IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727 PP 08-12
www.iosrjournals.org

Advanced Feature Extraction and speech recognition from Voice Encoded Signals

Dr.A.J.Rajeswari Joe
Associate professor
PG Department of Computer science
Thiruthangal nadar college, Chennai

Ms.A.Logalakshmi
Thiruthangal Nadar college

Abstract:
Speech recognition identifies the capability of software or hardware to receive a voice signal, Manipulates the speaker's features in the speech signal, and recognize the speaker thereafter. In general, the process of speech recognition involves three main criteria: acoustic processing, feature extraction, and classification/recognition. The main aim of feature extraction is to illustrate a speech signal using a predetermined number of systems needs a high computation speed. Processing speed plays a vital role in speech recognition in real-time systems. It requires the use of current technologies and wild algorithms that stimulate the acceleration in extracting the feature parameters from speech signals. The experimental results show that the proposed method successfully extracts the signal features. It also achieves unified classification presentation compared to other conventional ***speech recognition algorithms.***
Keywords: Speech Recognition, Neural Networks, Deep Learning, Machine Learning, Speech-to-text.

I. Introduction

Over the last few years, Voice Supporters with the reputation of Google Home, Amazon Echo, Siri, Cortana, have become ubiquitous. These are the examples of Automatic Speech Recognition (ASR) algorithms. This application start with a pin of spoken audio in some language. It can able to extract the words that were spoken as text. For this reason, they are also called as Speech-to-Text algorithms.

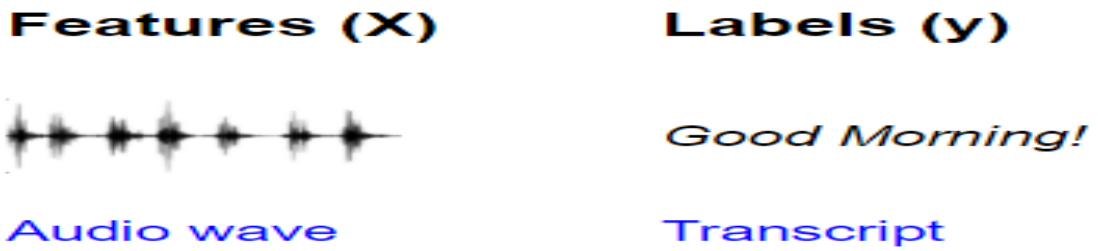

Figure 1: Audio waves and corresponding Transcript

In the above figure, Automatic Speech Recognition uses audio waves as input value and the text transcript as target labels.

The communication approaches between humans and computer technology is a critical task in modern artificial intelligence. One of the easiest methods for users to implement it is entering information through speech signals. Therefore, speech signal processing technology and its tools have become popular and necessary part of the information society. Speech signals contain semantic, personal, and environmental information [1].

The general approach to speech signal processing is to use short-term analysis, in which the signal is divided into time windows of a fixed size, assuming that the signal parameters do not change. An overlap is placed between windows to obtain a more accurate signal representation. Feature extraction algorithms such as spectral analysis and linear prediction are applied to each window. However, these processes should also consider speed and time.

One of the parameters that is especially important in machine learning algorithms is the training time. In this work, we found that the process of extracting features is important for the identification of speech

signals.FFT and DCT spectral transformations give good results in spectral analysis. The spectral analysis process is valid for each segment of the speech signal. This allows us to clearly distinguish these features from the speech signal. With the help of the OpenMP and TBB tools, a parallel computing algorithm was developed that made it possible to speed up the calculations. The segment size is important when dividing a speech signal into segments. [2]. During the experiments, it was found that the size of the segment of the speech signal depends on the cache memory of the central processor. A new parallel implementation method is proposed for spectral transformations of signals for feature parameter extraction from speech signals using a machine learning algorithm on multicore processors using TBB and OpenMP.

Speech-to-Text

As we can imagine, human speech is fundamental to our daily personal and business lives, and Speech-to-Text functionality has a huge number of applications. Basic audio data consists of sounds and noises. Human speech is a special case of that. speech is more complicated because it encodes language.[3]. Problems like audio classification start with a sound clip and predict which class that sound belongs to, from a given set of classes. For Speech-to-Text problems, the training data consists of:

- Input features (X): audio clips of spoken words
- Target labels (y): a text transcript of what was spoken

Voice comparison is a variant of speaker recognition or voice recognition. Voice comparison plays a significant role in the forensic science field and security systems. Precise voice comparison is a challenging problem. Traditionally, different classification and comparison models were used by the researchers to solve the speaker recognition and the voice comparison, respectively but deep learning is gaining popularity because of its strength in accuracy when trained with large amounts of data. This paper also discusses publicly available datasets that are used for speaker recognition and voice comparison by researchers. This concise paper would provide substantial input to beginners and researchers for understanding the domain of voice recognition and voice comparison. [4].

Preprocessing

Speech recognition systems consist of two main subsystems:

 The primary processing of speech signals (phonogram);
 The classification of acoustic symbols using an intelligent algorithm (spectrogram).

The first subsystem generates acoustic symbols as a set of informative acoustic properties of the signals and signal characteristics. The second subsystem converts the speech signal into a parametric form based on the obtained acoustic characteristics [5]. Speech signal processing consists of the following stages:

o Separation of the boundaries of the speech signal;
o Digital filtering;
o Segmentation;
o Application of a smoothing window;
o Spectral transformation, and
o normalization of spectral frequencies.

These steps are used extensively to extract feature parameters from speech signals, and their main features are considered in the following.

Separation of Boundaries

The extraction of only parts of speech from the incoming signal or the determination of the starting and ending moments of a sentence in a noisy environment is an essential task in speech processing. The following properties of the speech signal are used to solve this problem: the short-term energy of the speech signal, the number of points intersecting at zero, the density of the distribution of the values of the silence field within the signal, and the spectral entropy. [6]. Traditional speech recognition systems use techniques that are based on the transient and spectral energies of a signal (e.g., voice activity detection) to determine the speech boundaries from incoming speech signals.

Digital Filtering

In addition to a typical, useful signal, various types of noise are present. As noise negatively affects the quality of speech recognition systems, dealing with noise is a pressing issue. Two types of digital filters are used to reduce the noise levels in the system: a line filter and an initial filter. A linear filter can be considered a combination of low- and high-frequency filters as it captures all low and high frequencies. Initial filtering is

applied to minimize the impact of local disturbances on the characteristic markings that are used for subsequent identification. A speech signal must be passed through a low-pass filter for spectral. alignment. [7][10]

Segmentation

Segmentation is the process of dividing a speech signal into discrete, non-overlapping fragments. The signal is usually divided into speech units such as sentences, words, syllables, phonemes, or even smaller phonetic units. The segmentation of recordings that contain the utterances of numerous speakers may consist of attributing pieces of utterances to particular speakers. The term "segment-stations" is sometimes used to refer to a division of the speech signal into frames prior to its parameterization.

Spectral Transformation

It is necessary to distinguish the main features of speech signals that are used in the later stages of the speech recognition process. [8][9]. The initial features are determined by analyzing the spectral properties of the speech signals. The fast Fourier transform algorithm is commonly used to obtain the spectral frequency of a speech signal.

Normalization of Spectral Frequencies

The entire computational process in intelligent algorithms is based on moving decimal numbers. Therefore, the parameters of the objects that are classified using neural networks are limited to the range [0.0, 1.0]. The resulting spectrum is normalized between 0 and 1 to apply spectral processing using the neural network. To achieve this, each vector component is divided into its maximum components. Spectral processing methods enable the use of all data samples that are obtained from the speech signal. Many speech signals have a specific frequency structure and spectral properties. Spectral methods provide high-precision processing of speech signals. [11] [12]. The disadvantages of spectral processing include low flexibility in the local characteristics of the signals, a lack of spectral dimensions, and relatively high computational costs. Fourier transformation, wavelet analysis, and many other algorithms are used extensively in the spectral processing of speech signals. Fourier transformation is used in many fields of science, including speech processing. In speech signal processing, Fourier transformation converts the signal from the time field to the spectral field as a frequency component.

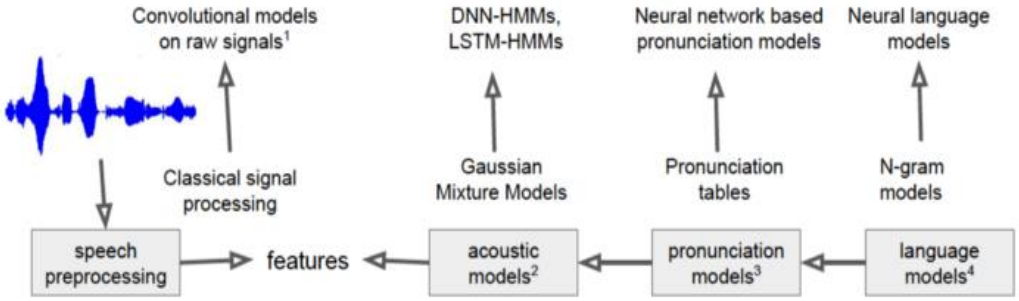

Figure 2: Schematic diagram of Feature extraction from Speech signals

II. Proposed method

This method enables excellent separation of sounds and spoken words while ensuring that speakers are insensitive to pronunciation patterns and changes in the acoustic environment. Most errors in word recognition are caused by a change in the pitch of the signal owing to a shift in the microphone or a difference in the pitch of the pronunciation. Another common cause of errors is random nonlinear deformations of the spectrum shape, which are always present in the speech signal of a speaker. Therefore, one of the most important tasks in creating effective speech recognition systems is the selection of a representation that is sufficient for the content of the analyzed signal as well as insensitive to the voices of speakers and various acoustic environments.

The system that is used for extracting feature parameters typically has the following requirements. The information content, that is, the set of feature parameters, must ensure the reliable identification of recognizable speech elements. Furthermore, the loudness, that is, the maximum compression of the audio signal, and the non-statistical correlation of the parameters must be minimized. Independence from the speaker must also be achieved, that is, the maximal removal of information relating to the characteristics of the speaker from the vector of characters. Finally, homogeneity, which refers to the parameters having the same average variance and the ability to use simple metrics to determine the affinity between character sets, must be provided. However, it is not always possible to satisfy all requirements simultaneously because such requirements are contradictory. The parametric description of the speech elements should be sufficiently detailed to distinguish them reliably and should be as laconic as possible.

In practice, the speech signal that is received from a microphone is digitized at a sampling rate of 8 to 22 kHz. Serial numerical values are divided into speech fragments (frames) with a duration of 10 to 30 ms, which correspond to quasi-stationary speech parts. A vector of features is computed from each frame, which is subsequently used at the acoustic level of speech recognition. At present, a wide range of methods is available for the parametric representation of signals based on autocorrelation analysis, hardware linear filtering, spectral analysis, and LPC.

III. Results and Discussion

Discrete Fourier transform (DFT), discrete cosine transform (DCT), short-term Fourier transform, and wavelet transform were applied for spectral analysis. These methods were used to transform fragments of a speech signal into the frequency domain and calculate the spectrum. The TBB and OpenMP packages were used to create parallel algorithms for spectral transformations. In these methods, the command divides the signal into frames; for example, N = 16, 32, . . . , or 4096 for each frame[4]. The size of each created frame is equal to the block size of the cache memory because cache memory is a factor that affects the efficiency of parallel processing. The acceleration results are depicted in the given Figure.

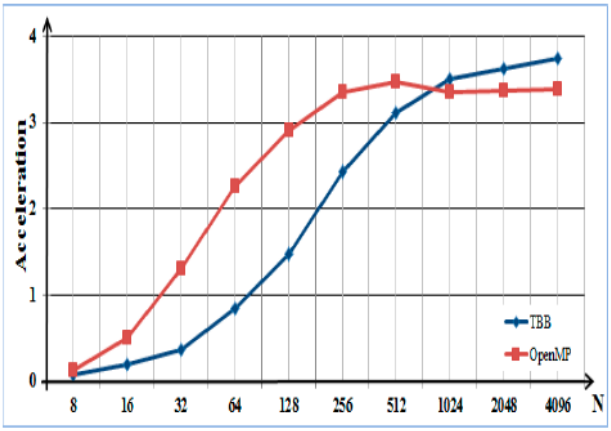

Figure 3: The acceleration results

Fourier analysis parameters are the basis of the methods that are used to generate feature vectors in most speech recognition systems in the digital processing of speech signals within the spectral domain. Mel-frequency cepstral coefficients (MFCCs) and linear predictive coding (LPC) techniques are used in Fourier analysis to generate spectrogram images from speech signals. The spectra that are obtained using Fourier analysis provide concise and precise information regarding a speech signal

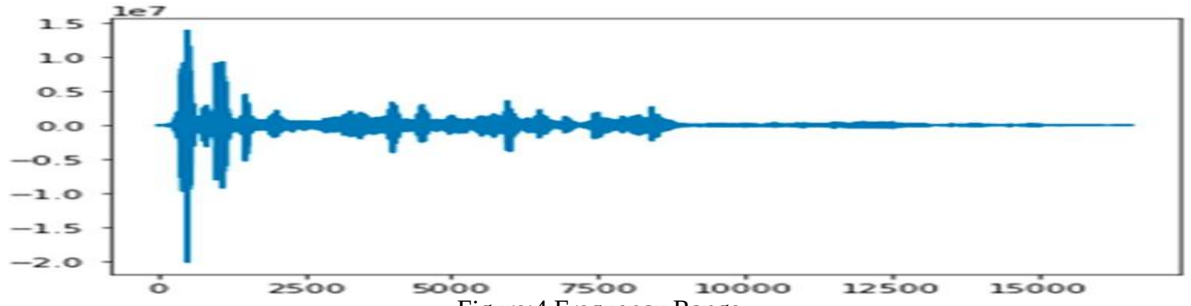

Figure:4 Frequency Range

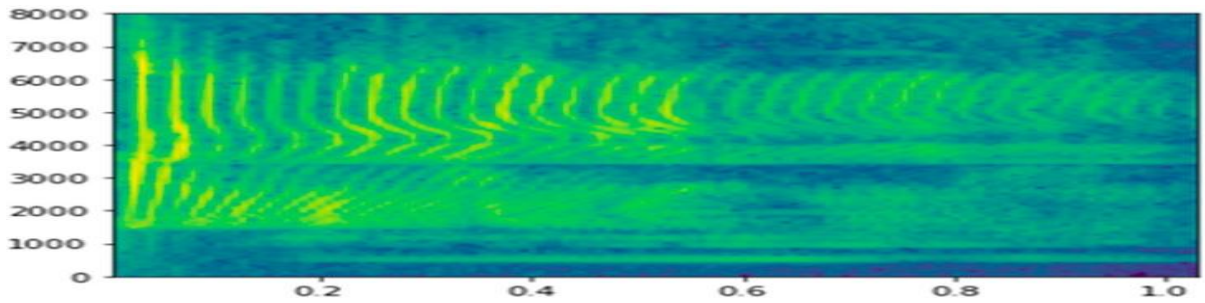

Figure:5 Corresponding Spectrogram levels

IV. Conclusion

In this paper, we have tried to introduce a simple approach which could be used to recognize connected speech and the person concerned. The speech features extorted are compared with related speeches in the database for identification. This approach utilizes it possible by the speech of the broadcaster and it will be simple to authenticate their independence. It generates control access to various applications like window speech recognition. Indeed, the application of this technique will certainly enhance smooth and perfect administrative innovations in the day today activates wherever manpower is entertained in multiples such as libraries banks and various workplaces etc. In the present study it is tried to develop a device which will enable to find the presence of a particular data from the cluster of datasets using python. In future this device can be taken to the next level by using Artificial Neural Network (ANN) which will lead us to work with incomplete knowledge on information related to the speech and the person concerned. [13]. Further in the new application the network layers will be built and trained to show the pictorial representation of in-built data.

References

[1]. Pahini A. Trivedi, "Introduction to Various Algorithms of Speech Recognition: Hidden Markov Model, Dynamic Time Warping and Artificial Neural Networks," International Journal of Engineering Development and Research, Volume 2, Issue 4, 2014.

[2]. M. S. Hossain and G. Muhammad, "Emotion recognition using deep learning approach from audiovisual emotional big data," Inf. Fusion, vol. 49, pp. 6978, Sep. 2019.

[3]. M. Chen, P. Zhou, and G. Fortino, "Emotion communication system," IEEE Access, vol. 5, pp. 326337, 2016.

[4]. Ondruska P., J. Dequaire, D. Z. Wang and Posner, End-to-end tracking and semantic segmentation using recwrent neural networks. Master Thesis, Cornell University, Ithaca, New York, USA, 2016.

[5]. N. D. Lane and P. Georgiev, "Can deep learning revolutionize mobile sensing?" in Proc. ACM 16th Int. Workshop Mobile Comput. Syst. Appl., 2015, pp. 117122.

[6]. J. G. Razuri, D. Sundgren, R. Rahmani, A. Moran, I. Bonet, and A. Larsson, "Speech emotion recognition in emotional feedback for human-robot interaction," Int. J. Adv. Res. Artif. Intell., vol. 4, no. 2, pp. 2027, 2015.

[7]. Subramanian Balaji, Yesudhas Harold Robinson, Enoch Golden Julie, "GBMS: A New Centralized Graph Based Mirror System Approach to Prevent Evaders for Data Handling with Arithmetic Coding in Wireless Sensor Networks," Ingenierie des Systemes

[8]. M. Fayek, M. Lech, and L. Cavedon, "Evaluating deep learning architectures for speech emotion recognition," Neural Netw., vol. 92, pp. 6068, Aug. 2017.

[9]. Q. Mao, G. Xu, W. Xue, J. Gou, and Y. Zhan, "Learning emotion discriminative and domain-invariant features for domain adaptation in speech emotion recognition,'" Speech Commun., vol. 93, pp. 110, Oct. 2017.

[10]. S. Zhang, S. Zhang, T. Huang, and W. Gao, "Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching,'" IEEE Trans. Multimedia, vol. 20, no. 6, pp. 15761590, Oct. 2017.

[11]. Qian, Y. and P.C. Woodland, "Very deep convolutional neural networks for robust speech recognition," in Proceedings of the 2016 IEEE International Workshop on Spoken Language Technology (SLT), San Diego, USA, ISBN:978-1-5090-4903-5, pp. 481-488, 2016.

[12]. S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, pp. 2227-2231, 2017.

[13]. Ji-Hae Kim, Byung-Gyu Kim, Partha Pratim Roy, Da-Mi Jeong "Efficient Facial Expression Recognition Algorithm Based on Hierarchical Deep Neural Network Structure," IEEE Access, vol. 7, pp. 41273-41285, 2019. https://doi.org/10.1109/ACCESS.2019.2907327