INDIAN INSTITUTE OF INFORMATION TECHNOLOGY TIRUCHIRAPPALLI

Department of Electronics and Communication Engineering

ECOE18- Digital Speech Processing

Assignment 02

Posted: Mar.24,2020 **Due: Mar.27,2020**

Note: You can use either MATLAB or Python. The objective is to design ASR to identify the digit spoken by the speaker. Here you must extract the MFCC features from the samples. Paste the coding and corresponding plots in a word, convert to pdf and upload in google classroom. Name the document with your Roll number. Execution must be shown in person once you come back.

Introduction:

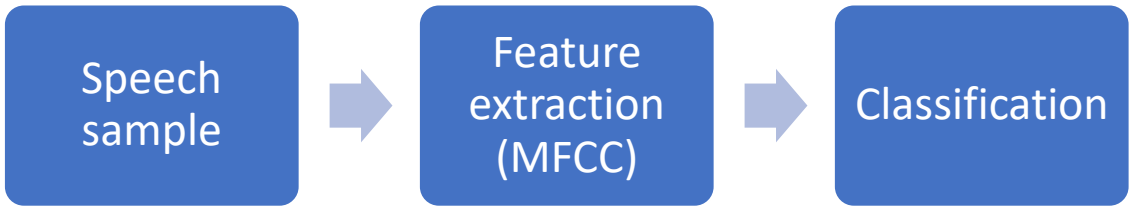
Automatic Speech Recognition (ASR) is defined as a computer-driven transcription of the spoken word into readable text. The main aim of ASR technology is to correctly identify the words spoken by a person. The importance of ASR is to allow a computer to recognize the words that are spoken by any human being independent of vocabulary size, noise, speaker characteristics or accent. We use Mel-frequency cepstral coefficients (MFCCs) for extracting features from speech.

ASR is a technology that allows a machine to recognize the person's spoken words and convert it into written text. The speech signals are slowly timed varying signals (quasi-stationary), but in a short period of time (5–100 ms), its characteristics are fairly stationary. Speech is the most natural form of human communication. The first step of speech recognition is speech signal pre-processing and feature extraction where the relevant information from the speech sample is extracted to characterize the time varying properties of an acquired speech sample. In feature extraction phase, feature vectors are extracted using feature extraction methods like cepstral coefficients of speech signal.

Digit recognition is the part of speech recognition with increasing importance among speech processing application. Digit recognition is getting more and more attention since last decade due to its wide range of application.

This is because it has an importance in several fields and used in checks in banks or for recognizing numbers on car plates, railway PNR or many other applications.

Proposed Methodology:



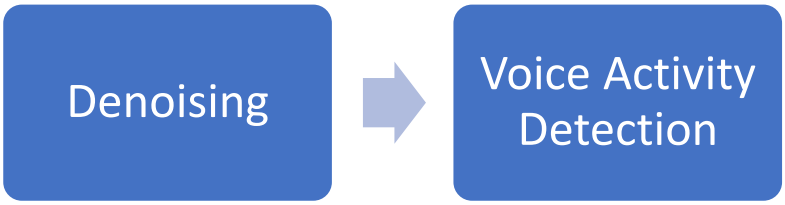
This is the actual methodology of digit recognition and in this assignment, you are supposed to do first two steps only.

Database Description

- The speech database is recorded in a lab environment. English digit zero to nine (0–9) has been recorded from 10 adult male speakers of age group 24–30 years where each speaker has repeated each word 20 times. The sampling rate is 16 kHz, and the sampling bit resolution is 16 bits/sample. For every digit, 200 data samples were recorded. The database is available on http://www.iitg.ernet.in/pkdas/digits.rar .

Or

- you can record your voice saying digits from 0 to 9 in the same conditions specified above. In that case you must preprocess the signals before framing



- Denoising –design suitable filters to remove unwanted interferences or noises
- Voice Activity Detection - to select voiced/ speech, the unvoiced/ non-speech sections and the silence regions in the signal
    - Using some of the classic parameters like zero-crossing rate, energy of the signal and autocorrelation function. (which was your first Assignment)

Framing & Windowing:

At the initial stage we analyse each frame of the signal in the short time instead of analysing the signal all at once. This is achieved by splitting the signal into several frames. The speech signal is mostly stationary for the range of 10–30 ms. 50 % overlap is applied to the frames. After framing a window function is multiplied with each frame of speech signal. Windowing is done to remove unnatural discontinuities in the speech segment and the distortion in the underlying spectrum. The choice of the window depends upon several factors. In speaker recognition process the most used window shape is the Hamming window.

Fourier Transform (FFT):

Fast Fourier Transform (FFT) converts the signal from time domain to frequency domain and prepare for the next stage (Mel frequency warping). The basis of the performing Fourier transform is to convert the convolution of the glottal pulse and the vocal tract impulse response in the time into multiplication in the frequency domain. The FFT gives the magnitude frequency response of each frame.

MFCC:

The Mel-frequency scale is linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz. For a given frequency f we can use the following formula to compute the Mels in Hz: Mel scale is defined as:

$$m_f = 1125 \ln\left(1 + \frac{f}{700}\right)$$

where f is the actual frequency in Hz.

Mel-frequency analysis of speech is based on human perception experiments. For frequencies lower than 1 kHz human ears hear tones with a linear scale instead of logarithmic scale which is used for the frequencies higher than 1 kHz. Low frequency components of the speech signal carry more information compared to the high frequency components. Mel scaling is performed in order to place more emphasis on the low

frequency components. Since Mel filter banks are non-uniformly spaced on the frequency axis so MFCC calculation has more filters in the low frequency regions and a smaller number of filters in high frequency regions Take logarithm of Mel filter bank energies. In the final step the log Mel spectrum is converted back to time domain. The result is called the Mel frequency cestrum coefficients (MFCCs). For the given frame analysis, the cepstrum representation of the speech spectrum is a good representation of the local spectral properties of the signal.

DCT:

Discrete Cosine Transform (DCT) is performed after DCT computation the set of coefficients is called an acoustic vector. These acoustic vectors can be used to represent and recognize the voice characteristic of the speaker.

Algorithm:

The calculation of MFCC includes the following steps

1. **Framing:** frame the speech signal at the initial stage.
2. **Windowing:** after framing the most used Hamming window is used for windowing to remove the discontinuity of the signal.
3. **Fast Fourier transforms (FFT):** calculate FFT to converts the signal from time domain to frequency domain.
4. **Mel-frequency warping:** calculate the Mel filter bank energies.
5. **Discrete cosine transform (DCT):** DCT is performed to convert the Mel spectrum to the time domain and capture the coefficients is called an acoustic vector (MFCCs).