J Sign Process Syst (2017) 88:259–271
DOI 10.1007/s11265-016-1148-z

Development of Multi-Level Speech based Person Authentication System

Rohan Kumar Das[1] · Sarfaraz Jelil[1] · S. R. Mahadeva Prasanna[1]

Received: 20 April 2015 / Revised: 19 May 2016 / Accepted: 24 May 2016 / Published online: 4 June 2016

Abstract This work presents the development of a multi-level speech based person authentication system with attendance as an application. The multi-level system consists of three different modules of speaker verification, namely voice-password, text-dependent and text-independent speaker verification. The three speaker verification modules are combined in a sequential manner to develop a multi-level framework which is ported over a telephone network through interactive voice response (IVR) system for aiding remote authentication. The users call from a fixed set of mobile handsets to verify their claim against their respective models, which is then authenticated in a multi-level mode using the above stated three modules. An analysis over a period of two months is shown on the performance of the multi-level system in attendance marking. The multi-level framework having combination of the three modules helps in achieving better performance than that of the individual modules, which shows its potential for practical deployment.

Keywords Speaker verification · Speech biometrics · Voice-password · Text-dependent · Text-independent

This work is supported by the project grant no. 12(6)/2012-ESD from the e-security division of Department of Electronics & Information Technology (DeitY), Govt. of India

✉ Rohan Kumar Das
rohankd@iitg.ernet.in

[1] Department of Electronics and Electrical Engineering, Indian Institute of Technology Guwahati, Guwahati, -781039, India

1 Introduction

The advancements in technology has lead us to use biometric technologies for recognizing a person for various applications [4, 20, 23]. Different types of human traits are used as biometric features, by which the unauthorized users are restricted from using an intended service. Face, fingerprint, iris, retina and DNA are the most often used biometric features which have got success in practical system deployment [19, 24, 27]. When speech is used as a biometric feature to verify a person, the task of speaker recognition comes into picture [8]. Speaker recognition can be broadly classified into two, namely speaker identification and speaker verification (SV). The former attempts to find out which one among a registered set of speakers best matches the speaker currently trying to be identified. This is found to be useful mostly in forensic applications, where given a test example its match to the registered speakers is computed. On the other hand, SV uses test speech with a claim to verify against the claimed model, which is of our interest and has significance towards a practical application oriented system. SV can be broadly categorized as text-dependent (TD) and text-independent (TI) SV depending on the content of the speech data used for the task [14, 25]. The TD SV requires the same set of text to be spoken during training as well as testing. The typical duration of sentence involved in this case is 2-3 seconds. As the amount of train and test data is less in this type of SV, it is more of comfort in a practical system implementation for the users. However if the training data is not proper, it may lead to poor performance. In the case of TI SV, it does not have any restriction over training and test data. User is not bound for any kind of fixed set of text for verifying his/her

identity claim, thus it relaxes the user from any bounding. In this kind of SV, typically 2-3 minutes of speech data is required for training and 20-30 seconds for testing. Speech biometric system using TI framework for remote person authentication has already demonstrated its significance [7, 9, 11, 12, 18]. For practical system based application this kind of SV has potential for short utterance case so that the duration does not become a hurdle on the way for implementation.

This work is motivated by the different modules of SV systems and an attempt to combine them into a common framework for practical implementation along with easement of the end users. The proposed multi-level SV system is a combination of three different SV systems. The first system, voice-password (VP) is inspired from TD based SV, where user specific fixed text is used for training as well as testing. As this system involves a speaker-specific phrase, it helps in reducing the spoofing attacks. Then the second system of TD based SV requires all the speakers to utter the same set of text during train and test sessions. Both of these two modules use commonly used acoustic feature *mel frequency cepstral coefficients* (MFCC) as basic features and *dynamic time warping* (DTW) algorithm to find a match between the train and test template [13]. The third system of the multi-level SV system is the TI SV module, where *i-vector* based modeling is used for verification of a claim [15]. This kind of modeling is helpful for compact representation, easy channel/session compensation which makes it feasible for deployment in a practical scenario. These three SV systems are fit into a common framework in a sequential manner and a callflow for SV system is designed over telephone network. The users are guided through an interactive voice response system (IVRS) for verifying their identity claim with attendance as an application. The performance of the system is evaluated for a period of two months, which showed its possibility towards practical system based implementation. The prime contribution of this work is to combine three different SV categories VP, TD and TI into a common framework for person authentication in a deployable scenario. Also the work addresses various issues associated with practical deployment setting, such as expecting proper phrase for TD framework, sufficient data from the speakers for each module, decision on the fly etc.

The rest of the paper in organized as follows: Section 2 presents an overview of the multi-level speech biometric based SV system. Section 3 describes the functionality and development of the three different modules VP, TD and TI SV. In Section 4, the development of multi-level SV system using stated three modules of SV is described. Section 5 shows the experimental results and analysis of the speech based multi-level biometric system. The summary and conclusion are given in Section 6.

2 Overview of Multi-Level Speech based Person Authentication System

Biometric systems can be classified into two broad categories which can be seen as unimodal and multimodal. The former refers to the systems that consider single biometric source for the task of authentication and the latter deals with involvement of multiple biometric source for establishing identity [6]. Thus, the proposed multi-level framework is a unimodal system that uses speech as a biometric source input. The speech signal is collected for three different categories of SV namely VP, TD and TI for verifying an identity claim of a person.

The systems can operate in different modes in combination of different measures or modeling approaches using the same biometric attribute. The combination can be done in serial, parallel, pipelining, hierarchical or sequential approach for system development [23]. In serial mode combination, the output of one modality goes to the next module. For parallel mode combination, the output of multiple modalities are used simultaneously for concluding into a final decision. The hierarchical mode uses a combination of both serial and parallel systems. Whereas in pipelining mode of operation, the advantage is obtained for the case of single sensor and feature extraction technique. During the time of feature extraction of the first modality, the input for the second modality is collected from sensor and so on, so that it becomes advantageous by pipelining. The sequential approach for combination is such that, if one modality rejects then it is checked by subsequent modalities and so on for concluding into a decision. This kind of combinational approach is useful to take advantage of each modality with optimum time complexity, which is required for deployment in a practical scenario using the the three different modules of SV for the proposed multi-level framework.

The three different modules of SV are fit into a common framework in sequential manner for the development of multi-level speech based authentication system. Thus, the proposed system is termed as multi-level in a sense that the decision is taken at each level in an online framework, which is combined in a sequential manner. The three different modules of multi-level framework depict three different SV categories in terms of difference in speech input, modeling and decision logic for verifying a claim. The VP module provides relaxation to the user to choose a phrase of own choice, which is easily remembered by him/her. This is unique for all the users in the system, hence it works as a passkey for each user with their voice as a biometric measure. The TD module uses a global phrase that is to be spoken by all the users during training as well as testing. As these phrases are of very short duration, the users are not burdened by uttering it for testing. For TI SV module, there is no restriction on the content of speech to be

spoken. This module provides freedom to the user to speak anything of his/her own choice. Thus the advantage of each of these modules are used for development of multi-level framework for speech biometric system for practical implementation. In this work, the multi-level speech based person authentication system is deployed for student attendance as an application. The details of each of the modules used in the multi-level SV system is described in the next section.

3 Modules of Multi-Level SV System

This section describes the three different modules of SV and their functionalities that are used in multi-level SV system. Each of these modules has its own uniqueness and different methodologies for validating a claim. The modules VP, TD and TI are designed in a way to fit to the multi-level framework with student attendance as an application.

3.1 VP SV Module

The VP based SV, which is inspired from the TD based SV provides a flexibility to the user to choose a phrase which is to be repeated during training and testing. Therefore, the length of the VP becomes significant and typically it should consist of 5-8 words so that it can capture about 4-6 seconds of speaker's speech. For better modeling of the speaker characteristics, multiple instances of the same the phrase are taken from the user for training.

3.1.1 Front-end Processing

The train and the test utterances are short term processed with a frame size of 20 ms with a shift of 10 ms. 39-dimensional MFCC (13-base + 13-Δ + 13-$\Delta\Delta$) features of those utterances are extracted for each Hamming windowed frame with 22 logarithmically spaced filters. Energy based end point detection is performed for detecting the begin and end points of the utterances. The average energy of the speech signal is calculated and then a threshold of 6 % of it is taken, which is then compared to each frame of the utterance. The begin point is marked when four consecutive frames have higher energy than the threshold. Similarly for detecting the end point, the signal is processed from the other side and similar decision logic is employed. The MFCC features within the begin and end points are taken, upon which cepstral mean variance normalization (CMVN) is performed [22]. The normalized features of train utterances are kept as reference template for each speaker.

3.1.2 Template Matching and Decision

The test speech features are compared with reference to the claimed model reference template by the DTW algorithm, which calculates the accumulated distance score between train template X and test template Y of different lengths by taking a warping path as,

$$d_\phi(X,Y) = \sum_{k=1}^{T} d(\phi_x(k), \phi_y(k)) m(k)/M_\phi \tag{1}$$

where, $d_\phi(X,Y)$ is the accumulated distance, $d(\phi_x(k), \phi_y(k))$ is short time spectral distortion, $m(k)$ is a non-negative path weighting coefficient and M_ϕ is the path normalizing factor. As VP based SV system involves speaker-specific fixed phrase, the distance score is compared to a speaker-specific threshold for taking decision with respect to a claim. This speaker-specific threshold is obtained from few test trials made against the model by considering the mean and standard deviation of the distance scores. Fig. 1 shows the structure of the VP based SV system.

The VP module of multi-level SV for student attendance considers *roll number and name* as the speaker-specific phrase as it is unique for each student. The roll number of the students consists of nine digits and instructions

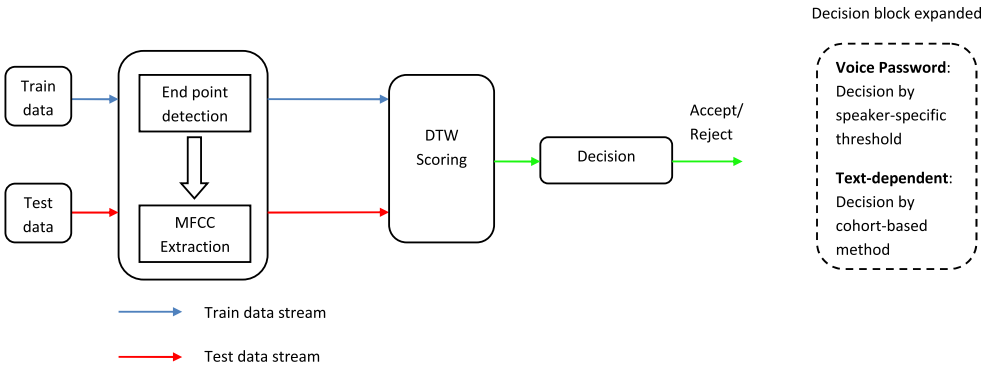

Figure 1 Structure of VP and TD based SV modules of multi-level framework.

are given for expanding the initials of their names. Thus, the roll number and name together capture about 4-6 seconds of data that is used for modeling each speaker. Three instances of the VP are recorded per user for speaker modeling. To fix the speaker-specific threshold, 10 thresholding trials are made by the users against respective model. The mean (μ) and standard (σ) deviation from the distance scores of thresholding sessions are taken to fix a speaker-specific threshold of ($\mu + 2\sigma$) for each speaker. In testing phase, the test speech is compared to the three training templates of the claimed speaker using DTW and then average of three distance scores is taken, which is compared to the speaker-specific threshold of the respective speaker for decision.

3.2 TD SV Module

The TD module of the multi-level framework deals with repetition of the same phrase of 3-4 seconds of duration during training and testing by all the users. This module is different from the earlier module of VP in the sense that the phrase is common across all the users in case of TD, whereas in VP based framework the phrases are speaker-specific. Thus the VP based module gives freedom to the user for choosing his/her own phrase, that he/she has to remember and produce the same during the testing sessions. As the phrase is speaker-specific in VP, it is more robust to the impostors unlike TD being more susceptible to the impostors. However, as the user is allowed to choose own phrase, the phrase may not be phonetically balanced. On the other hand, the TD based framework deals with repetition of prompted phrase, that are designed in a phonetically balanced manner inspired from the phrases of TIMIT database [26]. Thus, it can be seen that both the modules VP and TD though have some similarities, they are mainly different in terms of the way in which the speech data is collected from the users. The former deals with speaker-specific phrase and the latter deals with prompted phrase that is global across speakers, which makes to exploit different speaker characteristics and also to take advantage from the two categories in the multi-level framework.

3.2.1 Front-end Processing

The prompted phrase repeated during the train and the test sessions of TD module are processed in a similar manner to that of the VP module at the front-end. The train and the test utterances are considered in terms of Hamming windowed frame of 20 ms with a frame shift of 10 ms. The energy based end point detection is used to select the speech portion and 39-dimensional MFCC features are taken. These features are further normalized by CMVN approach to nullify the offset values [22]. The normalized features of the train and the test utterances are further considered for template matching for verification of a claim.

3.2.2 Template Matching and Decision

The test speech data is matched to the claimed model using DTW algorithm as given by Eq. 1 to give the distance score. As the fixed phrase is global for all the speakers, the decision is taken with respect to a set of four cohort speakers by comparing the distance score against them. The cohort speakers for each user are generated randomly from the enrolled set of speakers excluding the corresponding speaker and kept fixed for decision logic. The basic structure of the TD SV is shown in Fig. 1.

Three different sentences are considered for the TD module of the multi-level framework, which are:

– "*Don't ask me to walk like that*" (TD1 -prompt)
– "*Get into the hole of tunnels*" (TD2 -prompt)
– "*Lovely picture can only be drawn*" (TD3 -prompt)

Three instances each of these sentences are recorded for creating the speaker templates. During testing, one out of these three sentences is asked to the user through the IVR system callflow. This is done to check the reality of the user to some extent. The test speech features are compared to the respective claimed speaker for three instances of the specific sentence type using DTW algorithm to yield the distance score. This is compared to scores obtained from the set of four cohort speakers of the claimed speaker to accept/reject a claim.

3.3 TI SV Module

The TI module of the multi-level system is designed from a standard text-independent SV framework, that is quite different from VP and TD module. The processing and methodologies employed for verification of a trial is therefore different than those explained for the other two modules, which can be seen from the following descriptions.

3.3.1 Front-end Processing

The basic set of features are same like the earlier two modules VP and TD of the multi-level framework. The short term processing is made on the train as well as the test utterances in a similar way to extract 39-dimensional MFCC features. In this module feature selection technique employed is different from the former two modules. Here energy based *voice activity detection* (VAD) is done to perform speech/non-speech detection and the speech regions are only retained for considering the features from only those portions. The features from the speech regions are then normalized using cepstral mean subtraction

(CMS) followed by cepstral variance normalization (CVN) techniques [22].

3.3.2 Speaker Modeling and Decision

The i-vector based speaker modeling technique is used for modeling the speakers in the TI module of the multi-level framework [15]. Fig. 2 shows the TI framework using this technique. In this approach, Gaussian mixture model (GMM) mean supervectors of each utterance is represented by a low dimensional representation called i-vector [10]. It is done using a transformation matrix (T-matrix) that contains all the variabilities such as channel, session etc., which is trained using a set of development data. For an utterance that has GMM mean supervector M, the T-matrix T and universal background model (UBM) mean supervector m can be used to obtain the i-vector w as follows,

$$M = m + Tw \tag{2}$$

For an UBM having a weighted sum of C component Gaussian densities as $U = \{\mu_c, \Sigma_c, \eta_c\},\ c = 1, 2, ..., C$, where η_c, μ_c and Σ_c are the weight, mean vector and covariance matrix associated with mixture c, respectively and a sequence of L speech feature vectors $\{x_1, x_2, \ldots, x_L\}$ of dimension F, the 0^{th} order (N_c) and the centralized 1^{st} order (F_c) Baum-Welch statistics of the speech frames on the c^{th} component of the UBM are given by,

$$N_c = \sum_{t=1}^{L} P(c|x_t, U) \tag{3}$$

$$F_c = \sum_{t=1}^{L} P(c|x_t, U)(x_t - \mu_c) \tag{4}$$

where, $c = 1, 2, \ldots, C$ is the component index in the UBM, $P(c|x_t, U)$ is the posterior probability of the mixture component c generating the feature vector x_t and μ_c is the mean of UBM component c.

The total variability matrix T is learned from Baum-Welch statistics of the large amount of development data, computed using the UBM to capture different variabilities. For a given T, the estimated i-vector $\hat{w}$ is computed as,

$$\hat{w} = (I + T'\Sigma^{-1}N(u)T)^{-1}T'\Sigma^{-1}F(u) \tag{5}$$

where, $N(u)$ and Σ are diagonal matrix of dimension $CF \times CF$ whose diagonal blocks are $N_c I$ and Σ_c, respectively. $F(u)$ is a supervector of dimension $CF \times 1$ generated by concatenating all 1^{st} order Baum-Welch statistics (F_c) for a given utterance u.

Linear discriminant analysis (LDA) and *within class covariance normalization* (WCCN) are applied on i-vectors for channel and session compensation [3, 17]. LDA projects the feature vectors to a set of new orthogonal axes, where the intra-class variance caused by the channel is minimized and inter-class variance is maximized. The projection matrix is composed of the eigen vectors corresponding to the best eigen values of the eigen analysis equation as,

$$(W_c^{-1}B_c)v = \lambda v \tag{6}$$

where, W_c is the within-class covariance matrix, B_c is the between-class covariance matrix, v is an arbitrary vector, and λ is the diagonal matrix of the eigen values.

WCCN defines a set of upper bounds on the classification error metric to lower the error rate. A transformation matrix is used, by which the feature vectors are transformed to minimize the upper bounds on the classification error metric, which in turn minimizes the classification error. The transformation matrix B is obtained by Cholesky decomposition of the inverse of the within-class covariance matrix W as, $W^{-1} = BB^t$. As suggested in [15], when LDA is followed by WCCN better results are obtained, therefore W is calculated in the projected space of the LDA.

Figure 2 shows the structure of i-vector based TI SV module for the multi-level framework. The train and the test data undergo the same set of procedure as mentioned above to generate train and test i-vectors, respectively. For a pair of train and test i-vectors given by $\hat{w_{trn}}$ and $\hat{w_{tst}}$, the

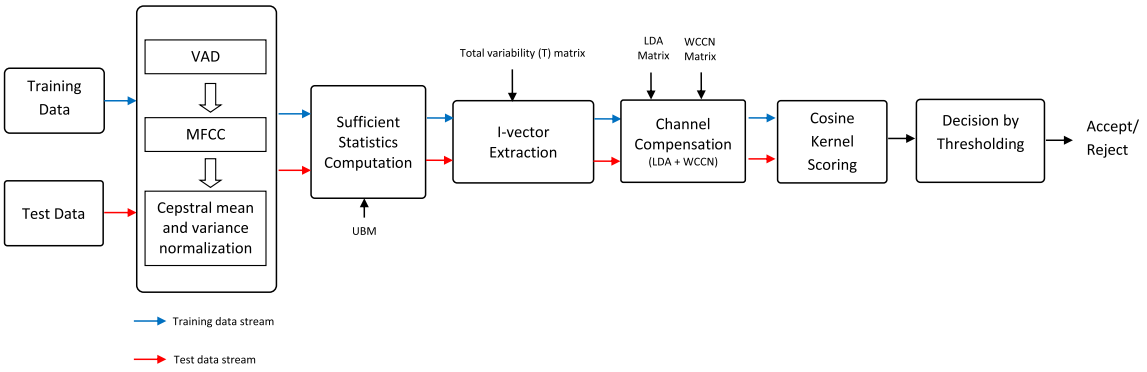

Figure 2 Structure of TI module based on i-vector framework of the multi-level system [21].

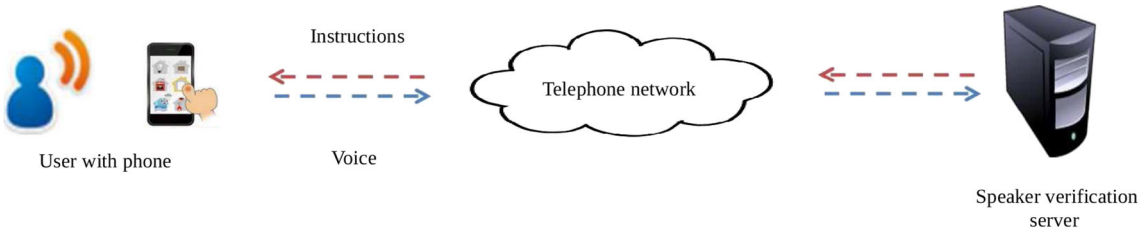

Figure 3 Multi-level SV system over telephone network.

verification of a claim is performed by computing the cosine kernel score between these two i-vectors as follows:

$$\frac{< \hat{w}_{trn}, \hat{w}_{tst} >}{||\hat{w}_{trn}||\ ||\hat{w}_{tst}||} \lessgtr \theta \ \text{(Threshold)} \tag{7}$$

The development data required for learning UBM, T-matrix, LDA and WCCN is taken from telephone channel data of NIST speaker recognition evaluation (SRE) 2010 database [1, 10]. The front-end analysis of the development data is also made in a similar way as it is done for train and test data. A subpart of the development data of about 90 hours is taken to build a gender independent UBM of 1024 components considering equal amount of male and female speech. Then a T-matrix of 400 columns, 250 dimensional LDA and full dimensional WCCN matrix are built using the development data.

During the training phase of TI SV module, 3 minutes of speech data is asked from the user which is then merged with the VP and TD data. This is used for modeling the speaker using i-vector modeling approach to give the train i-vector. In the testing phase, the speakers are asked VP and TD data for respective modules, which is combined together to treat as TI test speech to give the test i-vectors. Finally after performing the channel/session compensation by applying LDA and WCCN, cosine kernel between the test i-vector with respect to the claimed i-vector is taken. The resultant score is normalized by *test normalization* (t-norm) technique and then compared to a global threshold to accept/reject identity of a claim [16]. The global threshold is obtained by performing offline experiments on 30 speakers data collected in the same scenario to give minimum error for genuine and impostor trials.

4 Deployment of Multi-level SV Framework

This section describes the development of the multi-level framework using the three different modules of VP, TD and TI SV, the details of which are explained in the previous section. The three modules are combined to a common platform over a IVR system callflow through ISDN-PRI (integrated services digital network-primary rate interface) line that can handle telephone channel calls through computer telephone interface (CTI) card. It can handle up to 30 parallel calls over the telephone channel. The IVR is hosted on a voice-server which runs the Asterisk software. It is a software implementation of a telephone private branch exchange (PBX), that allows the server handle incoming calls and make outgoing calls from and to other public switched telephone network (PSTN) or voice over internet protocol (VoIP) services. This callflow is made for implementation of the attendance system of postgraduate students of our department at our institute. Students have to call to the toll-free number through which the IVR system callflow helps for enrollment as well as regular testings in a practical setting. Fig. 3 shows the overview of the multi-level SV system over telephone network which is developed for our speech biometric based attendance system. [1]

4.1 Enrollment Phase

The users can enroll into the multi-level system by calling to the server over the telephone network having IVR system. Fig. 4 shows the flowchart for enrolling into the multi-level system. The IVR system callflow guides a user for enrollment phase while the user calls to the toll-free number and selects the option of enrollment. It first generates a four digit unique speaker ID for the user, that works as personalized ID for a user. The user has to remember this ID as it is required during testing for claiming against that speaker model. The user is asked for his/her *roll number and name* together to be spoken as an input to the VP module. Then three sets of text, as mentioned in Subsection 3.2 are asked, that have to be repeated one at a time as the input for the TD module. The inputs for VP and TD module are asked to repeat for three times to have multiple reference templates for each user for these two modules. Finally, a 3 minutes of read text is asked from the user, which is considered as input to the TI module. There is a quality check at each of the modules of the multi-level framework to qualify the speech given for training the speaker models. This quality check refers to evaluating the input speech; whether sufficient amount of speaker's speech has been spoken for each of the module. The threshold for quality check in terms of

[1] An initial version of this work with text-independent speaker verification system framework is presented at the National Conference on Communication in February 2014. [21]

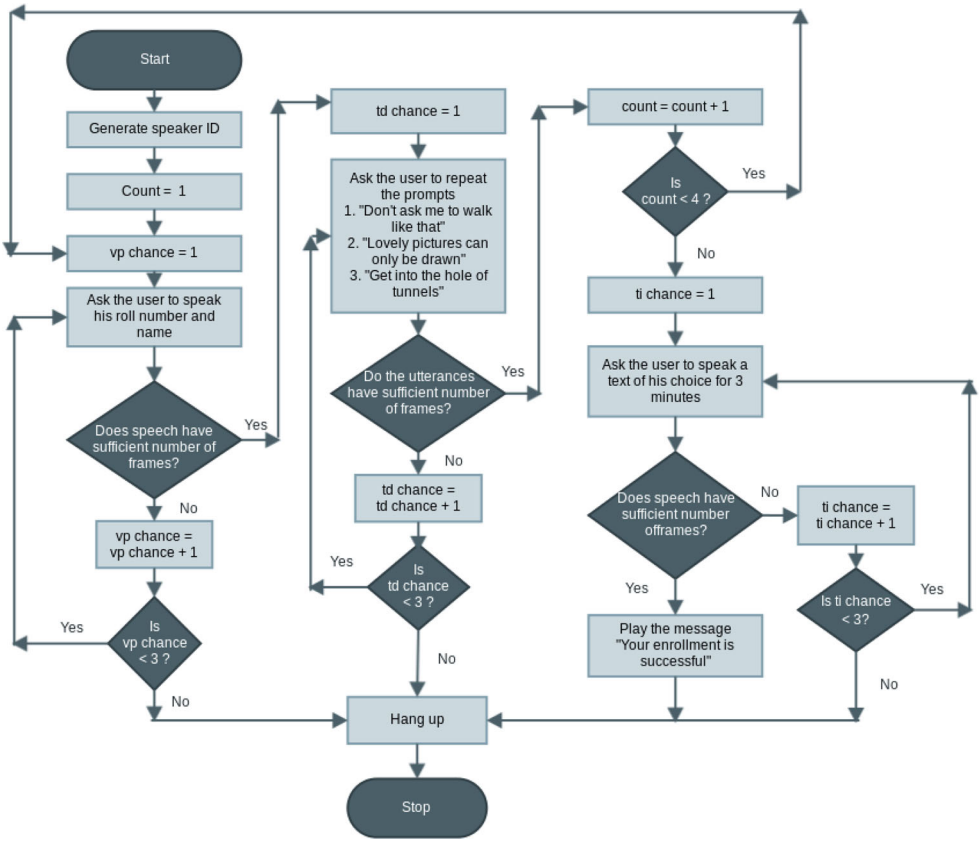

Figure 4 Flowchart for enrolling into multi-level attendance system.

number of speech frames is decided with an experimental analysis done over a small set of speakers. Once the input speech qualifies this threshold, the features are extracted for each of the modules. In case of VP and TD module the MFCC feature vectors after energy based end point detection are extracted and those are saved as reference templates of the speakers. For TI module, the train i-vector of the speakers after energy based VAD are extracted as explained in Subsection 3.3.

4.2 Testing Phase

The testing phase in the multi-level speech biometric authentication system refers to regular attendance marking by the students against their enrolled models. There are about 10 mobile handsets kept at the department office for attendance marking. The students call to the designated toll-free number and enter to the testing phase. Fig. 5 shows the flowchart of the testing process for the multi-level speech based attendance system in detail. The IVR callflow first asks the student to type his/her speaker ID, which is assigned during enrollment and used for verification of a claim. As a part of testing process, the student is asked for his *roll number and name* which then goes as a test speech for the VP module. The MFCC features are extracted after energy based end point detection and compared to the claimed speaker's model by DTW algorithm which gives a distance score to be compared to speaker-specific threshold of that speaker. If the claim is accepted at this module, the callflow announces to the user that the attendance is marked. On the other hand, if it fails then it goes for the TD module which asks for a randomly generated prompt (one out of TD1, TD2 and TD3) as explained in Subsection 3.2. The MFCC features of this utterance are extracted in a similar way to the VP module and compared to the claimed speaker model by DTW algorithm that gives a distance score for comparing against cohort speakers for that model. The acceptance of a claim at this module ends the callflow with declaring the speaker that attendance has been marked. For the speakers which are rejected in both VP and TD module, their claim is processed in the TI module. The features of the test speech data taken for VP and TD modules are combined together to treat them as input for TI module which is then tested using the i-vector based modeling with reference to the claimed speaker. If it accepts a claim then the user is marked present even if the claim is rejected in the preceding modules. Whereas, if the claim is rejected in all the three modules then the user is marked absent and allowed to sign in the register for considering the attendance. The testing phase has a quality check

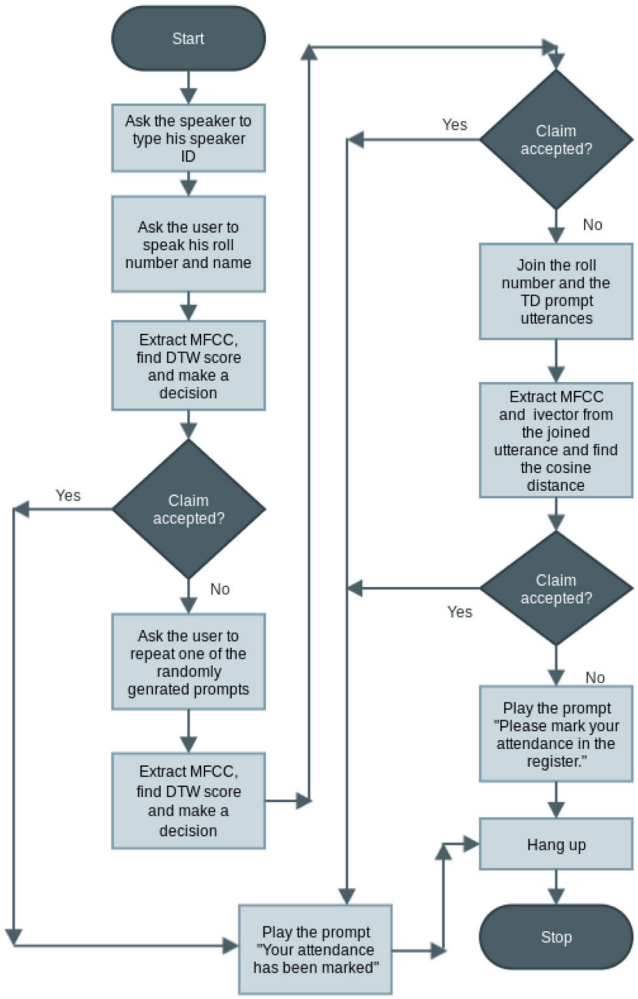

Figure 5 Flowchart for testing into multi-level attendance system.

submodule, similar to the one explained in the training phase for qualifying speech data as input for each module. It is to be noted that the testing from one module to another module transfers upon rejection by the preceding module. As soon as one module authenticates a claim, the user is marked present for that day.

Figure 6 shows the block diagram representation for sequential integration of the three different modules of SV namely, VP, TD and TI into one common framework for multi-level speech based person authentication system with student attendance as an application.

5 Results and Analysis

This section reports the performance analysis of each module of the multi-level person authentication system for student attendance application along with the overall multi-level framework, that is made with integration of the three different modules for SV. Also it includes some of the attempts that are tried offline to deal various practical issues for system deployment.

5.1 Practical Issues based Analysis and Refinements

Several issues came across while developing the multi-level framework for practical application oriented attendance system. One such issue is that if the user does not utter sufficient speech input to each of the modules properly as expected, it may lead to poor speaker modeling during training. Similarly insufficient and improper speech during testing may degrade the system performance. To handle this issue some offline analysis are performed on typical duration (number of speech frames) of speech input, that is required for each of the modules of multi-level framework. Using this information a threshold on minimum number of speech frames expected at each of the modules is set in the multi-level framework. The threshold is fixed by taking (mean + standard deviation) of number of speech frames in 30 cases of input taken for each of the modules. If the input to any of the module has lesser number of speech frames, the user is asked to repeat one more time so that proper speech input can be obtained for respective module. Table 1 shows the minimum number of speech frames that is set for each of the modules. It is to be noted that the speech signal is processed with 20 ms frame size with a shift of 10 ms.

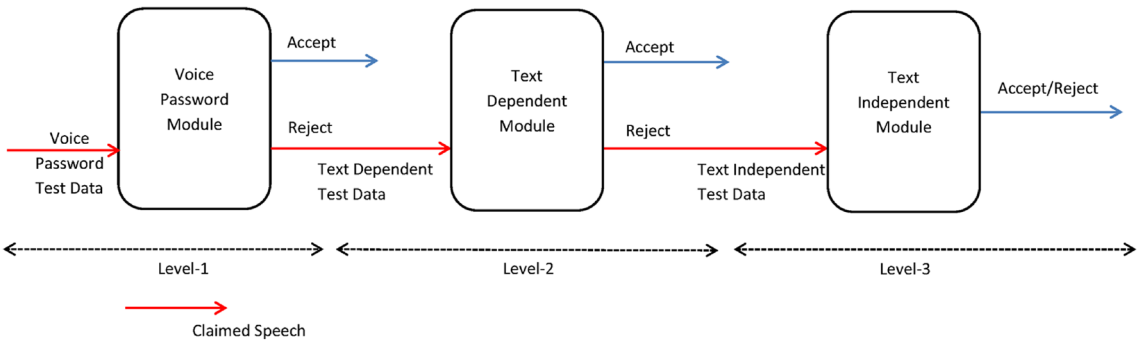

Figure 6 Structure of testing phase in multi-level framework for speech based attendance system.

Table 1 Table showing threshold for number of frames in each module of multi-level framework.

System	Time given to user	No. of speech frames expected
VP module- Training	6 sec	150
VP module- Testing	6 sec	150
TD module- Training	4 sec	90
TD module- Testing	4 sec	90
TI module- Training	3 min	7500
TI module- Testing	10 sec	240

5.2 Performance of Individual Modules: VP, TD and TI

The different methodologies used for the three different modules of the multi-level system are described in Section 3. Each of these modules are evaluated over standard databases to gain the confidence before deploying them into the multi-level framework of the online SV system. The performance of the three modules on standard database as well as on the data collected over multi-level system is discussed in this subsection.

5.2.1 Evaluation on Standard Databases

The VP and TD modules of the multi-level system are evaluated over the RSR2015 database, that is designed for the TD framework based studies [5]. The VP based framework is evaluated for Part I of the database. It contains 30 different fixed pass phrases from 300 speakers. For developing our VP based system, 30 male and 30 female speakers are considered so that each of them have unique pass phrases. Each of the speakers have 9 sessions of each pass phrase, out of which three are taken for training, three for calculation of speaker-specific threshold and remaining three for testing. The evaluation of TD module is also done on the Part I of the RSR2015 database. For this study, a common pass phrase is selected to fit to the TD framework from the 30 male and 30 female speakers. The verification of a claim for TD module is made by the methodology as explained in Section 3.2.

The TI module of the multi-level system based on the i-vector based framework is evaluated over national institute of standards and technologies (NIST) speaker recognition evaluation (SRE) 2012 database, that is designed for speaker recognition based task [2]. This evaluation consists of data collected from telephone as well as from microphone channel, which is from more than 2000 speakers in different conditions. The development set for this task is obtained from previous years SRE evaluation from 2006-2010. The dataset is evaluated under the core task of the evaluation, which has five different testing conditions. These conditions concentrate on test trials under clean as well as in addition of noise. The average performance of the test conditions are mentioned in in this work. The performance of each of the module VP, TD and TI are mentioned in Table 2 in terms of equal error rate (EER) which is gives the optimal performance, where the two kind of errors false rejection rate (FRR) and false acceptance rate (FAR) are equal. The three modules are found to give commendable results, that shows the efficacy to deploy the same methodologies for the practical multi-level system.

5.2.2 Evaluation on the Data Collected over Multi-Level Online System

The multi-level speech based person authentication system is developed for student attendance application for postgraduate students and research scholars of our department. The population of the students is 189 with age group in the range of 22-40. The system is deployed for a period of two months and then studies are made on the performance of each of the modules as well as the multi-level framework. As the performance evaluation of each module is done over a practical system, various issues came for the data collected over online system in a real world environment. These practical issues include prompts like *roll number and name* and text-dependent prompts not spoken properly by the user, failure of end point detection algorithm due to presence of background noise etc. We therefore pooled out a subset of data collected where the practical issues are not present and evaluated the performance of each of the system, which is shown in the second column of Table 3. However if we consider the data with practical issues as mentioned above and then evaluate the performance of each module, we get a poorer performance compared to the former case which can be seen from the third column of Table 3.

Table 2 Performance of individual module of multi-level SV on standard databases.

System	EER
VP module	2.22 %
TD module	3.06 %
TI module	5.34 %

Table 3 Performance of individual module of multi-level SV over the data collected over multi-level online system.

System	EER (actual)	EER (practical)
VP module	2.70 %	9.44 %
TD module	3.13 %	10.14 %
TI module	7.65 %	11.50 %

5.3 Performance of Multi-Level Framework

The three modules are combined in a sequential manner in the multi-level framework to get benefit from each of the modules. Table 4 shows the performance of the multi-level framework in terms of EER at each level. The EER of VP module under practical condition is found to be 9.44 %, which is the first level of multi-level framework. The claims rejected by this module goes to the TD module for verification and those which are rejected by both VP and TD modules, third level of verification takes place in TI module. In this way testings are performed in a sequential manner that helps in achieving EER of 4.23 % for the multi-level system. Thus, the combination of the three modules helps in achieving a significant improvement in EER than that of each system in a practical scenario with attendance as an application. Fig. 7 clearly shows that fusion of three levels provides a better performance for the multi-level framework.

Apart from the sequential way of fusion of the three modules of multi-level framework, the score level fusion of the three modules is carried out. In this regard before score fusion, transformation of scores is required as the modules VP and TD generates distance score which indicates less scores for the trials of same speaker and more scores for the trials from different speakers. Whereas the TI module generates cosine similarity score that indicates higher scores for likely speaker and vice versa. Therefore we transform the distance scores generated from VP and TD module to similarity scores in the range -1 to 1 by inverse mapping with resemblance to cosine kernel score of the TI module. The score fusion is first performed by averaging

Table 4 Performance of multi-level framework and sequential fusion in terms of EER.

System	EER
VP (Level-1)	9.44 %
TD (Level-2)	10.14 %
TI (Level-3)	11.50 %
VP+TD (Level-1+Level-2)	6.35 %
VP+TD+TI (Level-1+Level-2+Level-3)	4.23 %

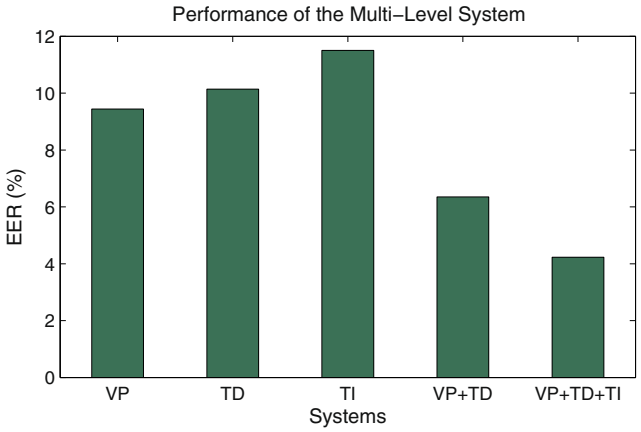

Figure 7 Performance of each module and their fusion of the multi-level speech based attendance system.

the scores for VP and TD module, then later including the scores obtained from three modules. The study is initially made over a background set of 30 speakers collected over online multi-level framework. A threshold is set on this set maximum separation of the genuine and the impostor trials. Once this threshold is set, that is applied on the dataset that is considered for sequential fusion of three modules. The performance under score fusion of multi-level framework is shown in Table 5. The result of score fusion indicates that the combination of VP and TD module is similar to the sequential way of fusion of two module. However the score fusion of all the three modules outperforms the sequential fusion by showing an EER of 2.67 %. Although the score level fusion is found to work in better way than the sequential way of fusion, the three modules are combined in a sequential manner to fit the framework according to less time complexity issue which is described in the next subsection.

5.4 Time Complexity of Multi-Level System

The time complexity issue is taken into account while developing the multi-level SV. Table 6 shows the testing time for each module of the multi-level SV system. The testing time for VP and TD module is very less compared to the TI module as only a single phrase is asked from the users for VP and TD modules. Whereas for TI module considers the combined speech input of VP and TD module as input

Table 5 Performance of multi-level framework under score fusion in terms of EER.

System	EER
VP+TD (Level-1+Level-2)	6.00 %
VP+TD+TI (Level-1+Level-2+Level-3)	2.67 %

Table 6 Table showing average testing time of each module in multi-level framework.

System	Avg. Testing Time
VP module	0.67 s
TD module	1.31 s
TI module	2.67 s

Table 7 Table showing recognition rate of the speech recognition module in the multi-level framework.

TD Prompt Category	Rec. Rate
TD1-Prompt	94.73 %
TD2-Prompt	93.50 %
TD3-Prompt	95.69 %

that makes more amount of speech data involvement under it. Also the complexity of VP and TD module is simpler as MFCC features of the test examples are matched to the reference speakers by a DTW algorithm to compute Euclidean distance. On the other hand in TI module, the MFCC features obtained from the speech input to VP and TD module are combined and then statistics are computed, upon which the T-matrix is projected to obtain low dimensional compact representation called i-vectors, that are matched to the claimed speaker i-vector by cosine kernel scoring after having channel/session compensation. Thus due to multiple steps involved in TI module, the processing time is slightly more than the other two modules. In this regard the three modules of the multi-level framework are combined in a sequential manner to arrive at a decision with less amount of time involved. A claim is accepted in the sequential flow of testing once accepted in any of the modules. If it is accepted in the first level then the callflow does not proceed further for second, third level and the user is announced that the attendance is marked. Similarly if the claim is accepted in second level after being rejected in the first level, then the third level of the system is not invoked in the callflow, thus optimizing the time involved for testing. However the score level fusion of the multi-level framework can be also done with more amount of time involved to give decision on the fly.

5.5 Integration of Speech Recognition Module into the Multi-Level Framework

A sub module of speech recognition is included in the TD module of the multi-level framework to check the validity of a real user to some extent. As described in Section 3.2, during testing of the multi-level framework one random TD prompt out of the three prompts is asked to the user through the IVRS callflow. Once the user utters the prompt, the speech recognition sub module is invoked in the TD module. The prompt uttered by the user is matched to the respective claimed speaker's three training templates for each of the TD prompt category by DTW algorithm. The distance score should be minimum with respect to the random prompt played by the IVRS. There is one more level of checking in this speech recognition sub module, where the distance score is checked with respect to the threshold of score for each of the prompt category. This threshold is obtained by considering mean and standard deviation of the scores obtained from several testing made against each of the prompt category by different users, where we define the threshold as (mean + standard deviation) for particular prompt category. This second level checking is done mainly to avoid out of context prompts in the TD module. The performance obtained in speech recognition by incorporating this method is shown in Table 7 for each TD prompt category, which is useful to check the reality of the user and also to discard the improper data uttered by the user.

6 Conclusion

This work describes an effort made to develop speech based person authentication system involving three different modules of SV under low security applications. A multi-level framework is designed using VP, TD and TI SV over the telephone network. The system is deployed for attendance application with the easement of end users considering lesser time complexity and evaluated for a period of two months. The multi-level system performs better in combination of all the three modules of SV compared to each module by taking the advantage of each of the categories of SV, which shows its potential for practical system based application. Also the functionality of each module can be moderated according to the type of application for which the system is designed. The future work will concentrate on exploring the deployment issues relating to the multi-level framework by improving user interface, robustness to spoofing attacks etc.

References

1. "The NIST Year 2010 Speaker Recognition Evaluation Plan", NIST, (2010).
2. "The NIST Year 2012 Speaker Recognition Evaluation Plan", NIST, (2012).
3. Hatch, A. O., Kajarekar, S., & Stolcke, A. (2006). Within-class covariance norMalization for svm-based speaker recognition. In *Proc. of ICSLP, pp. 1471–1474*.

4. Jain, A. K., Ross, A., & Prabhakar, S. (2004). An introduction to biometric recognition. *IEEE Trans Circuits Syst Video Technol*, *14*(1), 4–20.
5. Larcher, A., Lee, K. A., Ma, B., & Li, H. (2014). Text-dependent speaker verification: classifiers, databases and RSR2015. *Speech Comm*, *60*, 56–77.
6. Ross, A., & Jain, A. K. (2014). Multimodal biometrics: an overview. *In 12th European IEEE Signal Processing Conference*, 1221–1224.
7. Putra, B. (2011). Suyanto: Implementation of secure speaker verification at web login page using mel frequency cepstral coefficient-gaussian mixture model (mfcc-gmm). In *Instrumentation control and automation (ICA), 2011 2nd international conference on, pp. 358–363*.
8. Atal, B. S. (1976). Automatic recognition of speakers from their voices. *Proc IEEE*, *64*(4), 460–475.
9. Chakrabarty, D., Prasanna, M. S. R., & Das, R. K. (2013). Development and evaluation of online text-independent speaker verification system for remote person authentication. *Int J Speech Technol*, *16*(1), 75–88.
10. Reynolds, D. A., Thomas, F. Q., & Dunn, R. B. (2000). Speaker verification using adapted Gaussian mixture models. *Digital Signal Process*, *10*(1-3), 19–41.
11. Sarkar, G., & Saha, G. (2010). Real time implementation of speaker identification system with frame picking algorithm. *Procedia Computer Science*, *2*(0), 173–180. Proc. of the Int. Conference and Exhibition on Biometrics Technology.
12. Lee, K.-A., Larcher, A., Thai, H., Ma, B., & Li, H. (2011). Joint application of speech and speaker recognition for automation and security in smart home. In *INTERSPEECH, pp. 3317–3318*.
13. Muda, L., Begam, M., & Elamvazuthi, I. (2010). Voice recognition algorithms using mel frequency cepstral coefficient (mfcc) and dynamic time warping (dtw) techniques. *CoRR*, 2151–9617.
14. Hébert, M. (2008). Text-dependent speaker recognition. In *J. benesty, M. Sondhi, Y. Huang (eds.) Springer Handbook of Speech Processing, pp. 743-762. Springer Berlin Heidelberg*.
15. Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., & Ouellet, P. (2011). Front-end factor analysis for speaker verification. *IEEE Trans Audio Speech Lang Process*, *19*(4), 788–798.
16. Auckenthaler, R., Carey, M., & Thomas, H. L. (2000). Score norMalization for text-independent speaker verification systems. *Digital Signal Process*, *10*(1–3), 42–54.
17. Duda, R. O., Hart, P. E., & Stork, D. G. (2000). *Pattern Classification*: John Wiley and Sons.
18. Ramos-Lara, R., Lpez-Garca, M., Cant-Navarro, E., & Puente-Rodriguez, L. (2013). Real-time speaker verification system implemented on reconfigurable hardware. *Journal of Signal Processing Systems*, *71*(2), 89–103.
19. Rao, S., & Satoa, K. J. (2013). An attendance monitoring system using biometrics authentication. *International Journal of Advanced Research in Computer Science and Software Engineering*, *3*(4).
20. Bolle, R.d, & Pankanti, S.harath. (1998). *Biometrics, personal identification in networked society: Personal identification in networked society*. norwell, MA, USA: Kluwer academic publishers.
21. Dey, S., Barman, S., Bhukya, R. K., Das, R. K., Haris, B. C., Mahadeva Prasanna, S. R., & Sinha, R. (2014). Speech biometric based attendance system. In *National conference on communications*.
22. Furui, S. (1981). Cepstral analysis technique for automatic speaker verification. *IEEE Trans Acoust Speech Signal Process*, *29*(2), 254–272.
23. Sahoo, S. K., Choubisa, T., & Mahadeva Prasanna, S. R. (2012). Multimodal biometric person authentication : a review. *IETE Tech Rev*, *29*(1), 54–75.
24. Nawas, T., Pervaiz, S., Korrani, A., & Azhar-ud-din (2009). Development of academic attendence monitoring system using fingerprint identification. *International Journal of Computer Science and Network Security*, *9*(5).
25. Kinnunen, T.omi., & Li, H.aizhou. (2010). An overview of text-independent speaker recognition: From features to supervectors. *Speech Comm*, *52*(1), 12–40.
26. Fisher, W. M., Doddington, G. R., & Goudie-Marshall, K. M. (1986). The DARPA Speech Recognition Research Database: Specifications and Status. In *Proceedings of DARPA workshop on speech recognition, pp. 93–99*.
27. Kawaguchi, Y., Shoji, T., Lin, W., Kakusho, K., & Minoh, M. (2009). Face recognition-based lecture attendance system. Department of Intelligence Science and Technology, Graduate School of Informatics Kyoto University.

Rohan Kumar Das received B. Tech degree in Electronics and Communication Engineering from North-Eastern Hill University (NEHU), Shillong, India in the year 2010. Currently he is pursuing his doctoral studies at Department of Electronics and Electrical Engineering, Indian Institute of Technology (IIT) Guwahati, Guwahati, Assam from the year 2012. Prior to joining IIT Guwahati he worked as a Project Scientist at Assam Science Technology and Environment Council. His research interests are speech signal processing, speaker verification, machine learning and pattern recognition.

Sarfaraz Jelil was born in Guwahati, India, in 1987. He received his B.Tech and M.Tech degrees in Information Technology from North-Eastern Hill University (NEHU), Shillong, India in 2011 and 2015, respectively. In 2012, he joined the Department of Electronics and Communication Engineering, NEHU, Shillong, as a Project Engineer. Currently, he is pursuing his Ph.D in the Department of Electronics and Electrical Engineering, Indian Institute of Technology (IIT) Guwahati, Guwahati, India. His research interests include speaker verification, machine learning and pattern recognition.

S. R. Mahadeva Prasanna was born in India in 1971. He received the B.E. degree in electronics engineering from Sri Siddartha Institute of Technology, Bangalore University, Bangalore, India, in 1994, the M.Tech. degree in industrial electronics from the National Institute of Technology, Surathkal, India, in 1997, and the Ph.D. degree in computer science and engineering from the Indian Institute of Technology Madras, Chennai, India, in 2004. He is currently a Professor in the Department of Electronics and Electrical Engineering, Indian Institute of Technology, Guwahati. His research interests are in speech and signal processing.