EC-0108

SPEECH SIGNAL ANALYSIS ANDSPEAKER RECOGNITION BYSIGNAL PROCESSING

Abstract:

Recent developments in Digital Signal Processing (DSP) technology make it easier for scientists to develop powerful personal computer based data acquisition and analysis systems. Speaker recognition by signal processing technique is the process of automatically recognizing who is speaking on the basis of individual information included in speech waves.

Speaker recognition methods can be divided into text independent and text dependent methods. In a text independent system, speaker models capture characterstics of somebody's speech, which show up irrespective of what one is saying. On the other hand in a text dependent system the recognition of speaker's identity is based on the persons speaking one or more specific phrases like passwords, card numbers etc. This technique makes it possible to use the speaker's voice to verify their identity and control access to services such as voice dialing, banking by telephone, telephone shopping, database access services, information services, voice mail, security control for confidential information areas, and remote access to computers .Area of artificial intelligence where machine performance can exceed human performance – using short utterances and a large number of speakers , machine accuracy often exceeds that of human beings . Thus, with the help of MATLAB we reveal the above complex phenomena in speech signal analysis and recognition (of speaker). Using the provisions provided with MATLAB, we had designed the set of coding which reflects the above phenomena to recognize a person.

Keywords: *speech recognisation, telephone shopping, artificial intelligence, MAT lab,*

Conclusion:

Digital Signal Processing is one of the advancement in the field of Electronic and Communication Engineering which has lead to several critical and intelligent application. An attempt has been made here to recognize a person based on his speech. A speech signal is a low frequency signal having range of 2Hertz to 2Khertz. This signal may be processed to recognize a particular person.

INTRODUCTION:

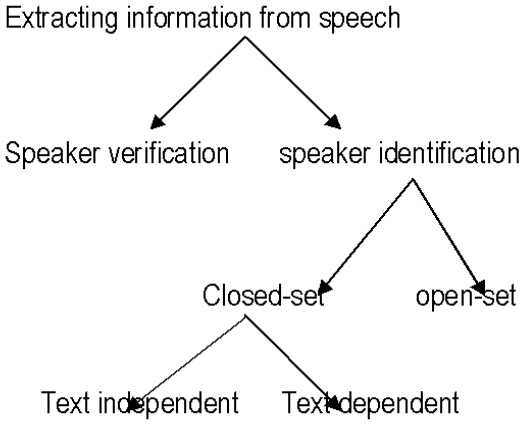
The human speech conveys different types of information. The primary type is the meaning or words, which speaker tries to pass to the listener. But the other types that are also included in the speech are information about language being spoken, speaker emotions, gender and identity of the speaker.

The goal of automatic speaker recognition is to extract, characterize and recognize the information about speaker identity. Speaker recognition is usually divided into two different branches, speaker verification and speaker identification. Speaker verification task is to verify the claimed identity of person from his voice. This process involves only binary decision about claimed identity. In speaker identification there is no identity claim and the system decides who the speaking person is.

PRINCIPLES OF SPEAKER RECOGNITION:

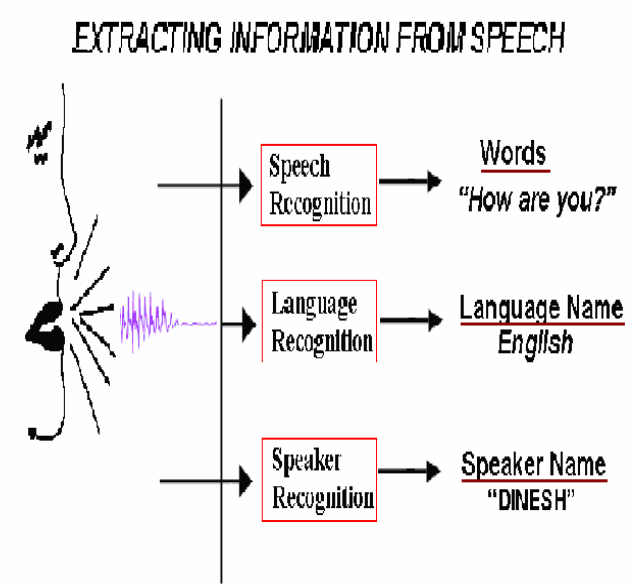
Speaker recognition can be classified into identification and verification. speaker identification is the process of determining which registered speaker provides a givenutterance. Speaker verification, on the other hand, is the process of accepting or rejecting the identity speaker's identity is based on his or her speaking one or more specific phrases, like passwords, cardnumbers, PIN codes, etc. The difference between text dependentand text independent is that in the first case the system knows the text spoken by the person while in the second case the system must be able to recognize the speaker from any text. All technologies of speaker recognition, identification and verification, text-Independent and text-dependent, each has its own advantages and disadvantages and May requires different treatments and techniques.The choice of which technology to use is application-specific. The system that we are going to develop comes under text-independent speaker identification system since its task is to identify the person who speaks regardless of what is saying. At the highest level, all speaker recognition systems contain two main modules: feature extraction and feature matching. Feature extraction is the process that extracts a small amount of data from the voice signal that can later be used to represent each speaker. Feature matching involves the actual procedure to identify the unknown speaker by comparing extracted features from his/her voice input with the ones from a set of known speakers. The identification taxonomy is shown below.

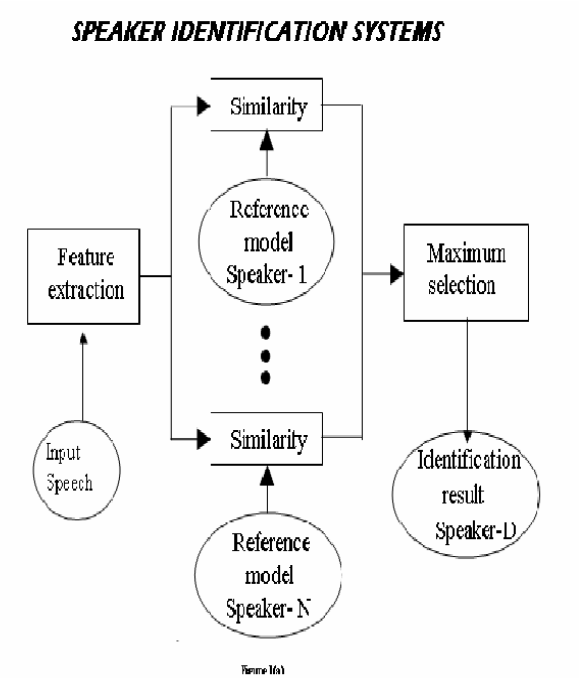
Identification Taxanomy



The process of speaker identification is divided into two main phases. During the first phase which is called speaker enrollment, speech samples are collected from the speakers, and they are used to train their models. The collection of enrolled models is also called aspeaker database. In the second phase which is also known as speaker database. In the second phase

which is also known asidentification phase, a test samplefrom an unknown speaker is compared against the speaker database. Both phases include the same first step, feature extraction, which is used to extract speaker dependent characteristics from speech. The main purpose of this step is to reduce the amount of test data while retaining speaker discriminative information. Then in the enrollment phase, these features are modeled and stored in the speaker database.





All speaker recognition systems have to serve two distinguish phases. The first one is referred to the enrollment sessions or training phase while the second one is referred to as the operation sessions or testing phase. In the training phase, each registered speaker has to provide samples of their speech so that the system can build or train a reference model for that speaker. In case of speaker verification systems, in addition, a speaker-specific threshold is also computed from the training samples. During the testing(operational) phase , the input speech is matched with stored reference model(s) and recognition decision is made. Speaker recognition is a difficult task and it is still an active research

area. Our speaker
recognition works based on the premise that a person's speech exhibits characteristics that are unique to the speaker. However this task has been challenged by the highly variant of input speech signals. The principle source of variance is the speaker himself. Speech signals in training and testing sessions can be greatly different due to many facts such as people voice change with time, health conditions (e.g. the speaker has a cold), speaking rates, etc. There are also other factors, beyond speaker variability, that present a challenge to speaker recognition technology. Examples of these are acoustical noise and variations in recording environments (e.g. speaker uses different telephone handsets).

Speech Feature Extraction:

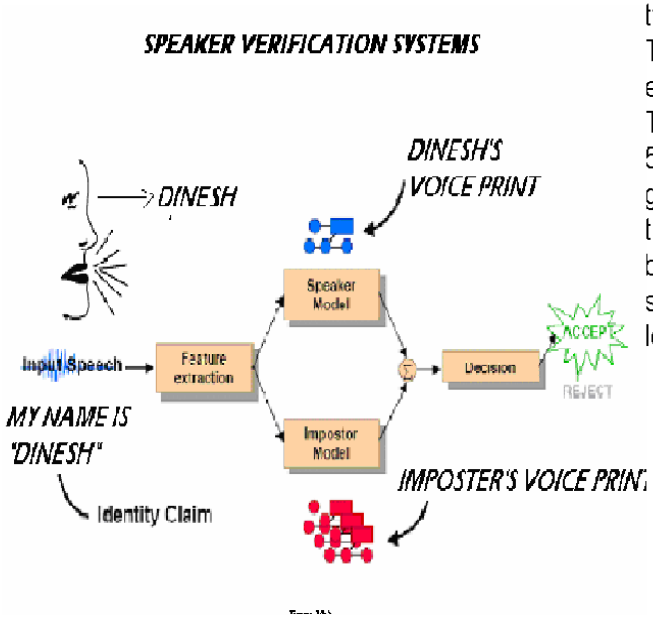
Introduction:

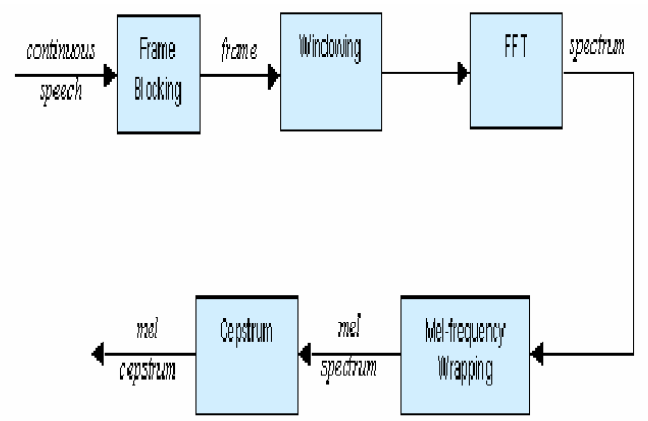
The purpose of this module is to convert the speech waveform to some type of parametric representation (at a considerably lower information rate) for further analysis and processing. This is often referred as the signal-processing front end. The speech signal is a slowly timed varying signal (it is calledquasi-stationary). An example of speech signal is shown in Figure 2.When examined over a sufficiently short period of time (between 5 and common way to characterize the100 msec), its characteristics are fairly stationary. However, overlong periods of time (on the order of1/5 seconds or more) the signal characteristic change to reflect the different speech sounds being spoken. Therefore, short-time spectral analysis is the most speech signal MFCC TECHNIQUE is perhaps the best known and most popular, and this will be used in this project. MFCC's are based on the known variation of the human ear's critical bandwidths with frequency, filters spaced linearly at low frequencies and logarithmically at high frequencies have been used to capture the phonetically important characteristics of speech. This is expressed in the Mel-frequency scale, which is a linear frequency spacing below 1000 Hz and alogarithmic spacing above 1000 Hz

.

Mel-frequency cepstrum coefficients processor:

A block diagram of the structure of an MFCC processor is given in Figure 3. The speech input is typically recorded at a sampling rate above 10000 Hz. This sampling frequency was chosen to minimize the effects of aliasing in the analog-to-digital conversion. These sampled signals can capture all frequencies up to 5 kHz, which cover most energy of sounds that are generated by humans. As been discussed previously, the main purpose of the MFCC processor is to mimic the behavior of the human ears. In addition, rather than the speech waveforms themselves, MFFC's are shown to be less susceptible to mentioned variations.





figure(3): BLOCK DIAGRAM OF MFCC PROCESSOR

1. Frame Blocking

In this step the continuous speech signal is blocked into frames of N samples, with adjacent frames being separated by M (M < N). The first frame consists of the first N samples. The second frame begins *M* samples after the first frame, and overlaps it by N - M samples. Similarly, the third frame begins 2M samples after the first frame (or Msamples after the second frame) and overlaps it by N - 2M samples. This process continues until all the speech is accounted for within one or moreframes. Typical values for N and M are N = 256 (which is equivalent to ~ 30 msec windowing and facilitate the fast radix-2 FFT) and M = 100.

2. Windowing :

The next step in the processing is to window each individual frame so as to minimize the signal discontinuities at the beginning and end of each frame. The concept here is to minimize the spectral distortion by using the window to taper the signal to zero at the beginning and end of each frame. If we define the window as W(n), 0 < = n <= N-1 where N is the number of samples in each frame, then the result of windowing is the signal Y(n) = X(n)W(n), 0 < = n < = N-1Typically the Hamming window is used which has the form

W(n)=0.54-0.46cos(2∏n/N-1), 0 < = n < = N-1

3. Fast Fourier Transform (FFT)

The next processing step is the Fast Fourier Transform, which converts each frame of N samples from the time domain into the frequency domain. The FFT is a fast algorithm to implement the Discrete Fourier Transform (DFT) which is defined on the set of N samples {xn} In general X n's are complex numbers. The resulting sequence {Xn} is interpreted as follow: the zero frequency corresponds to n = 0, positive frequencies 0 < f < Fs/2 correspond to values 1 < = n < = N/2 - 1 while negative frequencies -Fs/2 < f < 0 correspond to N/2 + 1 < = n < = N -1

Here, Fs denotes the sampling frequency. The result after this step is often referred to as spectrum or periodogram.

4. Mel-frequency Wrapping

As mentioned above, psychophysical studies have shown that human perception of the frequency contents of sounds for speech signals does not follow a linear scale. Thus for each tone with an actual frequency, f, measured in Hz, a subjective pitch is measured on a scale called the 'mel' scale .The mel-frequency scale is a linear frequency spacing below 1000 Hzand a logarithmic spacing above1000 Hz. As a reference point, the pitch of a 1 kHz tone, 40 dB above the perceptual hearing threshold, is defined as 1000 mels. Therefore we can use the following approximate formula to compute the mels for a given frequency f in Hz:

mel(f) = 2595 * log (1 + f / 700)

5. Cepstrum

In this final step, we convert the log mel spectrum back to time. The result is called the mel frequency cepstrum coefficients (MFCC). The cepstral representation of the speech spectrum provides a good representation of the local spectral properties of the signal for the given frame analysis. Because the melspectrum coefficients (and so theirlogarithm) are real numbers, we can convert them to the time domain using the Discrete Cosine Transform (DCT). Therefore if we denote those mel power spectrum coefficients that are the result of the last step are Sk, k =1,2,----,K

FEATURE MATCHING TECHNIQUE:

Vector Quantization:

Figure 5 shows a conceptual diagram to illustrate this recognition process. In the figure, only two speakers and two dimensions of the acoustic space are shown. The circles refer to the acoustic vectors from the speaker 1 while the triangles are from the speaker 2. In the training phase, a speaker-specific VQ code book is generated for each known speaker byclustering his/her training acousticvectors. The result codewords (centroids) are shown in Figure 5 by black circles and black triangles for speaker 1 and 2, respectively. The distance from a vector to the closest codeword of a codebook is called a VQ-distortion. In the recognition phase, an input utterance of an Unknown voice is vector-quantized_ using each trained codebook and the total VQ distortion is computed. The speaker corresponding to the VQ codebook with smallest total distortion is identified.

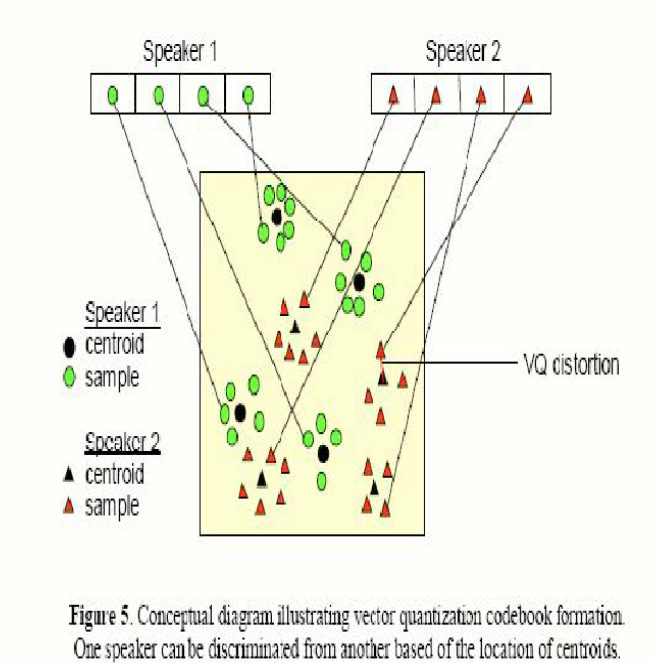


Figure 5. Conceptual diagram illustrating vector quantization codebook formation. One speaker can be discriminated from another based of the location of centroids.

SPEECH PROCESSING USING MATLAB:

In this phase you are required to write a Matlab function that reads a sound file and turns it into a sequence of MFCC (acousticvectors) using the speech processing steps .

The Matlab functions that you would need to use are: wavread, hamming, fft, dct and melfb(supplied function).

STEPS TO BE FOLLOWED FOR FEATURE EXTRACTION:-

* Read a sound file into Matlab Check it by playing the sound file in Matlab using the function: sound. What is the sampling rate? What is the highest frequency that the recorded sound can capture with fidelity? With that sampling rate, how many msecs of actual speech are contained in a block of 256 samples?

* Plot the signal to view it in the time domain. It should be obvious that the raw data in the time domain has a very high amount of data and it is difficult for analyzing the voice characteristic. Now cut the speech signal (a vector) into frames with overlap. The result is a matrix where each column is a frame of N samples from original speech signal. Applying the steps: Windowing and FFT to transform the signal into the frequency domain. This process is used in many different applications and is referred in literature as Windowed Fourier Transform (WFT) or Short-Time Fourier Transform (STFT). The result is often called as the spectrum or periodogram.

* After successfully running the preceding process, what is the interpretation of the result? Compute the power spectrum and plot it out using the images command. Note that it is better to view the power spectrum on the log scale. Locate the region in the plot that contains most of the energy. Translate this location into the actual ranges in time (msec) and frequency (in Hz) of the input speech signal.

* Compute and plot the power spectrum of a speech file using different frame size: for example N = 128, 256 and 512. In each case, set the frame increment M to be about N/3.

* The last step in speech processing is converting the power spectrum into mel- frequency cepstrum coefficients.The supplied function melfb facilitates this task.

*Plot out the mel-spaced filter bank. Compare the behavior of this filter bank with the theoretical

part. Finally, put all the pieces together into a single Matlab function, mfcc, which performs the MFCC processing.

* Compute and plot the spectrum of a speech file before and after the mel- frequency wrapping step.

STEPS TO BE FOLLOWED FOR FEATURE MATCHING:

* To inspect the acoustic space (MFCC vectors) we can pick any two dimensions (say the 5th and the 6th) and plot the data points in a 2D plane. Use acoustic vectors of two different speakers and plot data points in two different colors. Do the data regions from the two speakers overlap each other? Are they in clusters?

* With the help of a Matlab function, vqlbg that trains a VQ codebook using the LGB algorithm described before. Use the supplied utility function disteu to compute the pair wise Euclidean distances between the codewords and training vectors in the iterative process