Audio-Visual Person Authentication with Multiple Visualized-Speech Features and Multiple Face Profiles

Amitava Das, Ohil K. Manyam[1], Makarand Tapaswi[1]
Microsoft Research India. [1]*student interns at MSR-India.*
Email: amitavd@microsoft.com.

Abstract

We present an Audio-visual person authentication system which extracts several novel "Visualized-Speech-Features" (VSF) from the spoken-password and multiple face profiles using a simple user-interface and combine these features to deliver high performance and resilience against imposter attacks. The spoken password is converted to a string of images formed by several visualized speech features. A compressed form of these VSFs preserves speaker identity in a compact manner. Simulation results on an in-house 210-user AV-user-ID database collected with wide variations of users in real-life office environments demonstrate separable distributions of client and imposter scores (0% EER), while offering low storage and computational complexities compared to conventional AV user-recognition methods.

1. Introduction

Typical audio-visual person recognition systems use a frontal face image and a spoken password or speech samples and perform either person identification (given media samples, detect who the user is) or person authentication (given media samples and an identity-claim, verify whether the claimed user is the true user or not) tasks. In this paper, we propose an audio-visual user-authentication system which harnesses its power from multiple face profiles or poses of a person and the use of multiple spoken passwords. The proposed system not only offers a high degree of security and robustness against imposter attack but also delivers built-in liveliness detection capability highly useful for online access-control systems which need to prevent "bots" or non-human "agent-software" from establishing bogus accounts for malicious purposes. The proposed AV system therefore has a high potential to replace the traditional text-password based access control mechanism. Details of the proposed AV system reveal that it meets all the essential requirements of on-line user-authentication systems namely, a) "liveliness" detection (only humans should be able to log in and not machines), b) high resilience against imposter attacks and c) high accuracy and speed of operation. Other operational requirements of any authentication system will be: a) low complexity of computation and storage so that the system can be scaled to many users, b) enrollment of a new user should not require any massive re-estimation, c) ease of use of the system during enrollment and actual usage, d) ability to change "password", e) easy availability of sensors, and f) social acceptance of the biometrics. The proposed AV authentication system offer a number of unique advantages in these regards. First of all, it uses two biometrics (speech and face image), which people share quite comfortably in everyday life. No additional private information is given away. The AV sensors are virtually everywhere as many laptops these days are equipped with integrated web-cam and it is also cheap and easy to hook a web-cam to a regular PC. The combined use of an intrinsic biometric (face) along with a performance biometric (voice) also offers a heightened resilience against imposter attacks, while providing a way to "change password'. The requirement that the user needs to utter a password when indicated, also ensures liveliness. However this creates a problem that the password can be overheard by an imposter. Using the same password by an imposter may hurt the speech based authentication component a little (as we will see) but the integration of face feature eventually aids to robustness against imposter attacks. Usual real-life environmental noise also hurts the speech performance. Once again the joint AV mode helps to overcome this robustness issue too.

In this paper, we propose an AV person recognition framework suitable for online user-authentication, which exploits a number of powerful features extracted from the face images and the spoken-passwords of the user, delivering high accuracy and a significant robustness against imposter attacks. First of all, unlike traditional face recognition systems which use a single

978-0-7695-3476-3/08 $25.00 © 2008 IEEE
DOI 10.1109/ICVGIP.2008.106

frontal pose; our method uses multiple profiles or poses of a person's face. This captures the identity of a person much better than a single face profile. Secondly, the proposed framework extracts multiple "visualized-speech-features" or VSFs from the spoken-password. The 1-D speech signal is converted into a set of "images" formed by 2-D normalized feature-time representations. These images or VSFs of a user can be directly compared against stored VSFs of enrolled users. However, as we extract multiple VSFs from a single password and as we store up to 4 templates of the passwords per user, direct use of the VSFs pose a storage problem and complexity problem. Therefore, we use a compact representation of the VSFs using DCT which keeps complexity and storage requirements low while maintaining the high discrimination power of the VSFs.

For classification, we used a Multiple Nearest Neighbor Classifiers (MNNC) framework which integrates the multiple face-profiles and VSF features, using a set of feature-codebooks for each user. The MNNC framework offers excellent performance (0% EER as well as a non-overlapping distribution of the client and imposter scores) at low complexity as demonstrated on extensive evaluations done on a unique 210 people in-house AV biometric database created for this research. The use of multiple poses, multiple passwords and the dynamics of user-interaction inherent to the proposed AV system offers a very high resilience to imposter attacks and make it virtually impossible for any imposter to get access with photos of and/or recorded passwords of the client.

For the face modality, conventional methods use variations of PCA [11] with pose normalization using 2D or 3D models followed by a PCA-based dimension-reduction and a nearest neighbor matching of the reduced-feature template. Several studies [10, 13, 14] have shown that face recognition performance improves dramatically if a face video or a sequence of face images of a person is used as opposed to the use of a single frame. Spatial-temporal methods such as [15] also deliver good performance exploiting a sequence of face profiles. A good review can be found in [8]. Due to the complex model-based approach and sequence analysis requirements, all these methods require high complexity as well as processing of reasonably large sets of image frames before reaching a decision. In this paper, we present a multiple pose based face recognition method, which does not require the complex modeling and large training data as needed by some of the earlier methods mentioned here.

The paper is organized as follows: Section 2 gives an overview of the current AV person recognition methods. Section 3 describes the proposed visualized speech features for speech. Section 4 presents the multiple-profile face features and section 5 the multiple nearest neighbor classifier framework. Section 6 presents the database and experimental trials followed by the results and discussion presented in Section 7. Finally section 8 presents the conclusion and future work.

2. Overview of Current Audio-Visual User-Authentication Methods & Novelties of the Proposed Method

Majority of the recent audio-visual biometric authentication methods [1-7] use separate speech and face based classifiers and then combine these scores with apply various late fusion techniques such as sum, product, voting, mixture-of-experts, etc. A good survey of AV person authentication methods can be found in [1].

For the speech modality, two main types of algorithms are used a) text-independent (TI) and b) text-dependent (TD). TI methods [16, 19, 20] assume that the password users are uttering can be anything and treat the sequence of extracted features from the speech utterance as a bag of symbols. Speakers are modeled as distributions in the feature space by VQ codebooks [20, 7] or by GMM [19] and the task amounts to finding from which speaker distribution the test feature-vector-set is most likely to have originated. TD speaker recognition methods [17, 18] assumes that the system "knows" the password the user is saying and exploit the feature dynamics to capture the identity of the speaker. TD methods compare the feature vector sequence of the test utterance with the "feature-dynamics-model" of all the speakers, which can simply be the stored templates of feature vector sequence or can be HMMs trained by a large number of utterances of the same password. For classification, conventional TD methods use dynamic classification methods such as Dynamic Time Warping (DTW) [18] or HMM [17].

Important thing to note is that all the prevalent speaker recognition methods cut the 1-D acoustic signal of the spoken password into a set of overlapping frames, extract some feature (Mel-Frequency Cepstral Coefficient (MFCC) [16] being the most popular feature) and then handle these sets of extracted features as a feature-sequence or a bag of features. One significant novelty of our proposed method is to create a "holistic views" of the spoken password by creating a visual representation of speech or treat speech as a set of images, which we call "visualized speech feature" or VSFs. These VSFs capture the speaker-identity or the speaking style of a person efficiently by capturing speaker-specific speech dynamics typically exhibited in the co-articulation of various sound units. One well-known speech visualization approach is the "spectrogram" used by speech scientists, which are essentially magnitudes of short-term spectrum of frames of speech, stacked together to create an image. Spectrograms are widely used by speech scientists to detect speech sounds and speech pathological problems. We are not aware of any automated spectrogram processing method. The proposed Visualized Speech Features are inspired by spectrograms, the novelty being the methods proposed here facilitating the automated use of these "images" for speaker recognition purpose.

Finally, note that speech is a 1-D signal while image is a 2-D signal and the rates of processing of the two media and the resulting data rates for the two streams are different. The feature extraction methods proposed here overcome this by creating similar fixed-dimension feature vectors from the face-images as well as from the spoken-passwords. This avoids the problem of rate-asynchrony that hurts traditional AV person recognition methods. We describe the feature extraction part of our proposed AV method next.

3. Visualized Speech Features

The spoken password is converted into "Visualized Speech Feature" or VSF's as follows:

Step-1: VSF Formation: Given a speech segment of N_1 frames, a particular feature extraction method is employed to extract an L-dimensional feature vector from each frame. These N_1 vectors are then stacked to form a 2-D matrix of size N_1 x L. This matrix forms an image or a visual representation of the spoken password or the VSF. The VSF essentially captures the time-varying dynamics of the feature in an image. Figure 1 and 2 present two types of VSFs: VSF-MFCC using MFCC [17], and the Spectrogram using DFT magnitude as features.

Few important observations can be made from these two figures: a) The VSFs do look similar for the same person and different from person to person, and b) the X-axis (frame number) has different length since even if the same person utters the same password, one utterance will be a bit different from another in terms of duration. This requires some kind of normalization. Only then we can compare two such VSF images.

Step 2: VSF Normalization: Usual interpolation methods for images such as bilinear or bicubic interpolation [22] (the "imresize" command in Matlab for example) can be used to normalize the VSFs. If the resizing is done on the VSF of the entire password, the "matching" often is found to be poor due to bad alignment of internal components of the VSF image. An example is shown in Figure 4(a), where two spectrograms of two client passwords are time-normalized by resizing the entire image. Notice the misalignment of internal important structures.

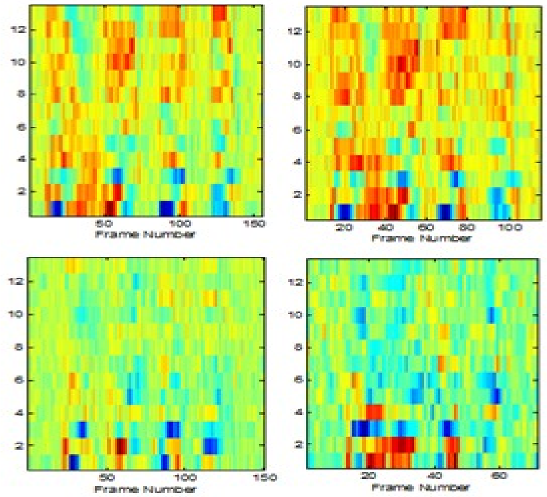


Figure 1: VSF-MFCC of two utterances of the password "9054" by client (top-row: 150 frame and 120 frame long) and the same password spoken by two imposters (bottom row: 150 and 80 frame long)

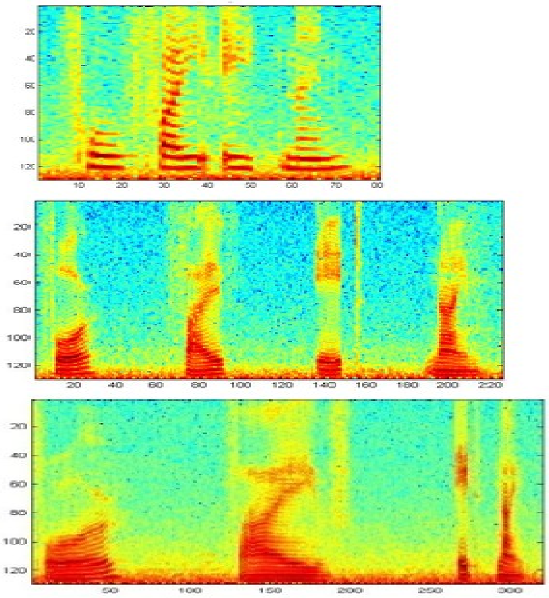


Figure 2: Spectrograms of utterances of an imposter (60 frames) and two versions by client (220 & 310 frames long) all uttering the same password.

We found that it helps to segment speech into simple segments using some acoustic feature. One simple acoustic feature, which can be detected with high accuracy is the “voiced/unvoiced” or the “V/U” feature. Speech segment in any language can be broken into these V type segments (e.g. “a”, “e”, “o”) which are associated with distinct periodicity and unvoiced segments (e.g. ‘s’, ‘t’) which are either noise like or noise associated with a short burst. A spoken password can be easily broken into a set of V/U “image” segments (see Figure 3). Details of such segmentation methods are beyond the scope of this paper.

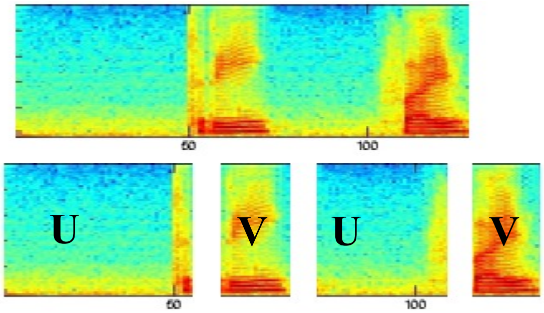
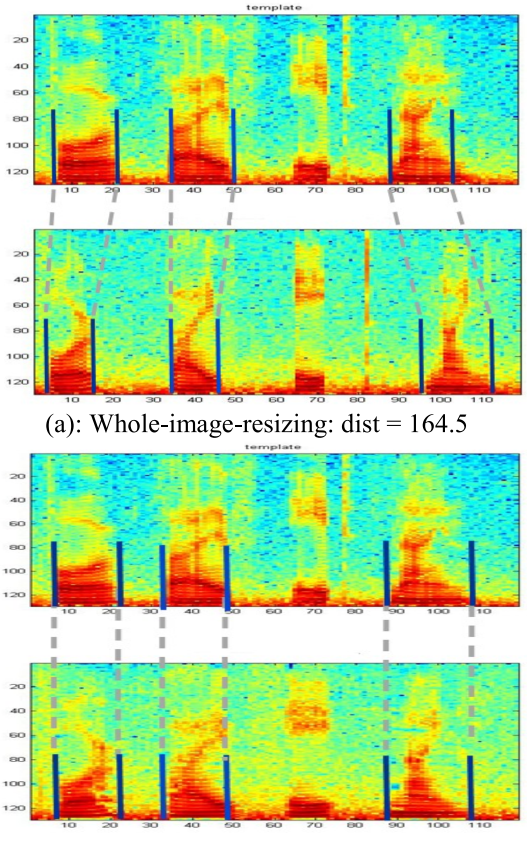


Figure 3: Breaking of Speech VSF (spectrogram here) into voiced and unvoiced segments.



(a): Whole-image-resizing: dist = 164.5

(b): Segment-based-resizing: dist=114.5

Figure 4: Various normalization methods for VSFs

Then individual segments are resized to similar components of a template as shown in Figure 4b. The resulting segment-based normalization works better than the whole-VSF-resizing approach as evident from the resizing results shown in Figure 4a.

Step 3: Compressed Feature Dynamics or CFD: Compact representation of the VSFs: The 2-D VSF representation creates a need of large storage and computational complexity, especially if we want to store many templates of a password per user or many passwords per user and a large number of users. Thus we create a compact 1-D representation of the VSF, we call compressed feature dynamics or CFD by applying a truncated DCT method. The compression power of DCT [21] captures the essential information contained in the 2-D VSF image in a small set of coefficients. We omit the DC value and keep the top $K=(m^2-1)$ coefficients in a zigzag scan (Figure 5) to form the K-dimension CFD vector. Thus a 3 second spoken password (8 KHz sampling rate) or 24000 numbers can be represented by a single 63-dimension (taking m=8) CFD feature vector or by 63 numbers.

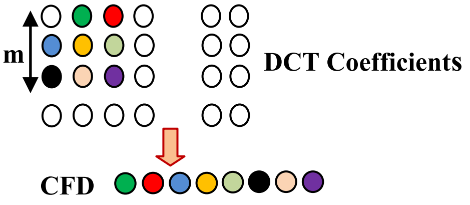


Figure 5: Creating CFD from VSF by truncated DCT

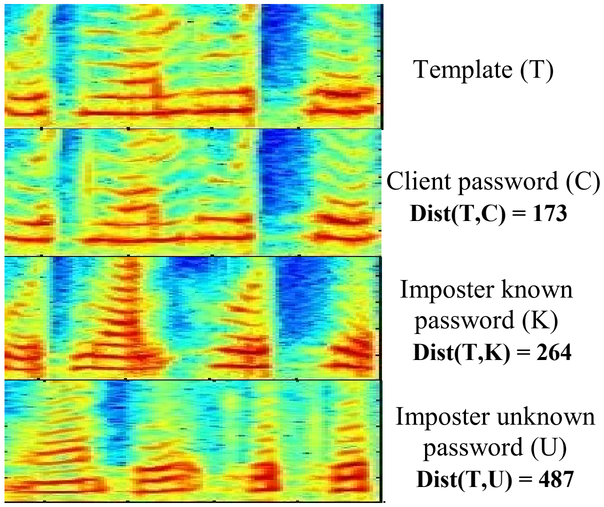


Figure 6: VSF(spectrograms) of various passwords. The marked one is used as template(T). Rest of them are resized & compared with this one (T).

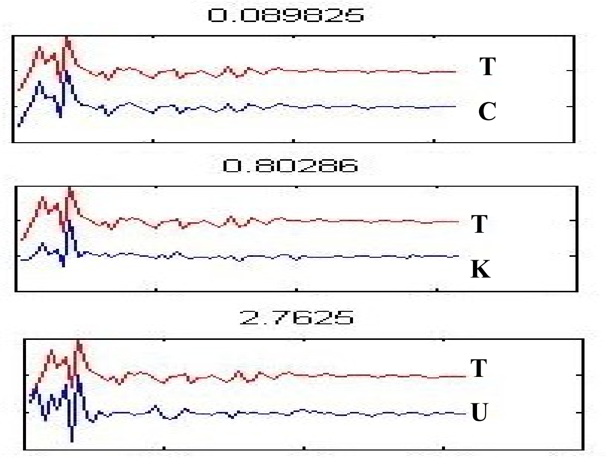


Figure 7: CFDs of the VSF(spectrograms) of Figure 6, showing the comparisons and distances from the template CFD (T) for client (C), known-password imposter (K) and unknown-password-imposter (U).

The CFD representation of the VSF does not lower the discrimination power as can be seen in Figure 6 and Figure 7 which show various VSFs (spectrograms) and the corresponding CFDs. To summarize, the proposed VSF-CFD feature extraction method for the spoken password offers the following advantages: a) speaker-identity is well preserved in a fixed-dimension CFD vector, b) compact representation: a 3 second password or 24000 numbers (for 8 KHz sampling rate) is represented by only K numbers (K, the dimension of CFD is typically chosen to be 63), c) fixed-dimension representation makes comparison of two variable-length spoken-passwords quite easy, and d) compact representation makes the storage of multiple password templates possible even if the number of users is large.

4. Multiple Pose and Transformed Face Profiles as Face Features

The AV method proposed here is being built for a kiosk/workstation based highly-secure user-authentication system. Therefore, camera-position, lighting, background noise, all of these will be in a controlled setting. The present set-up, database used for the simulation results presented here are all based on a laptop with an integrated web-camera.

In our proposed AV user authentication framework, multiple face profiles of each user are captured by the integrated web-cam, as the user follows a moving ball which goes to different locations on the screen (Figure 8). The same movement of the ball happens during training and testing/actual-usage. This way, during enrollment, our system collects a “bag-of-poses” or profiles for each user. Figure 8 shows the left, right and central profiles generated by our user interface.

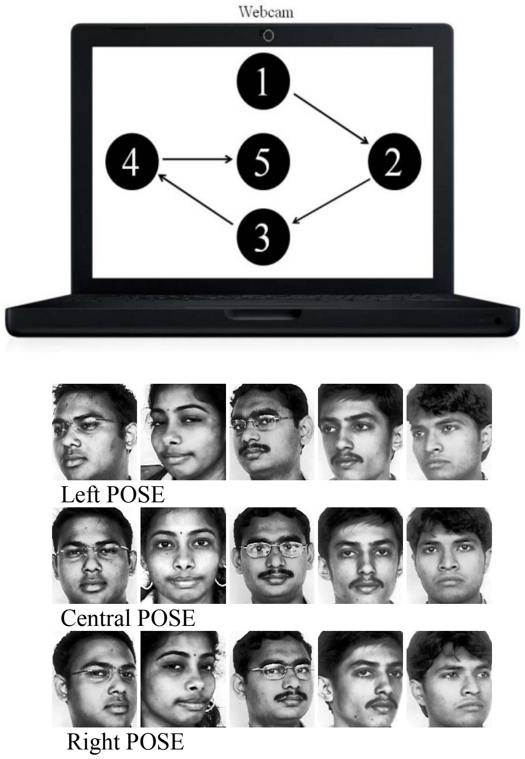


Figure 8: Capture of various face profiles by a single camera

The face features are extracted in a manner similar to the speech CFD. From each face profile image, face detection [12] is performed and histogram-equalization is done on the cut gray image. A set of selected DCT coefficients (as shown in Figure 5) then form the Transformed Face Profile (TFP) signature. The user can then be asked a set of questions and the answers are recorded as his/her spoken-passwords. Note that, each media sample (a face profile image or a spoken password) in the proposed method is now creating a fixed-dimension signature vector (CFD or TFP) capturing the user’s identity.

5. Multiple Nearest Neighbor Classifier

The multiple nearest neighbor classifier (MNNC) framework (Figure 9) proposed here offers an effective way to integrate the information from the multiple media samples. MNNC combines multiple Nearest-Neighbor-Classifiers (NNC), one per media sample. Each NNC has a set of codebooks, one for each user. During training, we extract T TFP’s/CFD’s from the T training-images/spoken-passwords for each user. For each i^{th} media sample, we use a dedicated NNC slice $\mathbf{NNC_i}$ which creates a score vector $\mathbf{R_i}$. Thus, if we are using N_S spoken passwords and a total of N_F x N_{FI} face images (N_F: number of face-profiles; N_{FI}: number of face-images/profile), then we will have L= $N_S + N_F$ x N_{FI} dedicated NNC slices generating L intermediate score vectors, $\mathbf{R_i}$ **;** i=1,2,3…L. A proper fusion method can then be used to combine these intermediate scores to create a final score vector $\mathbf{R_{final}}$. A suitable decision mechanism can be applied for the identification or authentication tasks.

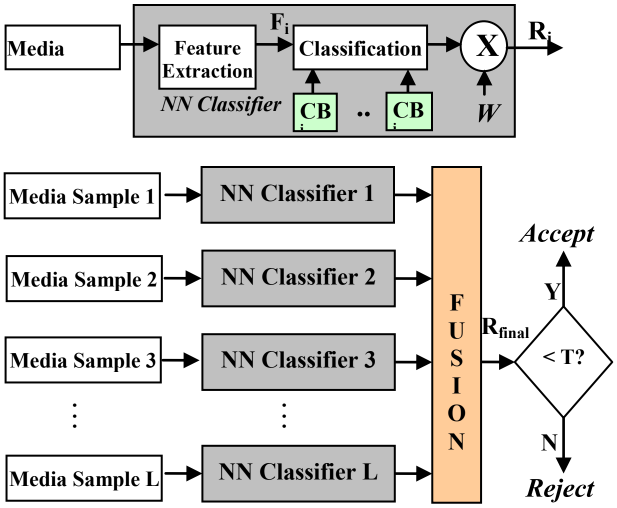


Figure 9: Architecture of the MNNC framework

We present next the details of a person authentication task (the identification task will be similar).

During test, a set of L media samples are presented (Figure 7), along with an identity claim, 'k'. We need to verify whether the feature set [F_1 F_2 F_3… F_L] belongs to person P_k or not.

Let us consider the i^{th} NNC slice, which has N codebooks CB^i_k ; k=1,2,...,N, for the N users enrolled to the system. Each CB^i_k has T codevectors, $CB^i_k = [C^i_{km}]$, m=1,2,..,T.

Given a feature vector F_i and a claim "k", we find the score R_i as follows:

Step1: Given the identity claim, k, compute two distances, D_{true} & D_{imp}, as follows:

D_{true} = minimum distance of $\mathbf{F_i}$ from the codebook $\mathbf{CB_k}$ of claimed person P_k, Dtrue = min $\{D_{km}\}$, where $D_{km} = \| \mathbf{F_i} - \mathbf{C^i_{km}} \|^2$, m=1,2,…,T

D_{imp} = minimum distance of $\mathbf{F_i}$ from the set of codebooks of all other persons except person P_k or D_{imp} = min $\{D_{im}\}$, where $D_{im} = \| \mathbf{F_i} - \mathbf{C_{nm}} \|^2$, m=1,2,,T; n=1,2,..N & n not equal to k.

Step2: Compute the interim score for i^{th} NNC slice as $\mathbf{R_i} = D_{true} / D_{imp}$

We compute this ratio R_i for each of the L features $\mathbf{F_i}$; **i**=1,2,..,L and then fuse them.

6. Database and Experimental Set Up

For this research, we collected a 210-user audio-visual biometric database in an office environment under controlled lighting. Our aim was to create the same environment as that of our target application of secure user authentication in a fixed office location.

We could not use any publicly-available AV biometric databases as none of them met our criteria of having multiple face profiles and multiple unique passwords per user. A password is a set of 4 words chosen by the user and spoken in the native language of the user. Some users selected 4-digit words as passwords. Thus, the passwords in our AV database are unique and quite different from each other. This database is available for research purpose. If there is an interest, the author may be contacted.

The following parameters constitute the system-parameter-set for the proposed AV person recognition framework: a) number of passwords (N_S) b) the number of face profiles (N_F), c) the number of face images per profile (N_{FI}), d) the dimensions of CFD and TFP, and e) the number of training samples T. These parameters need to be judiciously selected to balance performance and complexity.

Table 1 shows the identification accuracy as a function of TFP and CFD dimensions. Table 2 shows variation of face-based identification performance with respect to number of pose and number of images/pose.

Table 1: Impact of TFP/CFD on Identification Error

M	Image SID %Error	Speech SID %Error
15	5.82	2.45
63	1.57	0.40
143	1.89	0.19

Table 2: Variation of Performance with Number of Poses (N_F) and Number of Images used per Pose (N_{FI})

N_F	N_{FI}	ID %Error
1	1	4.26
3	6	0.00

Based on our evaluation (results of Table 1 and 2) we have selected the following system parameters: a) codebook size T=5 templates per user for speech and T=6x3 training images for face, b) spectrogram-CFD and a dimension of 63 for speech, c) image TFP dimension of 143 and 3 poses and 6 images per pose.

For performance measure we used the following metrics. For identification, we used the minimum score to identify the person and used “percentage-error” as the performance metric. For verification, we simulated two cases: a) unknown-speech password when the imposter does not know client-password and b) known speech-password – where an imposter overheard and uttering the actual password of the client. We used EER (Equal-Error-Rate), FAR (False-Acceptance Rate) and FRR (False Rejection Rate) as performance metrics. When the EER is zero (meaning that the target and imposter score distributions are separable) we also present the distance-of-separation (**D_{sep}** in Table 3) between the two distributions.

For testing, we used 5x1 speech samples per person and 6x3 sets of images per person and then combined these passwords and images to create 5x1 AV-trials per person for the AV experiments. We have 900 cases of known-password and 5x210=1050 unknown-password cases for verification. This makes a total of 5x210=1050 AV identification trials. For known-password-verification, we used 900 target and 900 imposter trials or a total of 1800 trials. For unknown-password-verification, we used 1050 target and 1050 imposter or a total of 2100 trials. We also did speech-only and face-only trials. For speech-only trials, there were 1050 identification trials and 1800 known-password and 2100 unknown-password verification trials. For face-only trials, there were 210 identification trials.

7. Results and Discussions

Table 3 presents the result of the combined AV person recognition method. Note that: a) the multiple-pose multiple image per pose face-only method itself is quite powerful creating separable client and imposter distributions and 0% identification and verification errors, b) when both speech and face is combined, the results are error-free, but the distance between the client and imposter distributions (Dsep) has increased, and c) the combined AV method successfully defends against imposter attacks even when the imposter knew the client passwords.

For a complexity analysis, let us compare the proposed AV person recognition method with a hypothetical AV biometric system using PCA for face and a high-performance text-dependent DTW based system (as in [18]) for speech. Table 4 shows the computational complexity (in terms of multiple-add per test-trial) and storage requirements (in terms of per test-trial) and storage requirements (in terms of number of floating-point numbers to be stored per user). As seen here, our proposed method is much simpler in both computational and storage complexity measures.

Table 3: Performance of the proposed AV Person Recognition System

Modality	ID (%)	EER (%)	FRR (%)	FAR (%)	D_{sep}
Face Image only	0	0	0	0	586
Speech only*	2.1	1.68	1.12	1.68	-
Face Img +Speech (Unknown-pwd)	0	0	0	0	1561
Face Img +Speech (Known-pwd)	0	0	0	0	648

*Note: For speech only trial, the verification result shown is for known-password case.

Table 4: Complexity Comparison (storage of training templates/user and per trial computational complexity)

Method	Computation Complexity	Storage Complexity
DTW+PCA	10^8	O(39000)
Proposed Method	O(2200)	O(2000)

Note: Assumptions: PCA+DTW: 5-template DTW; 4 second test/training utterance; 20 ms speech frame; 39 dimension MFCC used as speech feature; 40x40 size image; PCA dimension = 60;

Proposed method: TFP&CFD dimension=63; 3 passwords; 5 speech samples/user for training; 6 images x 3 pose for training; test: 6x3=18 images/test-trial; 3x1 spoken-passwords/test-trial;

8. Conclusions and Future Work

We proposed an AV user-authentication method suitable for on-line access control which offers high performance and strong resilience to imposter attacks while operating at significantly low complexity compared to conventional AV methods. The proposed method harnesses its power from strong person-identity traits extracted from multiple face profiles and a novel set of visualized speech features. A number of novel feature extraction methods are introduced which provide high discrimination power, allowing the classification part to remain much simpler than conventional high-performance AV methods.

As far as the weaknesses of the proposed method are concerned, we noted the following: a) good end-pointing of the password (silence separation) and good voiced-unvoiced segmentation are crucial (we do have robust algorithms for both), b) proper scaling and cutting of the face is important (we fixed this in our real-time system by an “adjustment” mode, when the user is asked to sit properly in front of the Laptop and adjust his/her position, so that the cut face fits a presized rectangle). Other factors like background noise and illumination-variation are controlled by a fixed lighting and reasonable office condition set-up.

We are working on further enhancements in several areas such as newer features and fusion methods, and robustness to practical challenges of variations of background noise, illumination, etc, found in real-life situations.

9. References

[1] P.S. Aleksic and A. K. Katsaggelos, “Audio-Visual Biometrics”, *Proc. IEEE*, vol. 94(11), pp. 2025-2044, Nov 2006.

[2] C. C. Chibelushi, F. Deravi and J. S. D. Mason, “A Review of Speech Based Bimodal Recognition”, *IEEE Trans. on Multimedia*, vol. 4(1), pp. 23-37, Mar 2002.

[3] A. Kanak, E. Erzin, Y. Yemez and A.M. Tekalp, “Joint Audio-Video Processing for Biometric Speaker Identification”, *Proc. ICASSP-03*, vol. 3(6-9), pp. 561-564, July 2003.

[4] S. Marcel, J. Mariethoz, Y. Rodriguez and F. Cardinaux, “Bi-Modal Face & Speech Authentication: A BioLogin Demonstration System”, *Proc MMUA-06*, May 2006.

[5] S. Ben-Yacoub, Y. Abdeljaoued and E. Mayoraz, “Fusion of Face and Speech Data for Person identity Verification”, *IEEE Trans. on Neural Networks*, vol.10(5), pp.1065-1074, Sep 1999.

[6] Z. Wu, L. Cai and H. Meng, “Multi-Level Fusion of Audio and Visual Features for Speaker Identification”, *Proc. ICB’06*, pp. 493-499, 2006.

[7] A. Das and P. Ghosh, “Audio-Visual Biometric Recognition by Vector Quantization”, *Proc.IEEE SLT-06*, pp. 166-169, 2006.

[8] W. Zhao, R. Chellappa, P. J. Phillips and A. Rosenfeld, “Face Recognition: A Literature Survey”, *ACM Computing Surveys*, vol. 35(4), pp. 399-458, Dec 2003.

[9] S. Zhou and V. Krueger, "Probabilistic Recognition of Human Faces from Video", *Computer Vision and Image Understanding*, vol. 91, pp. 214-245, 2003.

[10] K.C. Lee, J. Ho, M.H. Yang and D. Kriegman, "Video-based Face Recognition using Probabilistic Appearance Manifolds”, *Proc. CVPR-03*, vol. 1(18-20), pp. 313-320, June 2003.

[11] A. Pentland, B. Moghaddam and T. Starner, “Face Recognition using view-based and Modular Eigenspaces”, *Proc. SPIE’94 - Automatic Systems for Identification and Inspection of Humans*, vol. 2277, pp/ 12-21, Oct. 1994.

[12] P. Viola and M. Jones, ”Rapid Object Detection Using a Boosted Cascade of Simple Features”, *Proc. CVPR’01*, vol. 1, pp. 511-518, April 2001.

[13] Z. Biuk and S. Loncaric, “Face Recognition from Multi-Pose Image Sequence”, In *Proc. 2nd Intl. Symp. Image and Signal Processing,* pp. 319-324, 2001.

[14] V. Krueger and S. Zhou, “Exemplar-based face recognition from video”, In *Proc. ECCV-Part IV LNCS*, vol. 2353, pp. 732-746, 2002.

[15] S. Gong, A. Psarrou, I. Katsoulis and P. Palavouzis, “Tracking and Recognition of Face Sequences”, In *Proc. European Workshop on Combined Real and Synthetic Image Processing for Broadcast and Video Production*, pp. 97-112, Nov 1994.

[16] F. Bimbot, J.F. Bonastre, C. Fredouille, G. Gravier, I.M. Chagnolleau, S. Meignier, T. Merlin, J.O. Garcia, D.P. Delacretaz, D.A. Reynolds, “A Tutorial on Text-Independent Speaker Verification”, *Eurasip J. Appl. Speech Proc.*, vol. 2004(1), pp. 430-451, Jan. 2004.

[17] D. Falavigna, “Comparison of Different HMM Based Methods for Speaker Verification”, *EUROSPEECH-95*, pp. 371-374, Sept. 1995.

[18] V. Ram, A. Das, and V. Kumar, “Text-dependent Speaker-recognition using one-pass Dynamic Programming”, *Proc. ICASSP’06*, pp. 901-904, 2006.

[19] D.A. Reynolds, T.F. Quatieri and R.B. Dunn, “Speaker Verification using Adapted Gaussian Mixture Models”, *Digital Signal Processing*, vol. 10, pp. 19-41, 2000.

[20] T. Kinnunen, E. Karpov and P. Franti, “Real-Time Speaker Identification and Verification”, *IEEE Trans. ASLP*, vol. 14(1), pp. 277-288, Jan 2006.

[21] K. Rao and P. Yip, *Discrete Cosine Transform: Algorithms, Advantages, Applications*. AP. 1990.

[22] R. Keys, “Cubic convolution interpolation for digital image processing”, *IEEE Trans. on Signal Processing, Acoustics, Speech, and Signal Processing*, vol. 29(6), pp. 1153-1160, Dec 1981.