International Journal of
Emerging Research in Management &Technology
ISSN: 2278-9359 (Volume-4, Issue-7)

Research Article

July 2015

Speech Recognition and Elimination the Noise Based on MFCC

Karan Singh
Masters in ECE
GGS Kharar (PTU), India

Anil Bajaj
Master in CSE
LPU, India

Rekha Garg (Ass. Prof)
Master in ECE
GGS Kharar (PTU), India

Abstract:

The voice recognition is the method that calculates an optimal match between two given sequences with certain restrictions is called Dynamic time wrapping. The sequences are "warped" non-linearly in the time dimension to determine a measure of their similarity independent of certain non-linear variations in the time dimension. This sequence alignment method is often used in time series classification. Although DTW measures a distance-like quantity between two given sequences, it doesn't guarantee the triangle inequality to hold. The voice recognition is the ability of a machine to recognize the spoken words and convert them to any desired form. In the current scenario when we are moving towards the automated world, the applications of real-time voice recognition are increasing day by day. The voice recognition system is a good choice to give a voice command for any device which requires user inputs to operate. Lifts, television, gaming-stations, smart-phones and medical instruments are the few of the many such examples. The real time voice recognition system first requires some training and then is ready to recognize the real time voice data input. For a new incoming voice command, the system tries to match its features from the existing data set. The command is then classified into the 'best-matched' command from the existing data set. The technological advancement in the field of pattern recognition had made the voice recognition more reliable and user friendly.

Keywords: Mel-frequency cepstral coefficient (MFCC), Short time energy, Zero crossing Rate, Power spectral density, Dynamic time wrapping etc.

I. INDRODUCTION

In recent years, various types of pitch period extraction methods have been The voice recognition is the ability of a machine to recognize the spoken words and convert them to any desired form. In the current scenario when we are moving towards the automated world, the applications of real-time voice recognition are increasing day by day. The voice recognition system is a good choice to give a voice command for any device which requires user inputs to operate. Lifts, television, gaming-stations, smart-phones and medical instruments are the few of the many such examples. The real time voice recognition system first requires some training and then is ready to recognize the real time voice data input. For a new incoming voice command, the system tries to match its features from the existing data set. The command is then classified into the 'best-matched' command from the existing data set. The technological advancement in the field of pattern recognition has made the voice recognition more reliable and user friendly.[1]Voice Recognition Engine that can be used for interaction, as well as automating appliances with the help of simple components. The Voice recognition unit is discussed initially by analysing each of the components used and its functionality. It also provides the reasons based on which the model is framed.The proposed algorithm takes a voice signal input and provides the output class. Figure 1 shows the flowchart of the algorithm. First we take a voice input signal, then the silence part of the signal is removed, i.e. we extract the real input. The four important features viz. STE, ZCR, PSD and Pitch Period are extracted in the next step. A brief definition and mathematical formulation of these features are discussed below: A. Short time energy (STE) STE for each frame is the energy in a sound at a specific instance in time and it provides amplitude variation with frames. STE for the unvoiced part of the signal is very high compared to the STE for voiced part.[2] speech signal the pitch period can be though as the period of vocal cord vibration that occurs during the production of voiced speech.

Using this fact, we can separate the voiced and the unvoiced part of signal. We calculate S each frame independently using eq.1[2]

$$E_n = \sum_{m=-\infty}^{\infty} (x[n]\, w[n-m])^2 \qquad (1)$$

Zero Crossing Rate (ZCR) ZCR for a frame is the measure of number of times the signal changes it's sign. Practically, it provides the frequency content of the signal which is useful for separation of voiced and unvoiced part of the signal. ZCR for high speech signal is unvoiced and for low speech signal is voiced. Mathematically ZCR Zn calculated as in Eq. 2.

$$Z_n = \sum_{m=-\infty}^{\infty} |sgn(x[m]) - sgn(x[m-1])| w[n-m] \qquad (2)$$

Thus, Zn provide the number of zero crossing frame by frame.

1. **Power Spectral Density (PSD)**:PSD represents the power distribution of signal at different frequency components. PSD can be calculated by taking the Fourier transform of the autocorrelation function of a signal. Autocorrelation function r(T) for signal x[n] is given by Eq. 3.[2]

$$r(\tau) = \sum_{n=0}^{N-1-|i-j|} x[n] * x[n + |i-j|] = r(|i-j|) \quad (3)$$

2. PSD from autocorrelation can be determined by Equation 4:

$$PSD = \sum_{\tau=-\infty}^{\infty} r(\tau) e^{-j2\pi f \tau} \quad (4)$$

3. In order to get better performance, we use the log of PSD as a feature instead of PSD. D. Pitch Period Period for any signal can be defined the time required to complete one cycle. From human articulatory system we understand that voice generated due to the vibration in the vocal cord and these vibrations are constant short duration practically for 10 - 20 milliseconds. Thus, for applied. We have developed and implemented a new technique for determining the pitch using the short time cepestral coefficient. Cepestral coefficient is the log of Fourier transforms of speech signal calculated frame by frame. Here we are assuming that a female & male speaker speaks with the frequency of 50Hz-300Hz and 200Hz-450Hz respectively thus our region of interest is 50Hz-450HZ. Differential of cepestral coefficient gives us a constant for fixed spectral slope and peaks in the spectrum get well preserved, these peaks with respect to 50 Hz for each frame gives us the relative pitch frequency for that frame. The value of obtained by solving Eq. 5 gives us the relative pitch frequency.

$$\left(\frac{\partial}{\partial \omega}\left[\log\left|X(e^{j\omega})\right|\right]\right) = 0 \quad (5)$$

4. Finally, we are normalizing the whole pitch to make sure that the relative pitch doesn't make any difference. After extracting the features we take some samples as trainer with their known classes. Then we apply the k-NN algorithm for every new coming data point to find it's class.
5. Analytical hierarchical Process (AHP): AHP is a structured technique for organizing and analyzing complex decisions. We are applying this method to determine the weights to be assigned to different features based on their relative importance. For this purpose we prepare a pair-wise comparison matrix using a scale of relative importance. If a feature is compared to itself then it is assigned a value 1.[3]
6. F. K Nearest Neighbor (K-NN) Algorithm For efficient pattern classification we are using k Nearest Neighbor algorithm. K-NN algorithm measures the distance between real time voice sample and the set of stored voiced sample. Euclidian distance d of a new voice sample X from a stored sample is calculated as in the equation 6.

$$d = \sqrt{\begin{matrix} W_{ZCR}(X_{ZRC} - X_{1,ZRC})^2 + W_{STE}(X_{STE} - X_{1,STE})^2 + \\ W_{PSD}(X_{PSD} - X_{1,PSD})^2 + W_{PP}(X_{PP} - X_{1,PP})^2 \end{matrix}}$$

7. Cluster formation occurs depending upon the closeness of real time sample with different set of data samples. The class of the new data point is decided using majority voting of k-nearest neighbors.
8. How does speech recognition work?
9. The field of software engineering that arrangements with outlining PC frameworks that can perceive talked words. Note that voice acknowledgment suggests just that the PC can take transcription, not that it comprehends what is being said. Appreciating human dialects falls under an alternate field of software engineering called common dialect handling.
10. Various voice acknowledgment frameworks are accessible available. The most capable can perceive a great many words. In any case, they by and large oblige an augmented instructional course amid which the PC framework gets to be usual to a specific voice and accent. Such frameworks are said to be speaker subordinate.
11. Numerous frameworks additionally oblige that the speaker talk gradually and particularly and isolate every word with a short respite. These frameworks are called discrete discourse frameworks. As of late, extraordinary steps have been made in consistent discourse frameworks - voice acknowledgment frameworks that permit you to talk actually. There are currently a few persistent discourse frameworks accessible for PCs.
12. Due to their impediments and high cost, voice acknowledgment frameworks have customarily been utilized just as a part of a couple specific circumstances. For instance, such frameworks are valuable in occasions when the client is not able to utilize a console to enter information in light of the fact that his or her hands are possessed or incapacitated. As opposed to writing orders, the client can basically talk into a headset. Progressively, nonetheless, as the expense reductions and execution enhances, discourse acknowledgment frameworks are entering the standard and are being utilized as a distinct option for consoles.

II. LITERATURE REVIEW

- **Yogesh Kumar Sen, R. K. Chaurasiya et,al.(2014)** The classical front end analysis in speech recognition is a spectral analysis which parameterizes the speech signal into feature vectors. This paper proposes a voice recognition model that is able to automatically classify and recognize a voice signal with background noise. The model uses the concept of spectrogram, pitch period, short time energy, zero crossing rate, mel frequency scale and cepestral coefficient in order to calculate feature vectors. The proposed model applies k-NN algorithm for classification by using AHP to scientifically decide the weights of different features. We can conclude from the results section that our model is performing well if it is trained for a single user, and can be used in various commercial situations. But for multiuser case the accuracy is not very high.[1]
- **Hasan Serhan Yavuz, Hakan Çevikalp et,al.(2014**) This paper presents we aligned voices in a fully automatic manner and we obtained more reliable and realistic voice recognition rates. Experimental results showed that the automatic recognition rates can reach close to 90% correct recognition rates.[3]

- **G. Prashanthi, E. G. Rajan et,al.(2014)** This paper proposes a novel technique of sparsing speech data and compressing it in spectral domain. Discrete Transform is applied to voice data and the spectrum is sparsed by retaining the first component CPI (Cumulative Point Index) of the spectrum and forcing the other spectral components to zero. Thus the spectrum could be compressed to a maximum of 12.5% of the original data. As and when required the compressed spectrum could be synthesized using Inverse Discrete Transform and the reconstructed speech data analyzed for speaker recognition. Speaker recognition accuracy to a maximum of 93.5% has been obtained in this case.
- **Dian RetnoAnggraini et,al.(2012)** This research focus on developing a voice recognition system based on Principal Component Analysis(PCA) and unsupervised learning algorithm. The selected database for this research is Essex database that are collect at University of Essex which consist of 7900 voices taken from 395 individuals(male and female)[5]
- **Taketo Horiuchi et,al.(2011)** Biometrics is now recognized as an essential technology for establishing secure access control .It uses physiological characteristics of humans for for identifying individual and voice is one of the attributes usable for biometrics.

III. TECHNIQUES USED (RESEARCH METHODOLOGY)

MFCC(Mel-frequency cepstral co-efficient) Mel-frequency cepstral co-efficient Analysis In general MFCC, the frequency axis is initially warped to the mel-scale which is roughly below 2 kHz and logarithmic above this point. Triangular filter are equally spaced in the mel-scale are applied on the warped spectrum. The result of the filters are compressed using Log function and cepstral coefficient are computed by applying DCT to obtain MFCC feature vector for spoken words. MFCC is a signal perception model, which approximates the human auditory system's response more closely than the linearly-spaced frequency bands used in the normal spectrum.

The frame of MFCC algorithm is as follows:
Step1: The input signal x (t) which undergo a complex series of transformations in the early staged of auditory processing as in have to be transferred into x(ω) by the discrete fourier transform(DFT).
Step 2: Map the powers of the spectrum obtained above onto the mel scale, using triangular overlapping windows.
Step 3: Take the logs of powers at each of the mel frequencies.
Step 4: Take the discrete cosine transform (DCT) of the list of mel log powers.
Step 5: The MFCCs are the amplitudes of the resulting spectrum.

Dynamic time warping (DTW)Dynamic time warping is an algorithm for measuring similarity between two temporal sequences which may vary in time or speed. For instance, similarities in walking patterns could be detected using DTW, even if one person was walking faster than the other, or if there were accelerations and decelerations during the course of an observation. DTW has been applied to temporal sequences of video, audio, and graphics data in deed, any data which can be turned into a linear sequence can be analyzed with DTW. A well known application has been automatic speech recognition, to cope with different speaking speeds. Other applications include speaker recognition and online signature recognition. Also it is seen that it can be used in partial shape matching application.
In general, DTW is a method that calculates an optimal match between two given sequences (e.g. time series) with certain restrictions. The sequences are "warped"non-linearly in the time dimension to determine a measure of their similarity independent of certain non-linear variations in the time dimension. This sequence alignment method is often used in time series classification. Although DTW measures a distance-like quantity between two given sequences, it doesn't guarantee the triangle inequality to hold.

1) **Fast computation:**Computing the DTW requires O(N^2) in general. Fast techniques for computing DTW include Sparse DTW and the Fast DTW. A common task, retrieval of similar time series, can be accelerated by using lower bounds such as LB_Keogh or LB_Improved. In a survey, Wang et al. reported slightly better results with the LB_Improved lower bound than the LB_Keogh bound, and found that other techniques were inefficient.[3]
2) **Average sequence:** Averaging for Dynamic Time Warping is the problem of finding an average sequence for a set of sequences. NLAAF is the exact method for two sequences. For more than two sequences, the problem is related to the one of the Multiple alignment and requires heuristics. DBA is currently the reference method to average a set of sequences consistently with DTW. COMASA efficiently randomizes the search for the average sequence, using DBA as a local optimization process.[3]
3) **Supervised Learning:** Dynamic Time Warping is used as an elastic distance measure for the Nearest Neighbor Classifier.[4]

IV. SOFTWARE DETAIL

Matlab is a programming environment as well as a high level, interpreted, dynamically typed language. It is well suited for numerical computation, particularly computations involving matrix operations and linear algebra.

V. CONCLUSION

This research focus on eliminate the noise(wanted signal) present in the input signal and get the high freaming error rate(FER) , due to these user oprate the device without any problem also provide the security at high level.

ACKNOWLEEDGEMENT

Working on this thesis **of speech recognition and elimination of noise based on MFCC**provided a unique experience and analysis, I feel great pleasure and privilege in working over this research. I am deeply indebted to **"GGS Kharar"** for the guidance, support and motivation for the many other aids without which it would have been impossible to complete this project. I have no words to express my deep sense of gratitude for RekhaGarg (Mentor) mam for her enlightening guidance, directive encouragement, suggestions and constructive criticism for always listening to our problems and helping us out with their full cooperation. Last but not the least,Father ShyamLal MotherNeelam Devi Brothers Harkaran who have given me that much strength to keep moving on forward every time, we are greatly thankful to them and have no words to express my gratitude to them.

REFERENCES

[1] Yogesh Kumar Sen, R. K. Chaurasiya. IEEE International Conference on voice rcoginition-june2014,24:58-95.
[2] Daubechies, I. The wavelet transform, time-frequency localization andsignal analysis. IEEETransformation and Information Theory.2014,36: 961-1005.
[3] Hasan Serhan Yavuz, Hakan Çevikalp .A wavelet Tour of Signal ProcessingIEEE International Conference on signal processing june 2014,34:19-445.
[4] Tiecheng Yu. The Development State of the Voice Identification. The Development communication world.2005,2:56-59.
[5] Dian RetnoAnggraini .The development of a voice recognition system based on Principal Component Analysis(PCA) and unsupervised learning algorithm.2012,4:35-58.
[6] Jiqing Han, Lei Zhang, Tieran Zheng. Voice Signals Processing[M].Beijing: Tsinghua University Press 2004,3:67-94.

