VIETNAM NATIONAL UNIVERSITY

UNIVERSITY OF ENGINEERING AND TECHNOLOGY

SPEECH PROCESSING 2021
FINAL REPORT

VOICE-BASED TIME SHIFT AUTHENTICATOR

Course	INT3411 20	
Lecturer	Trần Quốc Long	
Date	6.6.21	
Member	1, Trần Quang Vinh - 17021357	Preprocessing Data Gaussian Mixture model
	2, Lê Minh Tâm - 17021332	Web interface Train model

1. Motivation

Voice has unique characteristics that can be used to identify a person, just like a fingerprint. Using voice is a new innovative tool – essentially offering a level up in security that simplifies the authentication experience for users. Speaker verification can be text-dependent or text-independent. The former requires users to

##Enrollment for speaker verification is text-independent - a process of verifying the identity without constraint on the speech content, in which users can speak freely to the system.

This application is built as an authentication system to help tracking worker's shift attendance. Users use their id and voice to check in/out their work shift. The system is text-independent so users can speak freely.

2, Features

Users need to add a username and speak some random sentences and submit. System will extract the utterance's features and pass them to the corresponding Gaussian Mixture model (GMM) model. If the model returns a score lower than the threshold, the user is successfully verified and the system will check in/out his/her shift. If not, this person has to re-record and submit again.

GMM models are trained using SP Class' Database; hence, the demo can verify most students in the class. Moreover, this application is text-independent, no fixed sentences are set so users can say freely whatever they want at the verification stage. We would like to be able to improve our own models' accuracy in order to put them in a near future demo release.

3. Tech Stack + Method

3.1 Tech Stack

- Python + Flask for main components;
- Scikit-learn for GMM model and feature extraction

- SQLite for database connection

3.2 Method

The verification system consists of pre-processing, feature extraction, and modelling stage.

3.2.1 Preprocessing

- Dataset:
  + The dataset is taken from shared data made by Speech Processing Class's students. 70 students with 699 record files of students speaking articles from VNExpress.
  + Name's label presented within the dataset.

- Pre-processing dataset:
  The pre-processing is used to reduce file size while retaining frequency range of human speech.
  + Convert stereo audio file to mono using librosa
  + Downsample wav files from ((44kHz - 48kHz) to 16kHz

		Datasize
Original Size		24 GB
Preprocessing	Stereo to mono & Downsample	**4,5 GB**

- Feature Extraction
  + Extract MFCC using window length of 25ms, hop length of 10ms, 13 coefficients => 13 features
  + Perform Cepstral Mean Subtraction to MFCC
  + Extract deltas and deltas deltas from MFCC => 13 and 13 features

=> Result: feature vector with size of (number of frames, 39)

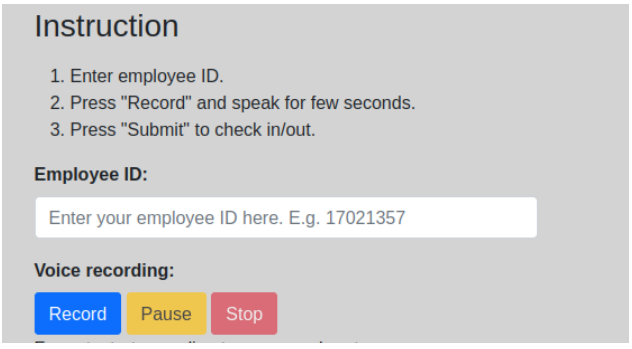
3.2.2 Modeling

- Use Gaussian Mixture Model (GMM) with 16 mixture components

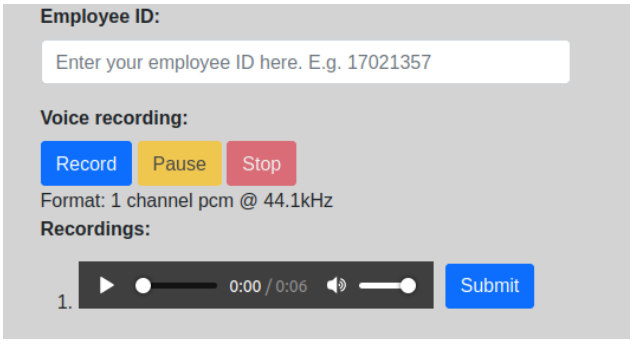
- For each speaker, the feature vectors of all files are vertically stacked into a 2D array. These arrays are the input of GMM.
- 70 models named studentID_studentName.gmm.
- Evaluation:
  + Used 20 files not from the dataset of 2 speakers (10 files/speaker) and 4 models (2 from the speakers, 2 from the imposter). These files are recorded at different times.
  + Evaluated with different thresholds from 1.0 to 3.0.
  + Metrics:
    + False Alarm Probability = # imposter accepted / # imposter attempts
    + Miss Probability = # legit speakers rejected / # legit speakers attempts

Threshold	False Alarm Probability	Miss Probability	Threshold	False Alarm Probability	Miss Probability
1.0	21.67%	75%	2.3 (chosen)	36.67%	45%
1.1 - 1.4	21.67%	70%	2.4	38.33%	45%
1.5 - 1.6	23.33%	70%	2.5 - 2.6	40%	45%
1.7	25%	60%	2.7	41.67%	35%
1.8 - 2.0	28.33%	60%	2.8	45%	30%
2.1	33.33%	55%	2.9-3.0	45%	25%
2.2	35%	50%			

4. Implementation

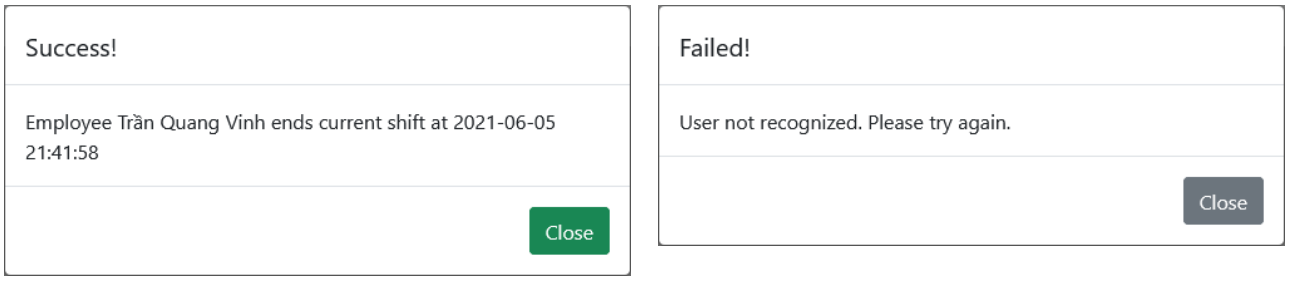


Home Page



User Recording

Input



If score under threshold | User Recording

Output

5. Difficulties

- Models' accuracy is mediocre.
    + If the environment is noisy, models usually output scores that are significantly higher than the threshold.
    + Slight change in the way the speaker speaks can cause the model to output scores that are slightly above the threshold.

6. Reference

[1] https://github.com/Atul-Anand-Jha/Speaker-Identification-Python/