A Study on Speech Recognition Technology

Arul.V.H[#1], Dr. Ramalatha Marimuthu[#2]

[#1] *Electronics & Communication Engineering Department, University of Calicut, India*

[#2] *Information Technology , kumaraguru College of Technology, Coimbatore, India*

[2]vharul@gmail.com

[2]ramalatha.marimuthu@gmail.com

Abstract— **The paper highlights a brief study on speech recognition technology, describing the various processing stages and results and also some primary applications as well. Following this review some of the vital strengths and speech processing steps will also discuss.**

Keywords— **Signal processing, sampling, cepstrum analysis**

I. INTRODUCTION

Speech is one of the attractive modality for human machine interaction. To the listener it conveys a wide range of information. Information likes, spoken language, emotions, gender and generally the proper identification. The main goal of the system is to recognize the word spoken by extracting, characterizing and finally recognizing the signal.

Recognition of speaker can be classified into two phases. First is speaker identification and second be the speaker verification. Speaker identification is the task of determining who is talking from a set of known voices or speakers [1]. If the speaker claims to be of a certain identity and voice is used to verify this claim, is called verification. Speaker recognition systems fall into two categories- text dependent and text independent. In a text dependent, system has a prior knowledge to the text to be spoken. But in text independent, no prior knowledge is needed. It is considered as more difficult and more flexible one. ie, text independent verification is a process of verifying the identity of the speech without any constraint. Several methods are used for recognizing the speaker over the last few decades. Certain approaches like spectrogram comparisons, to simple and time warping methods brings handy result. Above these statistical pattern recognition approaches like neural networks and Hidden Markov Models (HMM).

In this paper we discuss the areas of speech processing like sampling, bit resolution, identification of speech signals and some indications of performance. Following to this also discuss about the strength and weakness and also the applications.

II. SAMPLING FREQUENCY AND BIT RESOLUTION

Audio can be stored in either compressed or uncompressed format. Uncompressed audio format is the format which your module wants to work with. Sample represents how loud the audio at a single point. By shannon's sampling theorem the highest frequency component which is fed to the sampled system be equal to or less than half the sampling frequency. $F_s \geq 2Fm$. By controlling sampling frequency the quality can control more efficiently. ie, higher the sampling rate better the quality.

Sampling frequency is the parameter that controls the sampling process. Speech signal has frequency component in the audio frequency range (20 Hz-20 KHz) electromagnetic spectrum. The standard sampling frequency to sample the entire audio range is 44.1 KHz, because 20 KHz is the maximum frequency component. So 44.1 KHz is too high value and the useful information is about 7.5 KHz≈ 8 KHz. keeping this result, an optimum value of sampled frequency is around 16 KHz. ie, most of the speech signal has a frequency component up to 8 KHz. On communication, bandwidth is a vital resource. The speech signal then passes through an anti aliasing low pass filter with cutoff frequency 3.3KHz which is sampled at 8 KHz. Thereby speech signal collected over telephone network will have a bandwidth of 4 KHz. information up to this be sufficient for speech. Speech signal sampled at 8KHz be considered as narrow band speech and 16KHz as wide band speech, which is considered as an optimal frequency for speech.

A. *Digitization process*

It is the representation of sound, image, or a signal by a discrete set of points or samples. Simply it means capturing an analog signal in digital form. The term is generally used when diverse forms of information in to a single binary code known as ***bits***. The number of bits used for storing each sample of speech can be considered as ***bit resolution***. The number of bits/sample depends up on the number of quantization levels used during ADC. ***Quantization*** is a process by which the samples are rounded to a fixed set of numbers. If more

number of quantization levels, better will be the information preserved in the digitized form. All the speech signal processing applications are using 16 bits/sample as bit resolution. The most important thing is that if we are listening to a speech frame having a 1 bit resolution, we can still make out the information present in the frame with quantization noise. ie, information lies in the sequence ,not only by the amplitude value of samples.

III. IDENTIFICATION OF SPEECH

Speech is an acoustic signal produced from a speech production system. In speech production, nature of speech is identified by excitation phenomenon. Based on the input excitation it classifies in to three levels. One if, the input excitation is nearly periodic known as voiced speech and other will be in the random noise like nature (unvoiced speech) otherwise in no excitation state (silence region).

A. *Voiced speech*

As on the fig 1 If the input excitation is nearly periodic impulse sequence the corresponding speech looks like visually nearly periodic and is called as voiced speech.

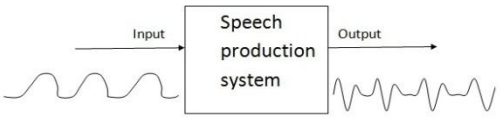

Fig 1: Speech system

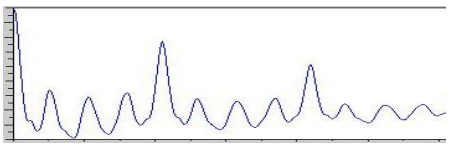

Fig 2: Autocorrelation sequence

The periodicity can be measured by autocorrelation analysis and the period is termed as pitch period. ***Pitch period*** is defined as the largest peak n the autocorrelation sequence from the beginning. Fig 2 shows the sample autocorrelation sequence of a speech frame. Voice speech is periodic in nature and there should be some frequency and harmonics in the spectrum of speech. From the spectrum as shown in fig: it is clear that the frequency is repeating after a particular interval. This will be the harmonic structure. The signal having no periodicity will be called as unvoiced speech. In unvoiced speech no harmonic structure is present. The region i identified by visual perception and automatic approach. In an intelligent speech the duration of the silence region is an important concern. But from the energy point of view it is less concerned. The fig shows a sample speech signal of word 'HelloHello'. The space in between the speech signal is the silent region and the set of varying signal be the voice region.

Fig 3: Spectrum

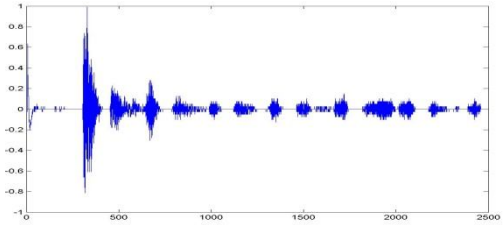

Fig 4: Speech signal waveform for 'HelloHello'

IV. SPEECH PROCESSING

Speech signal is non stationary in nature. Speech is produced from time varying vocal tract system with time varying excitation. Many tools like scilab, originlab, matlab are used in signals and systems for signal processing. In all the system is a time invariant system and time invariant excitation. In speech processing the most fundamental parameter is total energy computation. For stationary signals the total energy relation becomes

$$E_T = S^2(n) \quad n= -\infty \text{ to } \infty$$

But for speech processing it is not useful because of the time varying nature. Therefore a signal must be stationary when it is viewed in blocks of length 10-30msec. Such processing technique is known as ***short term processing*** (STP). This can be performed in either time or frequency domain. Processing domain is strictly depends on the information from the speech signal that are look into. Parameters like short term energy, short term zero crossing rate, short term auto correlation are computed in time domain. Speech signal is a combination of voiced, unvoiced or silenced region because of the nature of production. The energy concerned with the speech signal is large when compared with the rest of the frame. The relation for finding the short term energy can be derived from total energy relation.

$$E_T = S^2(n) \quad n= -\infty \text{ to } \infty$$

The short term energy, it weights a signal with a window function.

$$S_w(n) = S(m).\ W(n-m)$$

Total energy,

$$E_n = \sum(s(m).w(n-m)) \quad m= -\infty \text{ to} \infty$$

Where 'n' is the shift/rate in number of samples (normally half the frame size).

The second parameter be the ***short term zero crossing*** (ZCR), gives indirect information about the frequency content of the signal. If the numbers of zero crossing are more it is considered as high frequency otherwise low frequency. For stationary signal ZCR become,

$$Z = \sum \left| \text{sgn}(s(n) - \text{sgn}\ (s(n-1)) \right| \quad n= -\infty \text{ to} \infty$$

Considering a non- stationary signal like speech, ZCR is

$$Z(n) = 1/2N \sum S(m).\ W(n-m) \quad m= 0 \text{ to } N-1$$

The presence of factor 2 is to indicate that there will be two zero crossing per signal.

A. Windowing

Speech is considered as non-stationary signal, the property of speech mostly remain invariant around a time period of 10-100ms. For a short window time traditional signal processing approach can be followed. The short window of signal can be called as frame. During processing the shape is not at all a crucial factor, but usually selects soft windows like hanning, hamming, triangle etc. The reason for choosing such window is that the sideband lobes are substantially smaller when compares to rectangular.

The unnatural discontinuities and distortion in the spectrum can be avoided by performing window operation. In speaker recognition the most widely used window shape is hamming window. The primary reason for using such windows is to decreasing the high frequency components in each frame which is due to slicing of the signal. Windows can be used in some other applications like spectral analysis, filter design and for data compressions also.

$$W_H(n)= .54-.46\cos(2n\pi/N-1)$$

B. Cepstral Analysis

The main objective of cepstral analysis is to separate the speech into it's source and system components without any prior knowledge about the system. As per the source filter theory of speech production, voice sounds are produced by exciting time varying system with periodic impulse and unvoiced by random noise. The obtained speech signal can be considered as the convolution of respective excitation sequence and vocal tract filter.

$$S(n)= e(n)*h(n)$$

In frequency domain,

$$S(w)=E(w)\ H(w)$$

The above equation informs that the excitation and system components in the frequency domain for the convolution sequence as same in time domain. For deconvolution, multiplication of the two components in the frequency domain has to be converted into linear combination of the two. This is for we use cepstral analysis. Let the magnitude spectrum of the speech be,

$$\left| s(w) \right| = \left| E(w) \right| \left| H(w) \right|$$

And linear combination of E(w) and H(w) becomes,

$$\log \left| s(w) \right| = \log \left| E(w) \right| + \log \left| H(w) \right|$$

log operation converts '*' into '+' in frequency domain. And the separation of these two components can be achieved by taking IDFT of linearly combined log spectra of excitation and system. In simpler explanation, IDFT of log spectra transforms into quefrency/ cepstral domain which is similar to time domain.

$$C(n)=\text{IDFT} \log (\left| s(w) \right|)$$

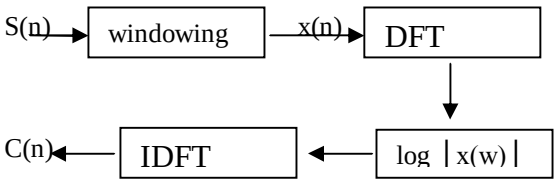

Fig 5: Cepstral functional block

Fig 5 shows, s(n) be the voiced frame and x(n) is window frame function. x(n) is obtained by multiplying s(n) with a window function. C(n) be the computed frame of s(n). Such spectrum contains vocal tract components which are linearly combined frames. From the block diagram it is clear that c(n) is derived from the log magnitude. So it is symmetrical in the quefrency domain.

For an unvoiced frame, the cepstrum computation varies according to upper/ lower quefrency region. The fast varying nature of the cepstrum towards the upper quefrency represents the excitation characteristics of the speech segment where as the variations in the lower quefrency region is mainly due to the vocal tract characteristics. Hence for extracting both the features independently from the frequency domain, ***liftering*** operation is performed. Liftering is just similar to the normal filtering operation. The desired region is selected by multiplying the entire cepstrum with a rectangular window function. The slow varying vocal tract characteristics can be estimated by low time liftering.

$$C_e(n)=W_e[n]c(n)$$

The extraction process is by applying DFT on the lower time filter sequence results a log magnitude spectrum.

$$\log (\left| H(w) \right|)= DFT[C_e(n)]$$

Pitch estimation can be performed by high time liftering. This can be obtained by window function. Pitch can be estimated at any instants corresponds to the highest peak in the high time liftered spectrum.

V. MERITS AND APPLICATIONS

Applications of speaker recognition technology are quiet growing and varied field. Many applications like in the fields of law enforcement, speech data management voice web customization and transaction authentication the impact of speaker recognition is well observed[1][2][3]. The primary strength of speaker verification technology is that the signal is natural and un obstructive to produce the signal. This is used for applications with remote users and speech interface tasks. On speaker verification side it is having high accuracy and low computational requirements. Robustness to channel variability is the biggest challenge to the current systems.

REFERENCES

[1] S. Furui. Recent advances in speaker recognition AVBPA97,pp 237-251,1997.

[2] J.P.Campbell, "*Speaker recognition; A tutorial*", Proceedings of the IEEE,vol.85,pp.1437-1462, September 1997

[3] Special issue on speaker Recognition,Digital signal Processing.vol 10,January 2000

[4] Douglas A Reynolds, "*An overview of automatic speaker recognition technology*" MIT Lincon Laboratory,Lexington,MA USA

[5] D.A.Reynolds and L.P.heek, "*Speaker verification: From research to reality*"ICASSP 2001

[6] Tommie Kinnunen, Haizhou Li,:An overview of text independent speaker recognition: from features to supervectors, speech communication 52 (2010)

