INTERSPEECH 2019: Show & Tell Contribution
September 15–19, 2019, Graz, Austria

SpeechMarker: A Voice based Multi-level Attendance Application

Sarfaraz Jelil[1], *Abhishek Shrivastava*[2], *Rohan Kumar Das*[3], *S. R. M. Prasanna*[1,4], *Rohit Sinha*[1]

[1]Dept. of Electronics and Electrical Engineering, Indian Institute of Technology Guwahati, India
[2]Dept. of Design, Indian Institute of Technology Guwahati, India
[3]Dept. of Electrical and Computer Engineering, National University of Singapore, Singapore
[4]Dept. of Electrical Engineering, Indian Institute of Technology Dharwad, India

{sarfaraz, shri, prasanna, rsinha}@iitg.ernet.in, rohankd@nus.edu.sg

Abstract

This work describes a multi-level speaker verification (SV) framework that is accessible via a graphical user interface (GUI) with attendance as an application. This framework has three different modalities of SV system, namely, voice-password, text-dependent and text-independent. The decision for attendance marking can be taken from each of the modalities or by fusion. There are two operating modes of the developed GUI, which are user and debug modes. The user mode is for general users to mark attendance, whereas the debug mode is to study the behavior of the three modalities from deployment point of view. The speech waveforms, different plots and scores can be analyzed in the debug mode for analysis. The system has been deployed successfully for regular attendance marking among a closed group in a laboratory environment.

Index Terms: speaker verification, voice-password, text-dependent, text-independent, multi-level

1. Introduction

The research in speaker verification (SV) has made significant progress in the recent decade to have very high performing systems. With such advancements, there have been efforts towards the development of application driven systems [1–4]. From the perspective of user comfort and performance efficacy, text-dependent SV is found to have an edge over text-independent SV. However, the former is quite vulnerable to spoofing attacks as the lexical content is same across all the users. On the other hand, text-independent SV is advisable for a practical system that captures speaker characteristics in a more generic manner as well as providing robustness across different variabilities.

In this work, three modalities of SV, namely, voice-password, text-dependent and text-independent are used in a common platform to develop a GUI with multi-level authentication for marking attendance. The developed GUI is referred to as SpeechMarker, which is an extension to our previous research work [5]. Each module of the multi-level system considers mel frequency cepstral coefficients (MFCC) as the common features. For voice-password and text-dependent module, the verification process is based on dynamic time warping (DTW) based temporal alignment between the MFCC features of train and test utterances [6]. On the other hand, the text-independent module is based on i-vector based speaker modeling that finds the match between the train and the test i-vectors to accept/reject a claim [7]. The three systems are combined in a sequential manner and a user interactive GUI is developed over laptop/desktop platform for multi-level SV. A user can enroll as well as mark the attendance using a head mounted microphone. This framework for attendance application has been deployed among a closed group in a laboratory environment.

The rest of the paper is organized as follows: Section 2 gives description of the attendance application SpeechMarker. In Section 3, the two modes of operation of SpeechMarker are detailed with conclusion in Section 4.

2. SpeechMarker: System Description

This section describes the three modules of the SpeechMarker application in a brief manner. The modules voice-password, text-dependent and text-independent constitute the multi-level framework of SpeechMarker.

2.1. Voice-password

The voice-password module deals with unique fixed phrases for all the users in the system. Each user enrolls by setting a unique phrase against his/her model, that is to be repeated during test sessions. During enrollment three sessions of the unique phrases are taken and MFCC features of those utterances are extracted to create their reference templates. During testing, the MFCC features of the test utterance are extracted and temporal alignment of the train and the test utterances is made via DTW algorithm that generates a distance score. This distance score is compared to the speaker-specific threshold of each user, which is computed by a few thresholding sessions from each user for taking decision with respect to a claim. In this multi-level framework, the name and phone number of each user are considered as the input for voice-password module that differs from one user to another.

2.2. Text-dependent

The text-dependent module considers a global phrase for all the users enrolled to the system. In this case, each user enrolls to the system using a common phrase for three sessions. The same phrase has to be produced during testing sessions. The verification methodology is similar to that of the voice-password module based on MFCC and DTW based framework. The decision is taken with respect to a set of four cohort speakers that are chosen randomly and kept against each of the speaker models. The claimed speaker score is compared to the scores obtained from the cohort speakers to conclude to a decision. In the context of text-dependent module, a set of three fixed phrases are used for enrollment. During testing, one out of the three phrases is displayed to the user to speak that ensures the liveness of the user to tackle spoofing attacks to some extent.

2.3. Text-independent

The text-independent module of the multi-level system considers i-vector based speaker modeling approach. The training session takes about 1 minute of speech data in the form of read

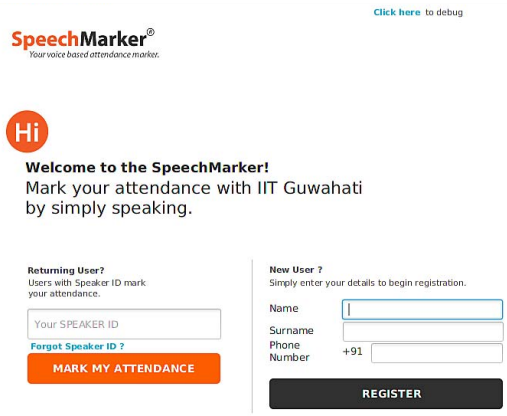

Figure 1: *Graphical user interface of the SpeechMarker.*

speech from each user. For this purpose, a text is displayed in the GUI which one has to read. During testing, few phrases of about 2-3 lines are displayed that the users have to read with a time around 10 seconds. As the three modules of the multi-level are in sequential manner, therefore the test data for voice-password and text-dependent module are merged along with 10 seconds of speech that is taken for text-independent module to consider the entire set as test example for text-independent module. Linear discriminant analysis (LDA) and within class covariance normalization (WCCN) are used for channel/session compensation [7]. The background models for this framework are learned using the NIST SRE 2010 database [8]. Finally, the similarity is measured between the train and test i-vectors to obtain a score. A global threshold is computed for concluding to a decision based on few examples of genuine and impostor trials of a background set speakers over the same framework.

3. Modes of Operation

This section explains the modes of operation of the GUI which is developed for attendance application. There are two modes on which it operates: user mode and debug mode. Figure 1 shows the home screen for GUI of SpeechMarker, where a new user can enroll by entering the name and phone number to get a four digit speaker ID at the end of enrollment session. The same home screen may be used by an enrolled user for testing against respective speaker ID. It also has a provision to retrieve the speaker ID if any particular user forgets. The modules of the multi-level system for attendance application can be also reconfigured as well as enabled as per requirement for decision making. A video[1] for demonstrating the use of the SpeechMarker has also been prepared.

3.1. User Mode

The user mode is basically designed for general users that test the attendance application on a regular basis to mark their attendance. Under this mode, there is a provision to check the input speech quality of the microphone and then proceed with the recordings. Speech/non-speech detection is done at each module and there is provision to play back the recorded speech to ensure the speech input, if the user has any ambiguity on it. Further, there is a help option at the GUI at each step to guide the user to smoothly enroll and test the system. If the user is verified in a module then, it is displayed to the user, else it goes to the subsequent module for testing.

[1] https://youtu.be/8LjMUPZY0eg

3.2. Debug Mode

The debug mode is mainly configured for the administrators and the persons who are willing to analyze each stage with a closer understanding. In this mode, the speech waveform is also displayed to the user while checking the quality of input speech. At each module of the multi-level system, the waveforms are displayed so that the user can view it. Further, under debug mode, provisions are there to modify the speech input in terms of cutting the waveform if there is background speech present at the start or at end, etc. During testing, the warping path obtained by using DTW algorithm for voice-password and text-dependent module is displayed to the user. For text-independent module, the score obtained between by comparing the train and test i-vectors is displayed alongwith the distribution of genuine and impostor scores which are used during threshold calculation for better understanding.

4. Conclusion

This work discusses about an interactive GUI, which is developed for attendance application using multi-level SV framework. Three different modalities of SV, namely, voice-password, text-dependent and text-independent are used in a common framework for development of the multi-level system. The users can use the GUI over desktop/laptop interface with a head mounted microphone for enrollment as well as testing. The verification of a claim can be made in fusion of all the three levels or with respect to each modality. The developed GUI is deployed successfully on a closed set of people for attendance marking.

5. Acknowledgement

This work is developed as part of a project grant 12(6)/2012-ESD for the project entitled “Development of Speech-Based Multi-level Person Authentication System” funded by the Department of Electronics and Information Technology (DeitY), Govt. of India.

6. References

[1] K.-A. Lee, B. Ma, and H. Li, “Speaker verification makes its debut in smartphone,” in *SLTC Newsletter*, February 2013.

[2] K. A. Lee, A. Larcher, H. Thai, B. Ma, and H. Li, “Joint application of speech and speaker recognition for automation and security in smart home,” in *INTERSPEECH*, 2011, pp. 3317–3318.

[3] D. Chakrabarty, S. R. M. Prasanna, and R. K. Das, “Development and evaluation of online text-independent speaker verification system for remote person authentication,” *International Journal of Speech Technology*, vol. 16, no. 1, pp. 75–88, 2013.

[4] S. Dey, S. Barman, R. K. Bhukya, R. K. Das, Haris B C, S. R. M. Prasanna, and R. Sinha, “Speech biometric based attendance system,” in *National Conference on Communications*, 2014.

[5] R. K. Das, S. Jelil, and S. R. M. Prasanna, “Development of multi-level speech based person authentication system,” *Journal of Signal Processing Systems*, vol. 88, no. 3, pp. 259–271, September 2017.

[6] L. Rabiner and B. Juang, *Fundamentals of speech recognition*. Prentice-Hall, Inc. Upper Saddle River, NJ, USA, 1993.

[7] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, “Front-end factor analysis for speaker verification,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, May 2011.

[8] NIST speaker recognition evaluations. [Online]. Available: www.itl.nist.gov/iad/mig//tests/spk/