded performance. Better accuracy in recognition and 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 have proposed a simple feature level concatenation technique of multimodal fusion as a means to exploit the correlation between audio and visual data. The proposed design offers superior performance as compared to audio-only, video-only and audio-visual systems which assume audio and visual data to be independent which shows that the proposed fusion technique effectively captures the correlation between audio and visual data and uses it to give better performance. We demonstrated that the proposed design demands low amount of memory and less computations which makes it suitable for a low-cost real-time DSP implementation. We also showed that the proposed design is capable of operating in real-time and it gives a reasonably good performance in real-time as well.

References

[1] Ross, A. and Jain, A. K., "Multimodal biometrics: an overview," *Proc. EUSIPCO*, pp. 1221-1224, Sept. 2004.

[2] Aleksic, P. S. and Katsaggelos, A. K., "Audio-visual biometrics," *Proc. IEEE*, vol. 94, no. 11, pp. 2025-2044, Nov. 2006.

[3] Sanderson, C., "Automatic person verification using speech and face information," *Ph.D. Thesis*, Griffith University, Queensland, Australia, 2003.

[4] Chetty, G. and Wagner, M., "Face-voice authentication based on 3D face models," *Proc. ACCV*, pp. 559-568, Jan. 2006.

[5] Chetty, G. and Wagner, M., "Speaking faces for face-voice speaker identity verification," *Proc. Interspeech*, pp. 513-516, Sept. 2006.

[6] Erzin, E., Yemez, Y., and Tekalp, A. M., "Multimodal speaker identification using an adaptive classifier cascade based on modality reliability," *IEEE Trans. Multimedia*, vol. 7, no. 5, pp. 840-852, Oct. 2005.

[7] Chetty, G. and Wagner, M., "Audio-visual speaker verification based on hybrid fusion of cross modal features," *Proc. PreMI*, pp. 469-478, Dec. 2007.

[8] Sanderson, C., *Biometric person recognition: face, speech, and fusion*. VDM Verlag, June 2008.

[9] Krishna, S., "OpenCV Viola-Jones face detection in MATLAB." http://www.mathworks.com/matlabcentral/fileexchange/19912.

[10] Sanderson, C. and Paliwal, K. K., "Polynomial features for robust face authentication," *Proc. ICIP*, pp. 997-1000, Sept. 2002.

[11] HTK v3.4, Cambridge University, http://htk.eng.cam.ac.uk.

[12] Garofolo, J. S., Lamel, L.F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., and Dahlgren, N. L., "The DARPA TIMIT acoustic-phonetic continuous speech corpus CDROM," NIST order number PB91-100354, 1992.