IOSR Journal of VLSI and Signal Processing (IOSR-JVSP)
ISSN: 2319 – 4200, ISBN No. : 2319 – 4197 Volume 1, Issue 1 (Sep-Oct. 2012), PP 43-45
www.iosrjournals.org

Speech Recognition Using Matrix Comparison

Vishnupriya Gupta
M. tech (electronics & communication) Amity University Noida

Abstract: *Speech recognition is a fascinating application of digital signal processing (DSP) that has many real-world applications. Speech recognition can be used to automate many tasks that previously required hands-on human interaction, such as recognizing simple spoken commands to perform something like turning on lights or shutting a door. To increase recognition rate, techniques such as neural networks, dynamic time warping, and hidden Markov models have been used. Recent technological advances have made recognition of more complex speech patterns possible.[1] Speech /voice recognition is a very difficult task to be performed by a computer system. Many speech /voice processing tasks, like speech and word recognition ,reached satisfactory performance levels on specific applications, and although a variety of commercial products were launched in the last decade, many problems remain an open research area, and absolute solutions have not been found out yet[1].*

Keywords: *speech, fourier transform, matrix*

I. Introduction

Speech is an acoustic waveform that conveys information from a speaker to a listener. Given the importance of this form of communication, it is no surprise that many applications of signal processing have been developed to manipulate speech signals. Almost all speech processing applications currently fall into three broad categories: speech recognition, speech synthesis, and speech coding. Speech recognition may be concerned with the identification of certain words, or with the identification of the speaker [2]. Isolated word recognition algorithms attempt to identify individual words, such as in automated telephone services. Automatic speech recognition systems attempt to recognize continuous spoken language, possibly to convert into text within a word processor. These systems often incorporate grammatical cues to increase their accuracy. Speaker identification is mostly used in security applications, as a person's voice is much like a "fingerprint"[4].

Speech coding is mainly concerned with exploiting certain redundancies of the speech signal, allowing it to be represented in a compressed form. Much of the research in speech compression has been motivated by the need to conserve bandwidth in communication systems. For example, speech coding is used to reduce the bit rate in digital cellular systems. In this lab, we will describe some elementary properties of speech signals, introduce a tool known as the short-time discrete-time Fourier Transform, and show how it can be used to form a spectrogram. Speech recognition (SR) aims at converting spoken language to text. Scientists all over the globe have been working under the domain, speech recognition for last many decades. This is one of the intensive areas of research. Recent advances in soft computing techniques give more importance to automatic speech recognition.

II. Earlier Methodology

Early attempts to design systems for automatic speech recognition were mostly guided by the theory of acoustic-phonetics, which describes the phonetic elements of speech (the basic sounds of the language) and tries to explain how they are acoustically realized in a spoken utterance[8]. These elements include the phonemes and the corresponding place and manner of articulation used to produce the sound in various phonetic contexts. For example, in order to produce a steady vowel sound, the vocal cords need to vibrate (to excite the vocal tract), and the air that propagates through the vocal tract results in sound with natural modes of resonance similar to what occurs in an acoustic tube. These natural modes of resonance, called the formants or formant frequencies, are manifested as major regions of energy concentration in the speech power spectrum. [3]In 1952, Davis, Biddulph, and Balashek of Bell Laboratories built a system for isolated digit recognition for a single speaker, using the formant frequencies measured (or estimated) during vowel regions of each digit. [9]These trajectories served as the "reference pattern" for determining the identity of an unknown digit utterance as the best matching digit.

III. Proposed Algorithm

Database collection

Database collection is the most important step in speech recognition. Only an efficient database can yield a good speech recognition system. As we know different people say words differently. This is due to the difference in the pitch, slang, pronunciation. In this step the same word is recorded by different persons. All words are recorded at the same frequency 16KHz. Collection of too much samples need not benefit the speech recognition. Sometimes it can affect it adversely. So, right number of samples should be taken. The same step is repeated for other words also.

Decomposition of speech signal

The next step is speech signal decomposition. For this we can use different techniques like LPC, MFCC, STFT, wavelet transform. Over the past 10 years wavelet transform is mostly used in speech recognition. Speech recognition systems generally carry out some kind of classification/recognition based upon speech features which are usually obtained via time-frequency representations such as Short Time Fourier Transforms (STFTs) or Linear Predictive Coding (LPC) techniques. In some respects, these methods may not be suitable for representing speech; they assume signal stationarity within a given time frame and may therefore lack the ability to analyze localized events accurately. Furthermore, the LPC approach assumes a particular linear (all-pole) model of speech production which strictly speaking is not the case

Feature vectors extraction

Feature extraction is the key for ASR, so that it is arguably the most important component of designing an intelligent system based on speech/speaker recognition, since the best classifier will perform poorly if the features are not chosen well. A feature extractor should reduce the pattern vector (i.e., the original waveform) to a lower dimension, which contains most of the useful information from the original vector.

The following features are used in our system:

- The mean of the absolute value of the coefficients in each sub-band. These features provide information about the frequency distribution of the audio signal.
- The standard deviation of the coefficients in each sub-band. These features provide
- information about the amount of change of the frequency distribution.
- Energy of each sub-band of the signal. These features provide information about the energy of the each sub-band.
- Kurtosis of each sub-band of the signal. These features measure whether the data are peaked or flat relative to a normal distribution.
- Skewness of each sub-band of the signals. These features are the measure of symmetry or lack of symmetry.
- These features are then combined into a hybrid feature and are fed to a classifier. Features are combined using a matrix. All the features of one sample correspond to a column.

Taking the fourier transform of the sound signals and finding the matrix related to them and comparing them with the standard signal helps in speech recognition.

The matrix size of the sample and the reference signal needs to be made same ie. (m,n) of reference = (a,b) of the sample signal.

The sample taken needs more concentration on the sound rof anf language. Foreg in word water the sound'wa' is more important since 'w' itself is composed of several sounds ie "da ba lu'. Hence we need to emphasis more on the sound rather than language chosen it can be the sound signal of the hindi or other language also which has single sound for a character.

IV. Test And Results

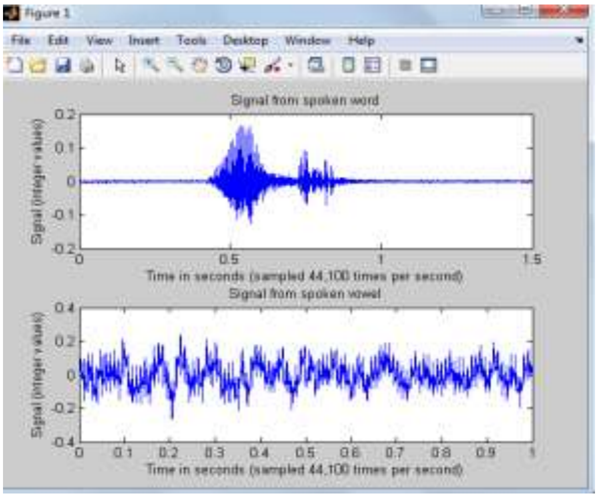

Fig1 signal from word and vowel

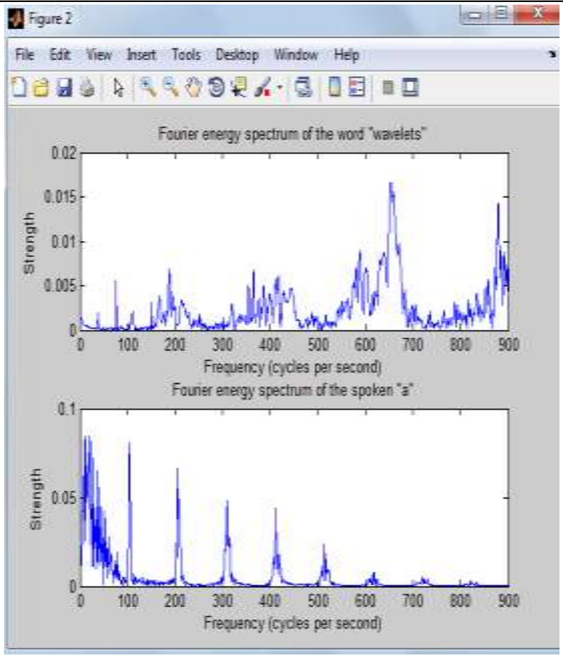

Fig2 fourier energy spectrum

V. Conclusion

In the health care domain, speech recognition can be implemented in front-end or back-end of the medical documentation process. Front-End speech recognition is where the provider dictates into a speech-recognition engine, the recognized words are displayed as they are spoken, and the dictator is responsible for editing and signing off on the document. Back-End or deferred speech recognition is where the provider dictates into a digital dictation system, the voice is routed through a speech-recognition machine and the recognized draft document is routed along with the original voice file to the editor, where the draft is edited and report finalised. Deferred speech recognition is widely used in the industry currently.

Many Electronic Medical Records (EMR) applications can be more effective and may be performed more easily when deployed in conjunction with a speech-recognition engine. Searches, queries, and form filling may all be faster to perform by voice than by using a keyboard. One of the major issues relating to the use of speech recognition in healthcare is that the American Recovery and Reinvestment Act of 2009 (ARRA) provides for substantial financial benefits to physicians who utilize an EMR according to "Meaningful Use" standards. These standards require that a substantial amount of data be maintained by the EMR (now more commonly referred to as an Electronic Health Record or EHR). Unfortunately, in many instances, the use of speech recognition within an EHR will not lead to data maintained within a database, but rather to narrative text. For this reason, substantial resources are being expended to allow for the use of front-end SR while capturing data within the EHR.

References

[1] Vimal Krishnan V.R, Athulya Jayakumar, Babu Anto.P, “Speech Recognition of Isolated Malayalam Words Using Wavelet Features and Artificial Neural Network”, 4th IEEE International Symposium on Electronic Design, Test & Applications

[2] Lawrance Rabiner, Bing-Hwang Juang, “Fundamentals Speech Recognition”, Eaglewood Cliffs, NJ, Prentice hall, 1993.

[3] Mallat Stephen, “A Wavelet Tour of Signal Processing”, San Dieago: Academic Press, 1999, ISBN 012466606.

[4] Mallat SA, “Theory for Multiresolution Signal Decomposition: The Wavelet Representation”, IEEE Transactions on Pattern Analysis Machine Intelligence. Vol. 31, pp 674-693, 1989.

[5] K.P. Soman, K.I. Ramachandran, “Insight into Wavelets from Theory to Practice”, Second Edition, PHI, 2005.

[6] Kadambe S., Srinivasan P. “Application of Adaptive Wavelets for Speech “, Optical Engineering 33(7), pp. 2204-2211, July 1994.

[7] Stuart Russel, Peter Norvig, “Artificial Intelligence, A Modern Approach”, New Delhi: Prentice Hall of India, 2005.

[8] S.N. Srinivasan, S. Sumathi, S.N. Deepa, “Introduction to Neural Networks using Matlab 6.0,” New Delhi, Tata McGraw Hill, 2006.

[9] James A Freeman, David M Skapura, “Neural Networks Algorithm”. Application and Programming Techniques, Pearson Education, 2006.

[10] E. Avci, and Z.H. Akpolat, Speech recognition using a wavelet packet adaptive network based fuzzy inference system, SinceDirect, vol.31, no. 3, 2006, pp 495- 503.