International Journal of Advanced Research in
Engineering and Applied Sciences

ISSN: 2278-6252

CONTENT BASED VOICE RECOGNITION AND RETRIEVAL SYSTEM USING FEATURE EXTRACTION

Jagdish Kukreja*

M.A. Ansari**

Abstract: *Voice recognition and speech recognition are the two words that have often being interchangeably used. In a technical sense, voice recognition is strictly about trying to recognize individual voices, not what the speaker said. It is a form of biometrics, the process of identifying a specific individual, often used for security applications. As we all have a distinct speaking style, computers can take a sample of voice and analyze it for distinct features. In this paper we take samples of different individuals and then extract the following features of the voice: Short Time Energy, Zero Cross Rate, Spectral Roll Off, Spectral Centroid, Spectral Flux, thus creating a "voice feature data base" that is unique to an individual in the same way a fingerprint is. Using this unique "voice feature data base" we identify a speaker thus calling it a voice recognition or a speaker recognition.*

Keywords: *Voice recognition, speech recognition, samples, finger print, voice feature data base, fingerprint.*

*GLBITM, G.Noida, India

** Gautam Buddha University, G. Noida, India

1. INTRODUCTION

For almost four decades, the voice recognition systems were made by various methods. In 1970, through experiments using statically features of voice spectra extracted from long terms averaged spectrum of a sentence-long utterance [1] and time averaged of log area ratios and fundamental frequencies derived from voiced portions of spoken words [2].The most difficult task in this decade was the inter-session variability (variability overtime) for a given speaker has a significant effect on recognition accuracy. In 1980,spectral dynamic in voice which include identification of consonants and vowel identification, perceptual critical points ,a voice of few ms which include max. Transaction includes important information about consonant. In 1990,a continuous -voice recognition using a newspaper corpus and broadcast news.In 2000,spontaneous voice projects which specify filled pauses, word fragments, reduced articulation or mispronunciation also includes laughter and coughing of a speaker[3,4].

The term "voice recognition" is sometimes used to refer to recognition systems that must be trained to a particular speak-er—as in the case for most desktop recognition software. Even many recognition systems have obtained good response but this performance degrades when noise is introduced due to environment and acoustic difference between different speakers. Conventional voice recognition system using feature extraction will face problem when test speaker's pitch frequency/speaking rate is very different from that of the speaker's data used during training. Voice recognition is a difficult problem, largely because of the many sources of variability associated with the signal. Sounders proposed a voice discrimination system based on zero cross rate [5]. Scheirer, most various features and developed a system based on GMM (Gaussian Mixture Model) and K-nearest neighbor method for voice recognition [6]. Some features like Spectral Centroid, Spectral Flux and pitch are used for voice discrimination [7, 8, 9].Other methods used for audio processing are MFCC and Hidden Markov Model [10] to do voice discrimination. In this paper, all the features are collectively taken to discriminate the audio and do audio indexing and detection.

A chapter related to the robust recognition of voice signal was also published by Richard M. Stern of Microsoft Corporation and Alejandro Acero,(et al.)[11] of the IBM Research Laboratory, in which there was a review on the ongoing research in the use of acoustical pre-processing to achieve robust voice recognition, discussing and comparing approaches based on direct cepstral comparisons. A point was made out that the development of a robust voice recognition system is necessary and at the same time is difficult. The research by them gives the result that even an automatic voice recognition systems that are designed to be speaker independent can perform very poorly when they are tested using a different type of microphone or acoustic environment from the one which they are trained. Information from speech recognition can be used in various ways includes the obvious use of recognized words to enable the use of text-dependent speaker modeling techniques when words spoken are not given [12]. Speaker recognition includes various techniques by verifying their identity and control access to services such as voice dialing, banking by telephone, telephone shopping, data base access service, information services, voice-mail, security control for confidential information areas, and remote access to computers[13]. Automatic voice recognition of surface electromyography signals that are generated due to human articulator muscles have been developed which also benefit from novel feature, which improve the F-score of AF classifiers from 0.467 to 0.686.This technique is applied to resolve the noise robustness[14].Voice recognition methods can be divided into text de-pendent and text independent method. The former require the speaker to as keywords or sentences having the same text for both training and recognition trials whereas the latter do not rely on a specific text being spoke [15].

1.1 Recognized Features

(a) Short Time Energy: As voice consists of a sequence of voiced and unvoiced sounds the short time energy measurement of a voice signal can be used to determine them. Voiced sounds are the sounds that are generated by vibrations of vocal chord and unvoiced sounds are generated when a speaker pronounces such letters as “s”, “f”, “p”, etc i.e. by expelling air through lips and teeth. Short time energy can also be used to detect the transition from unvoiced to voiced part and vice versa. The energy of voiced part is much greater than the energy of unvoiced part.

$$E_n = \sum_{m=-\infty}^{\infty} x^2(m)h(n-m) \qquad (1)$$

The short time energy (eq. 1) shows the concept of calculation of energy in which the value of power is multiplied with the window.

(b) Z**ero Cross Rate** : The zero-crossing rate is the rate of sign-changes along a signal, i.e., the rate at which the signal changes from positive to negative or back. This feature has been used heavily in both voice recognition and music information retrieval and is defined formally as

$$\frac{1}{T-1}\sum_{t=1}^{T-1}\{s_t s_{t-1}\langle 0\} \qquad (2)$$

where s is a signal of length T and the indicator function is 1 if its argument A is true and 0 otherwise. For monophonic tonal signals, the zero-crossing rate can be used as a primitive pitch detection algorithm.

(c)Spectral Roll Off: Spectral roll off is measure of the skewedness of the spectral distribution. Right skewed (higher frequencies) will have higher spectral roll off values. Music will typically have high spectral roll off values.

$$\sum_{k=o}^{M_C^R(j)}\left|X_{jk}\right| = \frac{C}{100}\sum_{k=0}^{S-1}\left|X_{jk}\right| \qquad (3)$$

It is defined as the second frequency bin $M_C^R(j)$ below which the c per cent (eg. C=90) of the magnitude distribution of the DFT X_R coefficient is concentrated for frame

(d) Spectral Centroid: The spectral centroid is a measure used in digital signal processing to characterise a spectrum. It indicates where the "center of mass" of the spectrum is. Perceptually, it has a robust connection with the impression of "brightness" of a sound. It is calculated as the weighted mean of the frequencies present in the signal, determined using a Fourier transform, with their magnitudes as the weights. Some people use "spectral centroid" to refer to the median of the spectrum. This is a different statistic, the difference being essentially the same as the difference between the unweighted median and mean statistics. Since both are measures of central tendency(there are many different descriptive statistics that can be chosen as a measurement of the central tendency of the data items. These include arithmetic mean, the median and the mode. Other statistical measures such as the standard deviation and the range are called measures of spread and describe how the data is spread out. An average is a single value that is meant to typify a list of values. If all the numbers in the list are the same, then this number should be used. If the numbers are not the same, the average is calculated by combining the values from the set in a specific way and computing a single number as being the average of the set) in some situations they will exhibit some similarity of behaviour. But since typical audio spectra are not normally distributed, the two measures will often give strongly different values.

$$\frac{\sum_{n=0}^{N-1} f(n)x(n)}{\sum_{n=0}^{N} x(n)} \qquad (4)$$

where x(n) represents the weighted frequency value, or magnitude, of bin number n, and f(n) represents the center frequency of that bin.

(e) Spectral Flux: Spectral flux is a measure of how quickly the power spectrum of a signal is changing, calculated by comparing the power spectrum for one frame against the power spectrum from the previous frame. More precisely, it is usually calculated as the 2-norm (also known as the Euclidean distance) between the two normalised spectra. Calculated this way, the spectral flux is not dependent upon overall power (since the spectra are normalised), nor on phase considerations (since only the magnitudes are compared).The spectral flux can be used to determine the timbre of an audio signal, or in onset detection, among other things. Power spectrum is the positive real function of a frequency variable associated with a stationary stochastic process or a deterministic function of time, which has dimensions of power per Hz, or energy per Hz. It is often called simply the spectrum of the signal. Intuitively, the spectral density captures the frequency content of a stochastic process and helps identify periodicities. Normalization is the process of uniformly increasing (or decreasing) the amplitude of an entire audio signal so that the resulting peak amplitude matches a desired target.

2. METHODOLOGY

A series of requisite steps are involved in the voice recognition process being collimated to which we can achieve successful voice recognition. The system recognises a person voice by an isolated word uttered into the microphone. The voice signal is digitised and recorded on a computer at a sample rate of 44.1KHz satisfying the Nyquist criteria. The voices are stored as wav file and a feature data base is created which plays an important role in voice identification and retrieval. Then the features, (spectral centroid, spectral flux , zero crossing rate, short time energy, spectral roll off) based on the frequency relations of the voice, which are unique for each isolated word is calculated and stored in a separate data base. After this a query sample is taken of any of the individuals, whose data is already taken in the data base, and the features are also calculated for this sample. Then the features values of this sample are matched with the values stored in the second data base. This process tells the individuals whose features matches with that of the query sample in the decreasing order of best matched sample.

The sound is recorded using a basic in build sound recorder of a computer as shown of Fig. 1.1.The green portion in Fig. 1.1 is the voice ant the length of this dark(green) portion gives the intensity of the voice.

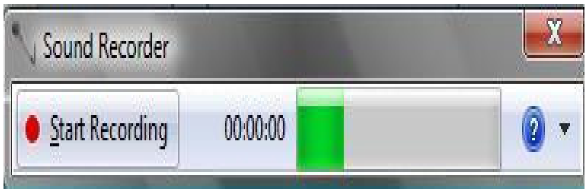


Fig. 1.1 Basic Sound Recorder

In Fig. 1.2 shows the schematic flowchart of the voice recognition presenting all the steps in a orderly manner.

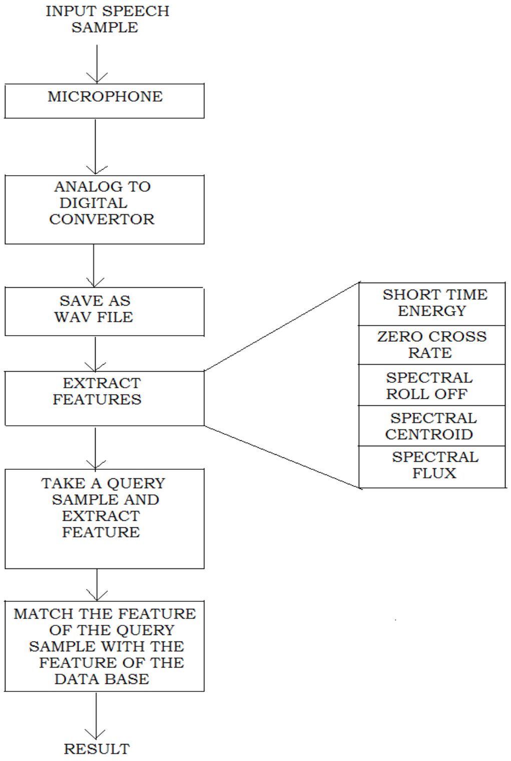


Fig.1.2 Flowchart for the methodology

3. STATISTICAL INTERPRETATION & RESULTS

When the process of recording the voice and then converting it to its digital process is completed then the digitized wave-form of the voice can be viewed in the form of its waveform. As the voice is different and unique for an individual so is its waveform. An example of the digital representation of a waveform is shown in Fig. 1.3

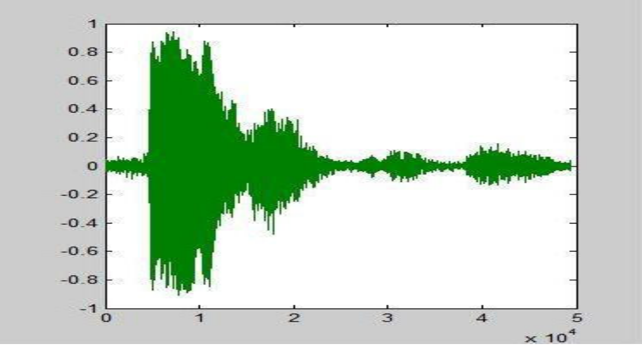


Fig. 1.3 Waveform the a sample as will be seen during the execution of the voice recognition process

For all the samples the parameters are calculated and the graph is plotted collectively. Graph is plotted for the short time energy (Fig. 1.4) of all the samples , zero cross rate of all the samples (Fig. 1.5) , spectral roll off (Fig. 1.6) of all the samples , spectral centroid (Fig. 1.7) of all the samples , spectral flux (Fig. 1.8) of all the samples and so on. For each of the five graphs (as there are five parameters) the mean, median, mode, max. and min. are calculated through the graph and recorded in a table (Table 1.1). The data statistics (mean, median, mode, max. and min) are shown in Table 1.1 which helps us to analyse the variation of the different parameters in different samples.

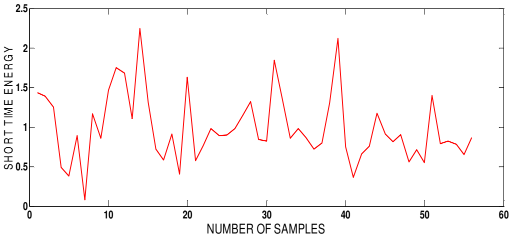


Fig. 1.4 Short time energy

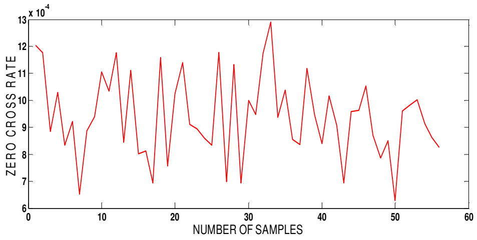


Fig. 1.5 Zero cross rate

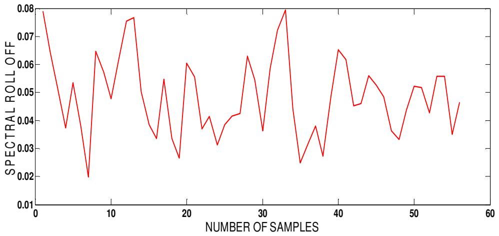


Fig. 1.6 Spectral roll off

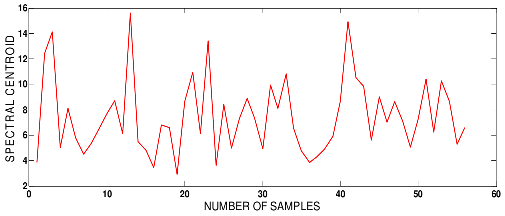


Fig. 1.7 Spectral centroid

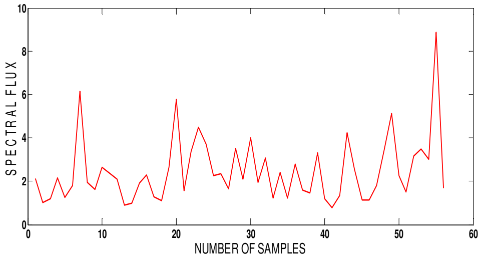
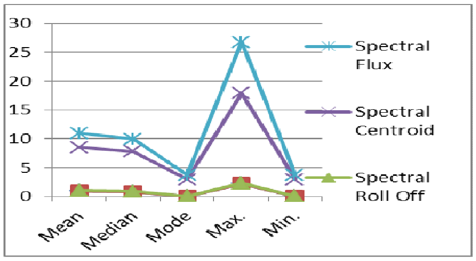


Fig. 1.8 Spectral flux

Table 1.1 Different data statistics for different parameters of the data base made from the samples collected of different individuals

Feature	Mean	Median	Mode	Max.	Min.
Short Time Energy	.9861	0.8752	0.0779	2.244	0.07779
Zero Cross Rate	0.0009402	0.0009295	0.00063	0.00129	0.00063
Spectral Roll Off	0.04855	0.04808	0.01982	0.07945	0.01982
Spectral Centroid	7.478	6.899	2.895	15.6	2.895
Spectral Flux	2.463	2.116	0.7706	8.862	0.7706

Graph of Table 1.1



At the end of the voice recognition process the matching of the features of the query sample with the feature data base created earlier in respect to a single word the result given by the program in terms of the resultant data base is shown in the figure(Fig. 1.8)

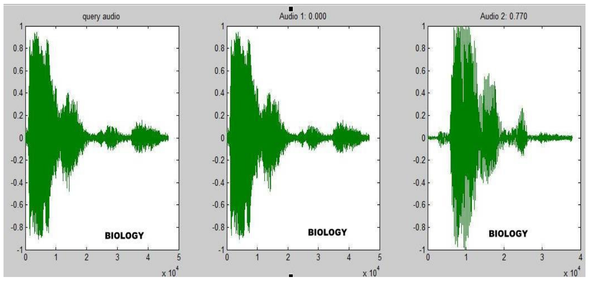


Fig. 1.8 Resultant Data Base

3.1 Male data base

When the data base of the samples is made in such a way that it contains only the samples of male individuals and if the approximation results for it are evaluated then the results calculated si shown in Fig. 1.9

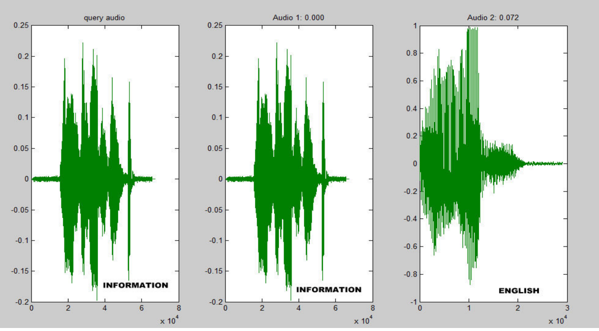


Fig. 1.9 Male Voice Samples

This figure shows the matching of the data base containing only the male voice samples

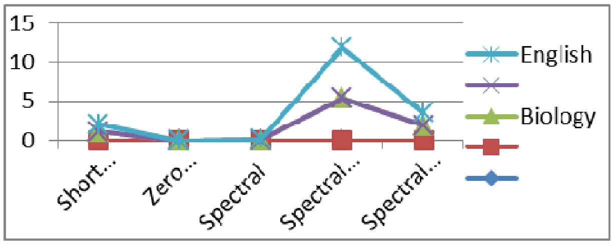
.

Table1.2 Numerical value of the features of a single individual

Word	Short time energy	Zero cross rate	Spectral Roll off	Spectral centroid	Spectral flux
Biology	1.168	0.001	0.065	5.356	1.943
English	0.856	0.001	0.057	6.549	1.6188
Information	1.468	0.001	0.048	7.679	2.653
Security	1.751	0.001	0.061	8.701	2..360

Four samples of a single individual is taken and there numerical value is presented in Table 1.2 and its graph is shown below :

Graph of Table 1.2



3.2 Female data base

If the data base is now modified in such a way that it contains only the voice samples for female individuals and after doing that the matching procedure is repeated then the result obtained is as per the figure Fig. 1.10

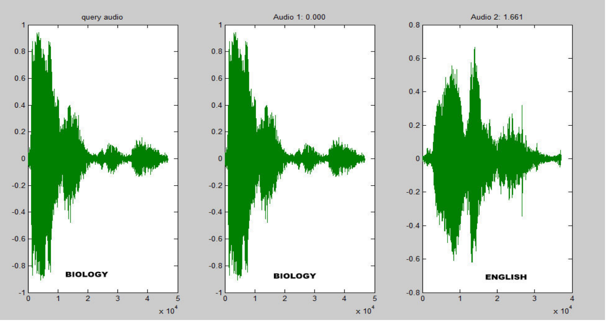


Fig. 1.10 Female Voice Samples

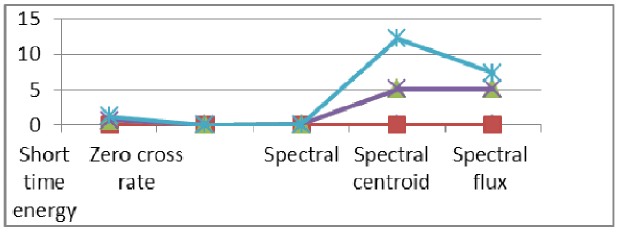
This figure shows the matching of the data base containing only the female voice samples

Table1.3 numerical value of the features of a single individual

Word	Short time energy	Zero cross rate	Spectral Roll off	Spectral centroid	Spectral flux
Biology	0.710	0.001	0.044	5.076	5.107
English	0.533	0.001	0.052	7.208	2.259
Information	1.394	0.001	0.052	10.416	1.498
Security	0.786	0.001	0..043	6.254	3.156

Four samples of a single individual is taken and there numerical value is presented in Table 1.3

Graph of table 1.3



4. CONCLUSION

In this work, a voice recognition system for real time application is developed based on feature extraction and retrieval. The voice recognition system is implemented to match both male and female voice uniquely based on isolated word.95% accuracy obtained for query based searching in the data base. Query data given to the system are matched on the basis of signal quantities like, Short time energy, Spectral roll off, Spectral flux, Spectral centroid, Zero cross rate. All these are compared by using Euclidean distance to give the final result. Therefore it is noble work as a real time voice recognition and retrieval with the software interaction.

REFERENCES

1. Furui,S.,”Itakura,F.,and Saito,S.,''Talker recognition by long time averaged speech spectrum”,Trans.IECE,55-A.pp.549-556.1972.
2. Furui,S.,''An analysis of long term variation of feature parameters of speech and its application to talker recogni-tion”,Trans IECE,57-A,pp.880-887,1974
3. Furui,s.,''Recent advances in spontaneous speech recognition and understanding ,”Proc.ISCA & IEEE Workshop on spontaneous Speech processing and recognitioN,Tokyo,Japan,pp.1-6,2003.
4. Furui,S.,'Recent progress incopus-based spontaneous speech recognition”,IEICE Trans.Inf&Syst.,E88-D,pp.366-375,2005.
5. J. Saunders, “Real-time discrimination of broadcast speech/music,”Proc. ICASSP, vol. 2, pp. 993–996, 1996.

6. M. J. Carey, E. S. Parris, and H. Lloyd-Thomas, "A comparison of features for speech, music discrimination," Proc. ICASSP, vol. 1, pp.149–152, 1999.

7. T. Giannakopoulos, A. Pikrakis, and S. Theodoridis, "A speech/music discriminator for radio recordings using bayesian network," Porc.ICASSP, pp. 809–812, 2006.

8. G. Tzanetakis, "Musical genre classification of audio signals," IEEETrans. Speech and Audio Processing, vol. 10, no. 5, pp. 293–302, 2002.

9. J. Ajmera, I. McCowan, and H. Bourlanrd, "Speech/music segmentation using entropy and dynamism features in a HMM classification framework," Speech Communication, vol. 40, pp. 351–363, 2002.

.10. K. Minami, A. Akutsu, and H. Hamada, "A sound-based approach to video indexing and its application," Trans. IEICE D-II (in Japanese),vol. J81-D-II, no. 3, pp. 529–537, 1998.

11. Richard M. Stern- of Microsoft Corporation, Alejandro Acero,Fu-Hua Liu, Yoshiaki Ohshima-IBM Research Laboratory

12.A.Stolcke, E. Shriburg, L. Ferrer, S.Kajarekar, K. Sonmez, G.Tur Speech technology and Research laboratory, SRI In-ternational, Menlo Park, CA, USA.

13. T. Matswi and S. Furui "Concatenated phoneme models for text-variable speaker recognition",Proceedings of ICASSP93,1993,pp 391-394

14. Szu-Chen Stan Jou and Tanja Schultz International Ceanter for advanced Communication Technologies, Carnegie Melton University, pittsburgh,PA,USA and Karlsruhe University, Karlsruhe, Germany

15. S.Furni,"Cepstral Analysis technique for automatic speaker verfication",IEEE Transactions on Acustics,Speech and signal processing,29(2),1981,PP.254-272