Analysis of Noise Filtering and Speaker recognition

Pragati Gaur

Abstract— In this advance technical era being able to speak to your personal computer, and have it recognize and understand what you say, would provide a comfortable and natural form of communication. It would reduce the amount of typing you have to do, leave your hands free, and allow you to move away from the terminal or screen. You would not even have to be in the line of sight of the terminal. It would also help in some cases if the computer could tell who was speaking. If you want to use voice as a new medium on a computer workstation, it is natural to explore how speech recognition can contribute to such an environment. Here we will review the state of speech and speaker recognition, focusing on current technology applied to personal workstations. The objective of this paper is to provide some explanation of the speech and speaker recognition data input of the user. An algorithm which efficiently determines the optimum coordination with H.M.M has been successfully designed with the help of MATLAB. Authentication of the user can be determined by the threshold value being set by the standard variance.

Index Terms— Minimum 7 keywords are mandatory, Keywords should closely reflect the topic and should optimally characterize the paper. Use about four key words or m, Speech, Speaker,Filter analysis, and MATLAB

---

1 INTRODUCTION

Speech recognition, the ability to identify spoken words, and speaker recognition, the ability to identify who is saying them, becoming commonplace applications of speech processing technology.Limited forms of speech recognition are available on personal workstations. Currently there is much interest in speech recognition, and performance is improving. Speech recognition has already proven useful for certain applications, such as telephone voice-response systems for selecting services or information, digit recognition for cellular phones, and data entry while walking around a railway yard or clambering over a jet engine during an inspection. Nonetheless, comfortable and natural communication in a general setting (no constraints on what you can say and how you say it) is beyond us for now, posing a problem too difficult to solve.[1] Fortunately, we can simplify the problem to allow the creation of applications like the examples just mentioned. Some of these simplifying constraints are discussed in the next section. Speaker recognition is related to work on speech recognition. Instead of determining what was said, you determine who said it. Deciding whether or not a particular speaker produced the utterance is called verification, and choosing a person's identity from a set of known speakers is called identification. The most general form of speaker recognition (text-independent) is still not very accurate for large speaker populations, but if you constrain the words spoken by the user (text-dependent) and do not allow the speech quality to vary too wildly, then it too can be done on a workstation Early attempts to design systems for automatic speech recognition were mostly guided by the theory of acoustic-phonetics, which describes the elements of speech and how they are realized to form a spoken language. In 1952, Davis et al. of Bell Laboratories built a system for isolated digit recognition for a single speaker, using the spectral resonances during vowel regions of each digit. In 1956, Olson and Belar of RCA Laboratories tried to recognize ten syllables of a single talker. At MIT Lincoln Laboratory, Forge and Forge built a speaker-independent ten-vowel recognizer in 1959, using time-varying estimates of the vocal tract resonance 2 Procedure for Paper Submission. 1960s, with emphasis on building a special hardware, severalJapanese laboratories also demonstrated their progress. Most notable among them were the vowel recognizer of Suzuki and Nakata of the Radio Research Lab in Tokyo, the phoneme recognizer of Sakai and Doshita of Kyoto University (noting the use of a speech segmenter to allow analysis and recognition of speech in different portions of the signal), and the digit recognizer of NEC Laboratories. One significant remark to be made is the year 1959 when Fry and Denes, at University College in England, attempted a phoneme recognizer to recognize four vowels and nine consonants. They incorporated statistical information about allowable phoneme sequences in English to enhance the overall phoneme recognition accuracy for words consisting of two or more phonemes.[2] This perhaps marked the first use of statistical syntax in automatic speech recognition The work of Martin's team at RCA Laboratories and that of Vintsyuk in the Soviet Union in the 1960s have particularly important implications on the research and development of automatic speech recognition. Martin recognized the need to deal with the non uniformity of time-scale in speech events and suggested realistic solutions, including detection of utterance endpoints, which greatly enhanced the reliability of the recognizer performance. Vintsyuk proposed the use of dynamic programming for time-alignment between two utterances in order to derive a meaningful matching score. Although his work was largely unknown to the West then, it appears to have pre-

ceded that of Sakoe and Chiba, as well as others who proposed more formal methods in speech pattern matching, generally known as dynamic time warping.

2 FOUNDATION OF SPEECH RECOGNITION AND UNDERSTANDING

The Spoken language processing encompasses a broad range of technical challenges, including recognition of words and phrases in the speech signal, extraction of keywords or key phrases in the utterance, and understanding of the spoken utterance for the machine to take actions. Conversation between people can take many different forms, many of which may be beyond the scope of the current scientific interest. For example, a casual conversation between two people can drift over an unbounded domain with no end result anticipated. We will not address this category of scenarios.[4] We will, however, assume that the common goal in speech recognition and understanding is to identify an important message, out of a finite set of possibilities, conveyed in the spoken utterance.

2.1 Basics of Linguistics and Acoustic Phonetics

Most of the classical speech-recognition research was based on the identification paradigm as discussed above. It requires extensive understanding of the properties of the object (i.e., the speech sound). It, thus, depends on and makes use of, almost exclusively, the acoustic-phonetic theory, which aims at building a framework for understanding speech by a human.[5]

Phoneticians and linguists decompose a spoken language into elements of linguistically distinctive sounds—the phonemes. The number of phonemes in a language is often a matter of judgment and is not invariant to different linguists. Phonemes are determined and taxonomically classified according to their corresponding articulator configurations. For example, a vowel is produced by exciting a vocal tract of an essentially fixed shape with quasi-periodic pulses of air, caused by the vibration of the vocal cords. Front vowels (/i/, /I/, /e/, and /æ/) are vowels produced with a tongue hump in the front portion of the vocal tract. Other phoneme categories include diphthongs, semivowels, nasals, stops, fricatives, affricates, and whisper. As in many classical studies, the taxonomy was established for a systematic investigation of the properties of the "element" of speech sounds. Such properties of sounds are often referred to as acoustic-phonetic features.[6] An alternative way to classify the phonemes is to use the broad phonetic class according to key acoustic-phonetic feature dimensions.

2.2 Factors affecting speech recognition

Modern speech recognition research began in the late 1950s with the advent of the digital computer. Combined with tools to capture and analyze speech, such as analog-to-digital converters and sound spectrograms, the computer allowed rearchers to search for ways to extract features from speech that allow discrimination between different words. The 1960s saw advances in the automatic segmentation of speech into units of linguistic relevance (such as phonemes, syllables, and words) and on new pattern-matching classification algorithms. By the 1970s, a number of important techniques essential to today's state-of-the-art speech recognition systems had emerged, spurred on in part by the Defense Advanced Research Projects Agency speech recognition project. These techniques have now been refined to the point where very high recognition rates are possible, and commercial systems are available at reasonable prices.

3 TECHNOLOGY COMPONENTS OF AUTOMATIC SPEECH RECOGNITION AND UNDERSTANDING

Most computer systems for speech recognition include the following five components.

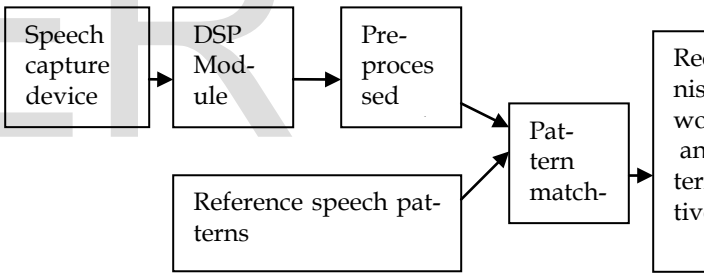

Fig. 1. Components of a typical speech recognition system

This usually consists of a microphone and associated analog-to-digital converter, which digitally encodes the raw speech wave form. The DSP module performs endpoint (word boundary) detection to separate speech from non speech, converts the raw waveform into a frequency domain representation, and performs further windowing, scaling, filtering, and data compression. The goal is to enhance and retain only those components of the spectral representation that are useful for recognition purposes, thereby reducing the amount of information that the pattern-matching algorithm must contend with. A set of these speech parameters for one interval of time (usually 10-30 milliseconds) is called a speech frame.

Here, the preprocessed speech is buffered for the recognition algorithm. Stored reference patterns can be matched against the user's speech sample once it has been prepro-

cessed by the DSP module. This information is stored as a set of speech templates or as generative speech models. The algorithm must compute a measure of goodness-of-fit between the preprocessed signal from the user's speech and all the stored templates or speech models. A selection process chooses the template or model (possibly more than one) with the best match.

4 SPEECH SIGNAL MODELLING

The statistical method, as discussed in the previous sections, requires that a proper, usually parametric, distribution form for the observations be chosen in order to implement the MAP decision. Using the task of isolated-word speech recognition as an example, we have to determine the distribution form for the speech utterance of each word before we employ an estimation method to find the values of the parameters.

Speech is a time-varying signal. When we speak, our articulator apparatus (the lips, jaw, tongue, and velum) modulates the air pressure and flow to produce an audible sequence of sounds. Although the spectral content of any particular sound in speech may include frequencies up to several thousand hertz, our articulator configuration (the vocal-tract shape, the tongue movement, etc.) often does not undergo dramatic changes more than ten times per second. During the short interval where the articulator configuration stays somewhat constant, a region of "quasi-stationary" in the produced speech signal can often be observed. This is the first characteristic of speech that distinguishes it from other random, non stationary signals. The temporal variation is manifested in several ways: the timing of voicing onsets, the vowel duration, etc.

Furthermore, speech is not a memory less process due to articulator and phototoxic constraints. According to the phonological rule of a language, there is a certain dependency between sound pairs that occur in sequence; some occur more often than others, while some are simply nonexistent in the language. The speech model or distribution needs to have provisions to permit characterization of this sequential structure, ideally in a manner consistent with the slowly varying nature (i.e., "quasi-stationary") of the speech signal.

5 RESULT AND DISCUSSIONS

To examine the effect of speech and speaker recognition a program in MATLAB has been written. There are few assumptions which has been considered as

1.Frequency of operation is 4 KHz.

2.No noise is introduced through internal or external source.

3.Already *.dat file is stored as a data base for the system.

4.Only absolute value is considered.

5.Value of zero and once matrix is constant.

6.Only 2 sec are allotted for the execution of the program and checking of data base.

7.Threshold can be set by using the value of's'.

8.Constant supply for MIC and computer system.

9. FILTER SPECIFICATION:

Sampling frequency(Hz)=48000Hz
Nyquist frequency(Hz)=24000Hz
Bandpass frequency range (Hz)=700Hz-12000 Hz
Passband frequency range(Hz)=12000 Hz-18000Hz
➢ Stopband frequency range(Hz)= 14000Hz-16000Hz (for IIR elliptical filter)

5.1 Experiment

During the first experiment a program has been written in MATLAB to verify the characteristic of speech and speaker recognition. During the program first we enter the name to be verified, which should be done within two second. There is a choice of entering the data voice again also, by pressing 1 else the program will verify the voice with the stored data.

During the first program we have speech the correct word, and the output is as:

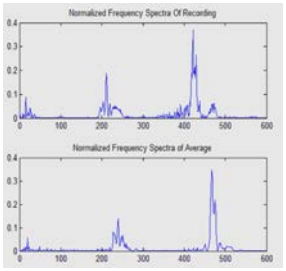

Fig. 2. Voice match with pre recorded voice

During the second program we have speech the wrong word, and the output is as:

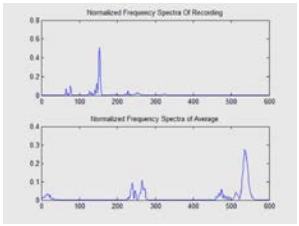

Fig. 3. Voice do not match with pre recorded

voice

6 Filter analysis

Butterworth filter

BW filter is a signal processing filter that is designed to have as flat a frequency response as possible in the pass-band. It is also referred to as a maximally flat magnitude filter. For an N-order low pass filter, the flat characteristic is the derivative of (2N-1) order of analog function in ω=0. The other feature of BW filter is its amplitude frequency characteristic which is always monotonically decreasing function of frequency, and its analog function is closer to the ideal low pass filter with N increase.

Bandpass filter

A band-pass filter can be characterised by its Q factor. The Q-factor is the inverse of the fractional bandwidth. A high-Q filter will have a narrow passband and a low-Q filter will have a wide passband. These are respectively referred to as narrow-band and
wide-band filters.. The cutoff frequencies, f_{start} and f_{end}, are the frequencies at which the output signal power falls to half of its level at f_0, the center frequency of the filter. The value f_{end} – f_{start},expressed in hertz (Hz), kilohertz (kHz),megahertz (MHz),
or gigahertz (GHz), is called the filter bandwidth. The range of frequencies between f_{start} and f_{end} is called the filter passband.

The maximum frequency that is representable in our signal is the sampling frequency divided by 2. This is called the Nyquist frequency. The sampling frequency of our audio file is 48000Hz, which means that the maximum frequency represented in your audio file is 24000Hz. fft stands for Fast Fourier Transform. Think of it as a very efficient way of computing the Fourier Transform. The traditional formula requires that perform multiple summations for each element in your output. The FFT will compute this efficiently by requiring far less operations and still give the same result

I used fftshift so that the centre maps to 0Hz, while the left spans from 0 to -24000Hz while the right spans from 0 to 24000 Hz. This is intuitively how I see the frequency spectrum. Ideally, the frequency distribution for a negative frequency should equal the positive frequency. When plot the frequency spectrum, it tells how much contribution that frequency has to the output. That is defined by the magnitude of the signal and find this by taking the abs() function.

A band-pass filter can be characterised by its Q factor. The Q-factor is the inverse of the fractional bandwidth. A high-Q filter will have a narrow passband and a low-Q filter will have a wide passband. These are respectively referred to as narrow-band and
wide-band filters.. The cutoff frequencies, f_{start} and f_{end}, are the frequencies at which the output signal power falls to half of its level at f_0, the center frequency of the filter. The value f_{end} – f_{start},expressed in hertz (Hz), kilohertz (kHz),megahertz (MHz),
or gigahertz (GHz), is called the filter bandwidth. The range of frequencies between f_{start} and f_{end} is called the filter passband.

The maximum frequency that is representable in our signal is the sampling frequency divided by 2. This is called the Nyquist frequency. The sampling frequency of our audio file is 48000Hz, which means that the maximum frequency represented in your audio file is 24000Hz. fft stands for Fast Fourier Transform. Think of it as a very efficient way of computing the Fourier Transform. The traditional formula requires that perform multiple summations for each element in your output. The FFT will compute this efficiently by requiring far less operations and still give the same result

I used fftshift so that the centre maps to 0Hz, while the left spans from 0 to -24000Hz while the right spans from 0 to 24000 Hz. This is intuitively how I see the frequency spectrum. Ideally, the frequency distribution for a negative frequency should equal the positive frequency. When plot the frequency spectrum, it tells how much contribution that frequency has to the output. That is defined by the magnitude of the signal and find this by taking the abs() function.

Gaussian filter

A Gaussian filter is a filter whose impulse response is a Gaussian function (or an approximation to it). Gaussian filters have the properties of having no overshoot to a step function input while minimizing the rise and fall time. This behavior is closely connected to the fact that the Gaussian filter has the minimum possible group delay. It is considered the ideal time domain filter, just as the sinc is the ideal frequency domain filter. These properties are important in areas such as oscilloscopes and digital telecommunication systems. It modifies the input signal by convolution with a Gaussian function. The choice of sigma depends a lot on what I want to do. Gaussian smoothing is low-pass filtering, which means that it suppresses high-frequency detail (noise), while preserving the low-frequency parts.

IIR elliptical filter

An elliptic filter is a signal processing filter with equalized ripple (equiripple) behavior in both the pass band and the stop-band. The amount of ripple in each band is independently adjustable, and no other filter of equal order can have a faster transition in gain between the pass band and the stop-band, for the given values of ripple (whether the ripple is equalized or not). Alternatively,

7 CONCLUSIONS

The outcome of experimental result of this paper is to provide some explanation of the speech and speaker recognition data input of the user. An algorithm which efficiently determines the optimum coordination of H.M.M has been successfully designed with the help of MATLAB. Authentication of the user can be determined by the threshold value being set by the standard variance. One advantage of this algorithm over conventional algorithm is that false acceptance and false rejection can also be controlled by threshold value. It give the brief analysis of various filters using butterworth filter, bandpass filter, Gaussian filter and IIR elliptical filter. For the current thesis, hence domain signals. Degradation of signals by the application of Gaussian noise is performed. Background noise was successfully removed from a signal by the application of a 3rd order Butterworth filter. I choose n = 7 and cut-off frequency=0.05 to start off. It need to normalize your frequencies so that the Nyquist frequency maps to 1, while everything else maps between 0 and 1.

We apply a band pass filter to get rid of the low noise, capture most of the voice and any noisy frequencies on the higher side will get cancelled as well. For a band pass filter, I choose beginFreq and endFreq map to the normalized beginning and ending frequency In our case, that's beginFreq = 700 / Nyquist(fs/2) and endFreq = 12000 / Nyquist(fs/2), where fs is a variable used for wav file read. There are a lot of spikes around the low frequency range. This corresponds to our humming whereas the voice probably maps to the higher frequency range and there isn't that much of it as there isn't that much of a voice heard

We apply a Gaussian filter for the minimum possible group delay. It is considered the ideal time domain filter where I choose sigma=0.335. it suppresses high-frequency detail (noise), while preserving the low-frequency parts.

We also apply a IIReliptical filter where cutoff slope is shaper. It has the structure with a feedback loop. The operation of IIR filter is usually composed by delay. I choose the Wp value on the ellipord function between 12000 and 18000hz and Ws value between 14000 and 16000 hz and other value rp=3 and rs=60 for noise removal. Code was constructed for only ten recordings. Only 2 sec are allotted for the execution of the program and checking of data base. Authentication of the user can be determined by the threshold value being set by the standard variance.the analysis table given below:-

Sample	Sample size	Bw filter (fs)	Gaussian filter	Bandpass filter	IIR elliptical filter
Pragati gaur	1000hz	400hz	450hz	700-12000hz	350hz
Archana	1200hz	398hz	450hz	700-12000hz	375hz
Girija shanker	1100hz	395hz	550hz	700-12000hz	370hz
Manju sharma	1200hz	395hz	550hz	700-12000hz	370hz
Gunav gaur	1100hz	395hz	550hz	700-12000hz	370hz
lado	1050hz	395hz	440hz	700-12000hz	370hz
gaayant	1100hz	395hz	452hz	700-12000hz	370hz
Ravindra solanki	1000hz	400hz	450hz	700-12000hz	350hz
renu	950hz	350hz	420hz	700-12000hz	360hz

Fig 4 Analysis of different filter

Sample	Frequency of a sample	Butterworth filter analysis on different frame length									
		fs	Time	Fs/2	Time	fs/4	time	Fs/8	time	Fs/12	time
Pragati gaur	1000hz	400	[0 2.5]	600	[0 4.5]	650	[0 9]	750	[0 18]	850	[0 30]
Archana	1200hz	398	[0 2.3]	600	[0 4.3]	650	[0 9]	750	[0 18]	880	[0 30]
Girija shanker	1100hz	395	[0 2.1]	620	[0 4.3]	700	[0 8.7]	820	[0 17.5]	900	[0 27]
Manju sharma	1200hz	395	[0 2.2]	670	[0 4.3]	740	[0 8.8]	820	[0 17.5]	900	[0 27]
Gunav gaur	1100hz	400	[0 2.1]	620	[0 4.3]	700	[0 8.7]	800	[0 17]	880	[0 26]
Lado	1050hz	395	[0 2.2]	600	[0 4.3]	700	[0 8.6]	800	[0 17]	900	[0 26]
Gaayant	1100hz	395	[0 2.2]	620	[0 4.3]	700	[0 8.7]	800	[0 17]	880	[0 26]
Ravindra solanki	1000hz	400	[0 2.5]	600	[0 4.5]	650	[0 9]	750	[0 18]	850	[0 30]
Renu	950hz	350	[0 1.8]	680	[0 4]	780	[0 9.5]	850	[0 18.5]	930	[0 31]

Fig5 Butterworth Filter analysis for different frame length

Now from the above table we can see that among the BW filter-Gaussian filter,BandPass filter,IIR elliptical filter the BW filter gives the best result and have a more efficiency to identify a speaker.

8 Future Scope

Over three decades of research in spoken language processing have produced remarkable advances in automatic speech recognition and understanding that helps us take a big step toward natural human–machine communication. Signal-processing techniques led to a better understanding of speech characteristics, providing deep insights into acoustic-phonetic properties of a language. The introduction of a statistical framework not only makes the problem of automatic recognition of speech tractable but also paves the road to practical engineering system designs. It was found that a particular probabilistic measure, the HMM, provides a speech modeling formalism that is powerful and yet easy to implement. Coupled with a finite state representation of a language, hidden Markov modeling has become the underpinning of most of today's speech-recognition and understanding systems under deployment. To accomplish the ultimate goal of a machine that can communicate with people, however, a number of research issues are awaiting further study. Such a communicating machine needs to be able to deliver a satisfactory performance under a broad range of operating conditions and have an efficient way of representing, storing, and retrieving "knowledge" required in a natural conversation. With the current enthusiasm in research advances, we are optimistic that the Holy Grail of natural human–machine communication will soon be within our technological reach.

Acknowledgment

Author would like to thanks Professors. & Head of department of Computer Science and Engineering of IET Alwar for all his support. I am grateful for the encouragement of my internal guide Assistant Professor Nitin Sharma of IET Alwar that made me write and publish paper. The author will also like to express sincere appreciation and gratitude to their family and friends for their support and also for their valuable advice.

References

[1] Huffman, Larry. "Stokowski, Harvey Fletcher, and the Bell Labs Experimental Recordings". www.stokowski.org. Retrieved February 17, 2014.)

[2] Jump up ^ Juang, B. H.; Rabiner, Lawrence R. "Automatic speech recognition–a brief history of the technology development" (PDF). p. 6. Retrieved 17 January 2015

[3] "Speaker Independent Connected Speech Recognition- Fifth Generation Computer Corporation". Fifthgen.com. Retrieved 2013-06-15K

[4] Jump up ^ "British English definition of voice recognition". Macmillan Publishers Limited. Retrieved February 21, 2012.

[5] Jump up ^ "voice recognition, definition of". WebFinance, Inc. Retrieved February 21, 2012.

[6] Jump up ^ "The Mailbag LG #114". Linuxgazette.net. Retrieved 2013-06-15.

[7] O. Castillo, O. and P. Melin, "A New Approach for Plant Monitoring using Type-2 Fuzzy Logic and Fractal Theory", International Journal of General Systems, Taylor and Francis, Vol. 33, 2004, pp. 305-319.

Pragati Gaur is currently pursuing masters degree program in computer science engineering in IET Alwar,RTU,India.
E-mail: pragatigaur91@gmail.com