International Journal of Electronics, Communication & Soft Computing Science and Engineering
ISSN: 2277-9477, Volume 2, Issue 6

Speech Recognition using Digital Signal Processing

Mr. Maruti Saundade Mr.Pandurang Kurle

Abstract: **- Speech recognition methods can be divided into *text-independent* and text dependent methods. In a text independent system, speaker models capture characteristics of somebody's speech, which show up *irrespective of what one is saying*. In a text-dependent system, on the other hand, the recognition of the speaker's identity is based on his or her *speaking one or more specific phrases*, like passwords, card numbers, PIN codes, etc. This paper is based on text independent speaker recognition system and makes use of Mel frequency cepstrum coefficients to process the input signal and vector quantization approach to identify the speaker. The above task is implemented using MATLAB. Digital Signal Processing (DSP) is one of the most commonly used hardware platform that provides good development flexibility and requires relatively short application development cycle. DSP techniques have been at the heart of progress in Speech Processing during the last 25years.Simultaneously speech processing has been an important catalyst for the development of DSP theory and practice. Today DSP methods are used in speech analysis, synthesis, coding, recognition, enhancement as well as voice modification, speaker recognition, language identification. Speech recognition is generally computationally-intensive task and includes many of digital signal processing algorithms.**

I. INTRODUCTION

The objective of human speech is not merely to transfer words from one person to another, but rather to communicate, understanding a thought, concept or idea. The final product is not the words or phrases that are spoken and heard, but rather the information conveyed by them. In computer speech recognition, a person speaks into a microphone or telephone and the computer listens. Speech processing is the study of speech signals and the processing methods of these signals. The signals are usually processed in a digital representation. So speech processing can be regarded as a special case of digital signals processing applied to speech signals. Automatic Speech Recognition technology has advanced rapidly in the past decades.

Speech recognition is a vast topic of interest and is looked upon as a complex problem. In a practical sense, speech recognition solves problems, improves productivity, and changes the way we run our lives. Reliable speech recognition is a hard problem, requiring a combination of many techniques; however modern methods have been able to achieve an impressive degree of accuracy [1]. Real-time digital signal processing made considerable advances after the introduction of specialized DSP processors.

II. LITERATURE SURVEY

Every speech recognition application is designed to accomplish a specific task. Examples include: to recognize the digits zero through nine and the words “yes” and “no” over the telephone, to enable bedridden patients to control the positioning of their beds, or to implement a VAT (voice-activated typewriter). Once a task is defined, a speech recognizer is chosen or designed for the task.

Recognizers fall into one of several categories depending upon whether the system must be “trained” for each individual speaker, whether it requires words to be spoken in isolation or can deal with continuous speech, whether its vocabulary contains a small or a large number of words, and whether or not it operates with input received by telephone. Speaker dependent systems are able to effectively recognize speech only for speakers who have been previously enrolled on the system. The aim of speaker independent systems is to remove this restraint and recognize the speech of any talker without prior enrolment. When a speech recognition systems requires words to be spoken individually, in isolation from other words, it is said to be an isolated word system and recognizes only discrete words and only when they are separated from their neighbours by distinct interword pauses. Continuous speech recognizers, on the other hand, allow a more fluent form of talking. Large-vocabulary recognizers are defined to be those that have more than one thousand words in their vocabularies; the others are considered small-vocabulary systems. Finally, recognizers designed to perform with lower bandwidth waveforms as restricted by the telephone network are differentiated from those that require a broader bandwidth input.[4]Digital signal processors are special types of processors that are different from the general ones. Some of the DSP features are high speed DSP computations, specialized instruction set, high performance repetitive numeric calculations, fast and efficient memory accesses, special mechanism for real time I/O, low power consumption, low cost in comparison with GPP. The important DSP characteristics are data path and internal architecture, specialized instruction set, external memory architecture, special addressing modes, specialized execution control, specialized peripherals for

DSP.[6]At the beginning of each implementation process is an important decision: the choice of appropriate hardware platform on which a system of digital signal processing is operated. It is necessary to understand the hardware aspects in order to implement effective optimized algorithms. The above hardware aspects imply several criteria for choosing the appropriate platform: It is preferable to choose a signal processor than a processor for general use. It may not be decisive a processor frequency, but its effectivenes.DSP tasks require repetitive numeric calculations, alternation to numeric, high memory bandwidth sharing, real time processing. Processors must perform these tasks efficiently while minimizing cost, power consumption, memory use, development time. To properly select a suitable architecture for DSP and speech recognition systems, it is necessary to examine well the available supply and to become familiar with the hardware capabilities of the "candidates". In the decision it is necessary to take into account some basic features, in which processors from different manufacturers differ. Most DSPs use fixed-point arithmetic, because in real world signal processing the additional range provided by floating point is not needed, and there is a large speed benefit and cost benefit due to reduced hardware complexity. Floating point DSPs may be invaluable in applications where a dynamic range is required. To implement speech recognition different algorithms like Linear predictive coding, Advantages of MFCC (Mel Frequency Cepstrum coefficient) methods are it is capable of capturing the phonetically important characteristic of speech, band limiting can easily be employed to make it suitable for telephone application.

III. FUNCTIONAL DESCRIPTION

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 given utterance. Speaker verification, on the other hand, is the process of accepting or rejecting the identity claim of a speaker. Figure shows the basic structures of speaker identification And verification systems. At the highest level, all speaker recognition systems contain two main modules (refer to 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.

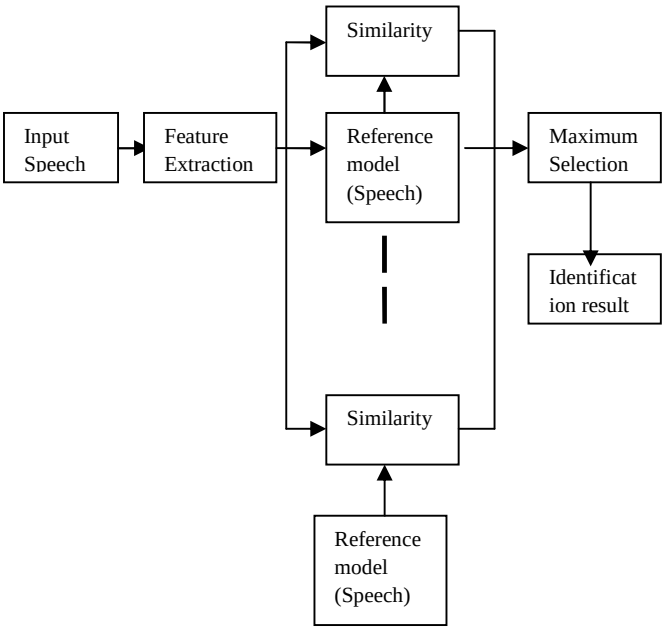
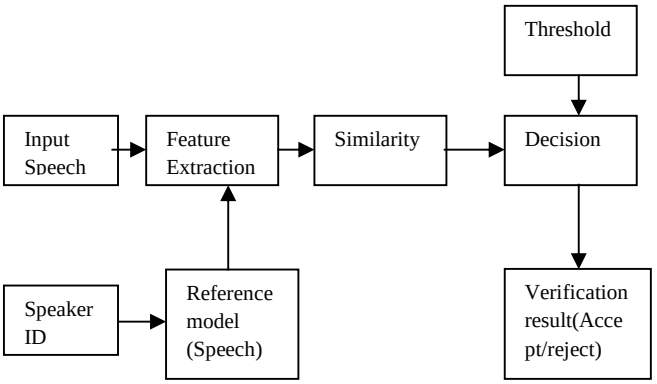


Figure a: (speaker identification/recognition)



b: Speaker verification (Speech verification)

Figure 1.Basic structures of speaker recognition process

All speaker recognition systems have to serve two distinguishes 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 phase (Figure 1), the input speech is matched with stored reference model and recognition decision is made.

SPEECH FEATURE EXTRACTION:

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 called quasi-stationary). An example of speech signal is shown in Figure 2. When examined over a sufficiently short period of time (between 5 and 100 ms), its characteristics are fairly stationary. However, over long periods of time (on the order of 1/5 seconds or more) the signal characteristic change to reflect the different speech sounds being spoken. Therefore, short-time spectral analysis is the most common way to characterize the speech signal.

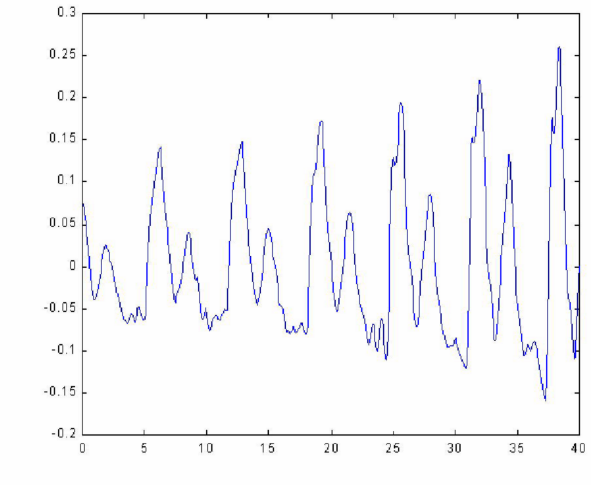


Figure 2. An example of speech signal

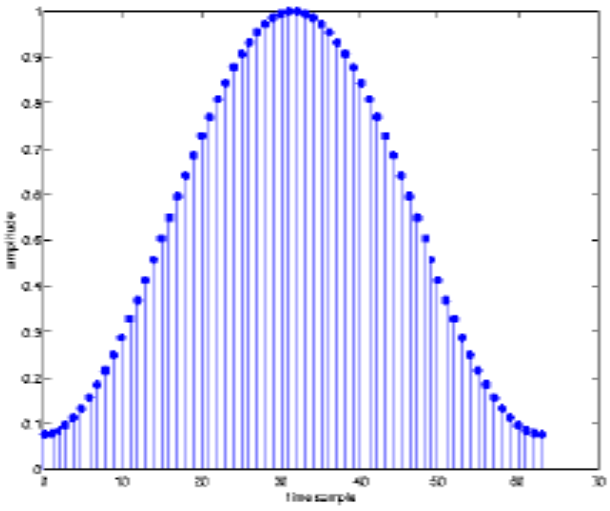


Figure 3: Speech signal in time domain

MEL-FREQUENCYCEPSTRUM COEFFICIENTS PROCESSOR:

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 a logarithmic spacing above 1000 Hz.

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.

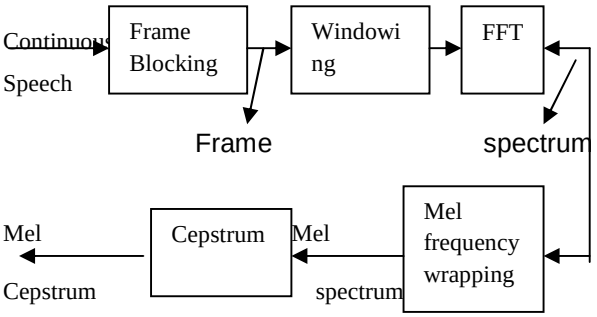


Figure 3. Block diagram of the MFCC processor

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 M samples after the second frame) and overlaps it by $N - 2M$ samples. This process continues until all the speech is accounted for within one or more frames. Typical values for N and M are $N = 256$ (which is equivalent to ~ 30 m sec windowing and facilitate the fast radix-2 FFT and $M = 100$

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 y1(n)=x1(n)w(n), 0<n<N-1 Typically the *Hamming* window is used, which has the form

W (n) = 0.54-0.46 Cos (2πn/N-1), $0 \le n \le N-1$

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 implement the Discrete Fourier Transform (DFT) which is defined on the set of *N* samples {*x (n)*}, as follow:

$$X_n = \sum_{k=0}^{N-1} x_k e^{-2\pi j k n / N}, \qquad n = 0,1,2,\ldots,N-1$$

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 linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz. As a reference point, the pitch of a 1 kHz tone, 40 dB above the perceptual hearing threshold, is defined as 1000 Mels.

Mel (f) =2595*$\log_{10}$ (1+f/700)

One approach to simulating the subjective spectrum is to use a filter bank, spaced uniformly on the Mel scale. That filter bank has a triangular band pass frequency response, and the spacing as well as the bandwidth is determined by a constant Mel frequency interval.

CEPSTRUM

In this final step, the log Mel spectrum is converted back to time. The result is called the Mel frequency cepstrum coefficients (MFCC). The cepstrum representation of the speech spectrum provides a good representation of the local spectral properties of the signal for the given frame analysis. Because the Mel spectrum coefficients (and so their logarithm) 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

$$\tilde{S}_k, \quad k = 1,2,\ldots,K$$

We can calculate the MFCC as

$$\tilde{c}_n = \sum_{k=1}^{K} (\log \tilde{S}_k) \cos\left[n\left(k - \frac{1}{2}\right)\frac{\pi}{K}\right], \qquad n = 1,2,\ldots,K$$

FEATURE MATCHING

The state-of-the-art in feature matching techniques used in speaker recognition include Dynamic Time Warping (DTW), Hidden Markov Modeling (HMM), and Vector Quantization (VQ). In this paper HMM used due to ease to implementation and high accuracy.

CONCLUSION

Even though much care is taken it is difficult to obtain an efficient speaker recognition system since this task has been challenged by the highly variant input speech signals. The principle source of this 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. Because of all these difficulties this technology is still an active area of research.

REFERENCES

1) Performances of speech Recognition Devices and acoustics, speech and signal IEEE international conference

2) Performances of Isolated word Recognition system acoustics speech and signal IEEE international conference

3) L.R. Rabiner and B.H. Juang, Fundamentals of Speech Recognition, Prentice-Hall, Englewood Cliffs, N.J., 1993.

4) L.R Rabiner and R.W. Schafer, Digital Processing of Speech Signals, Prentice-Hall, Englewood Cliffs, N.J., 1978.