A Novel Approach of Voice Recognition Using MFCC and GMM, Speech Recognition and Text Recognition to Assist for Email Communication for Visually Impaired People

Senthuja Karunanithy
University of Jaffna
senthunithy@gmail.com

SLSWCS Student Workshop on Computer Science

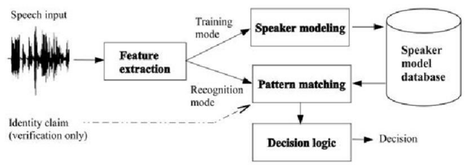
Abstract

Nowadays, Human-computer interaction plays a prominent role in the day to day life. However, it has become a challenging task for visually impaired people to get involved with computers in their day-to-day activities because of limited accessibility to the input mechanism. This work proposes speech-to-text, text-to-speech, and voice recognition techniques giving access to blind people to interact with Email communication. Voice recognition helps to recognize the voice of a specific person from the audio recording as voice is different from each other than the fingerprint where speech recognition helps to disregard the language and meaning to detect by the person behind the speech. The proposed model is based on the classification of MFCC coefficients obtained from speech signals with GMM for voice recognition. The proposed method is evaluated using VoxForge Dataset; containing the 340 voices of 34 speakers and obtained the result with 100% success.

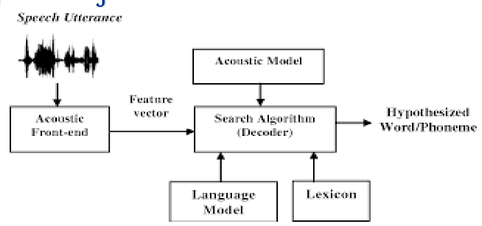
Keywords: Text Recognition, Speech Recognition, Voice Recognition, Mel Frequency Cepstral Coefficient (MFCC), Gaussian Mixture Modelling (GMM)

Introduction

Speaker Recognition



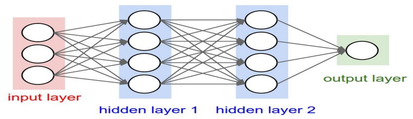
Speech Recognition



Objective

The objective of the research is to develop a voice-based email system that would help blind people to access email. The system will not let the user makes the use of the keyboard instead will work on speech recognition and voice recognition.

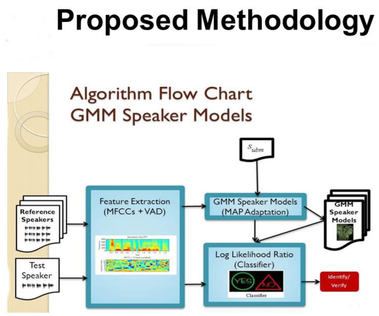
Background



Methodology

The methodology can be summarized in six basic phases:

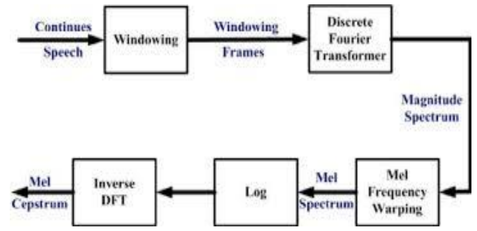
- Data Acquisition
- Data pre-processing
- Feature Extraction
- Model Training
- Perform Testing (identification)
- Application



Data Acquisition
Methodology is assessed with the VoxForge DATASET

Data preprocessing
The data must be preprocessed in order to achieve better outputs and prediction results. This is to ensure that the model is trained with minimum errors. Vox-Forge dataset was already clean and noise free.

Feature Extraction: Mel frequency Cepstral coefficient (MFCC) estimation



Functions provided in python_speech_features module

```
pythonspeechfeatures.base.fbank(signal,
samplerate=16000, winlen=0.025, winstep=0.01,
nfilt=26, nfft=512, lowfreq=0, highfreq=None,
preemph=0.97,winfunc=<function <lambda>>)
```

Model Training: Front-end processing
The objective in the front-end processing is to modify the speech signal, so that it will be more suitable for feature extraction analysis. The front-end processing operation based on noise cancelling, framing, windowing and pre-emphasis.

Speaker modelling
The objective of modeling technique is to generate models for each speaker using specific feature vector extracted from each speaker. It performs a reduction of feature data by modeling the distributions of the feature vectors. The speaker recognition is also divided into two parts that means speaker dependent and speaker independent. In the speaker independent mode of the speech recognition the computer should ignore the speaker specific characteristics of the speech signal and extract the intended message .on the other hand in case of speaker dependent mode speech recognition machine should extract speaker characteristics in the acoustic signal.

Speaker database
The speaker models are stored here. These models are obtained for each speaker by using feature vector extracted from each speaker. These models are used for identification of unknown speaker during the testing phase.

Decision logic
It makes the final decision about the identity of the speaker by comparing unknown speaker to all models in the data base and selecting the best matching model.

Perform Testing
The log-likelihood for each GMM of every speaker was calculated in the model training phase. It was stored as a database in a separate folder. This data dictionary is used for matching 1:N speaker's file. The speaker with the highest score is chosen and identified.

Results

Vox-Forge: 34 speakers each accompanied seven voice samples for training data, and three voice samples for testing data totaling 340 cleaned and preprocessed voice sample Data. Each voice sample was around 4 sec. in length, and thus the default value of `nfft=512` in `mfcc()` worked fine.

Training corpus: It has been developed from audios taken from 'on-line Vox-Forge speech database' and consists of seven speech utterances for each speaker, spoken by 34 speakers (20-30 seconds/speaker).

Test corpus: This consists of remaining three unseen utterances of the same 34 speakers taken in train corpus. All audio files are of 10 seconds duration and are sampled at 16000 Hz. Thus, speaker identification was successfully conducted with an outstanding result on the dataset. The accuracy was 100% in case of VoxForge Dataset. MFCC- GMM model gives satisfactory results.

Why it is specific

It is unique, because here we do not consider the language of the speaker. Whatever the language spoken by the user is not the matter. This research mainly focus on the tone, frequency *etc*.

Discussion & Conclusion

MFCC algorithm is used in our system as it has the least false acceptance ratio. In order to improve system performance and also to achieve high accuracy GMM model can be used in the feature matching technique.

Reference

Alif Khan, Shah Khusro, Badam Nasi, Jamil Ahmad, Iftikhar Alam and Inayat Khan, *"Tetramail: a usable email client for blind people"*, Universal Access in the Information Society, September 2018.

Acknowledgement

Foremost, I would like to sincerely thank my supervisor Dr. S. Mahesan and all other lecturers who continually supported me during my research.