ANALYSIS OF AUTOCORRELATION DOMAIN PROCESSING FOR ROBUST FEATURE EXTRACTION TECHNIQUES OF SPEECH RECOGNITION

[1]G RAMESH, [2]G LAXMI PRIYANKA, [3]CH SWATHI
[1,2,3]Assistant professor, ECE Department, St.Martin's Engineering College

ABSTRACT

Speech analysis involves various methods such as speech transcription, synthesis of speech, identification of speech and identification of speakers. Speaking processing has many uses in the area of digital signal processing and it is still a field of intensive study. Two basic operations, including feature extraction and classification, are often done by speech processing. For the successful speech processing system, the key criterion is the selection of a practical extraction method that plays an important part in system accuracy. This work requires three stages from the input speaker signal recognizing the text, including preprocessing, extraction of features and multiple Vector Support Machine (SVM).The signal is processed and the signal is noise free by storing and removing the characteristics. Various optimization algorithms are used to automate these functions. The best features, including peak signal strength, tri-spectral function and discrete wave transformation (DWT) are obtained by this algorithm in the APSO technology.

1. INTRODUCTION

Speech is the most frequent method of personal communication and speech processing is one of the most exciting research areas for signal processing. Speech processing is nothing more than learning the signals and methods for interpreting them. Signals are typically electronically processed, and speech processing is a single case of digital signal processing that is used for speech signal. The digital signal processing.Voice Recognition is one of the dipping fields of language processing (voice), also known as automatic speech recognition. A computer can pay attention to individual voice commands and to understand individual languages with speech recognition technology. Speech recognition is the process by means of an algorithm which is carried out as a computer program to transform a given input signal into a number of words. In other words, the speech recognition program helps a machine to understand and translate words a human says on a microphone or telephone in reading.

Speaking is one of the traditional ways of expressing oneself. Such speech signals are currently also used in biometric recognition technology and computer communication.The speech signals are slowly coordinated (quasi-stationary) with various signals. Their characteristics are very stationary when tested over a relatively short period of time (5-100 msec). Nevertheless, when the characteristics of the signal shift for some time, it represents the various speech sounds.The knowledge in the speech signal is in reality a short-term wave-form amplitude. Through this way, features can be derived from speech (phonemes) by using a short-term amplitude array.

All the sound extracted features are used for classifying the audio class and are very high, but it decreases slowly as class numbers increase. So, if we choose the feature set, the number of classes that that trigger additional problems by choosing the wrong feature in relation to the problem definition, the result will allow you to make a comparison and direct you when using which one.

Speech is the most typical way for human communication. Speech also contains the speaker's information. There are features in the voice signal to identify the speaker of the speech. These extracted features are useful for the voice recognition model research. Functional extraction is the cornerstone of audio processing. In speech recognition and processing systems, the importance of feature extraction can never be ignored[8]. However, these features extracted must follow these requirements when understanding the voice. The following are the standards[9]:

- Easy to measure extracted speech features
- Not be susceptible to mimicry
- Perfect in showing environment variation
- Stability over time

The combination of speech features, which often occurs between different speakers and in different methods for the same speaker [8]. In recent years, research into speech synthesis has moved from using speech synthesis technology for unit selection[9]. In factors such as production pattern, perception, bandwidth, loudness and intensity, digital speech is different from audio signal. The right method for protecting and monitoring digital media [10] is digital watermarking. Emotional recognition from speech has been used in recent years in speech processing systems, including speech tutoring systems, the medical emergency environment for the diagnosis of stress and pain, robotics experiences, video games and call centers[11]. The main purpose of this work is to investigate the classification of speech signals in the SVM method with optimum selected functional extraction procedure. This optimum functional selection method is used to boost the accuracy of the classification method through various swarm intelligence optimizations.

2. LITERATURE REVIEW

BhuvaneshwariJolad, A new area of research is an automated speech recognition, which enables a normal and user-friendly communication between the person and the system. The appreciation of speech is the capacity to pay attention, understand and execute behavior based on the information we speak about. This article discusses the identification of language and various strategies including MFCC, LPC and PLP for speech recognition. Of the three methods, that is to say. MFCC, LPC, PLP, Mel frequency cepstral coefficient (MFCC) are used frequently in language recognition processes because they are nearest to the actual acoustic opinion.

Dr. Vilas Thakare, All audio information is transmitted in the time domain of a speech signal. On the basis of the waveform itself, very little can be known from the phonological point of view. However, previous work in mathematics, acoustics and speech technology presented various methods for data conversion that can be regarded as information if correctly interpreted. It is important to have mechanisms to reduce the information in each segment in the audio signal to a relatively small amount of parameters or features, to distinguish certain statistical information from incoming data. Such characteristics should define each segment in a way that can combine other similar segments by comparing their characteristics. The speech signal in terms of parameters is incredibly interesting and exceptionally defined. While they all have their strengths and their disadvantages, we have shown their value to some of the most used methods.

Isra Khan, Since the world is moving to a new age called "artificial intelligence," in which certain things can be automatically controlled by a lot of sources such as face and thumb lock, we can monitor things by sound as the technology progresses every day, and the technique is increasingly increasing but not explored. We explore the sound and the techniques of its extraction by means of which we can extract features from different types of sound and make them relevant, since this paper presents an examination of the extraction of the features for comparative analysis of features such as noisy data, difficulty, accuracy and extraction process. Function extractions have a direct connection to every algorithm of the machine learning.

Authors of based on comparative analyzes and the analysis concluded that the PLP is derived from the logarithmic spaced filter bank, in accordance with a configuration of a human hearing device, and it has improved its results than the LPC.

2011 F. L. Huang suggested an efficient Chinese vocabulary recognition method focused on the Hidden-Markov-Model (HMM) Individual language recognition for Chinese words. The characteristics of words are formed by Chinese character sub-syllables. There are a total of 640 speech samples of 4 men and 4 women who talk often. Preliminary findings from internal and external studies are respectively 89.6% and 77.5%.

For the speech emotion detection discussed at length, authors have extracted MFCC function. MFCC function is extracted, tested out and equipped very effectively to detect the emotion of speech detection.

The authors of focused on the isolated speech recognition by the use of the MFCC and Dynamic Time Wrapping (DTW). In this work, MFCC was used to extract characteristics for the isolated speech recognition.

Research defined and focused on optimizing acoustic characteristics for automatic speech recognition provided by Ant Colony Optimization. Speech signal is regarded as the input in this study and feature extraction is performed with MFCC extraction method by means of this signal, a total of 39 coefficients are extracted by means of MFCC.

3. PROPOSED METHODOLOGY

In interpretation of the speaking signal through estimation of the contents, the proposed speech recognising approach includes three steps such as pre processing, selection of features and multifunctional vector system (SVM).

The paper contains existing features, such as the input speed signals, Peak frequency modularity, MFCC, Tri Spectral Features and Discrete Wavelet Transformation (DWT). This text incorporates all existing features. In the input voice signal certain steps are taken on the wavelet-transform DWT cycle.
In the process of speech recognition the above-mentioned functions are optimized. To achieve optimal functionality, optimization algorithms such as genetic algorithms (GA), genetic adaptive algorithms (AGAs), PSO optimisation, HARMONY SECHARGE (HS) and adaptive particulate swarm optimisation (APSO) are used. Such optimal feature is used to predict the text in speaker-dependent process, using the SVM-process linear kernel-function.
Figure 1 demonstrates our proposed speech technique block diagram.

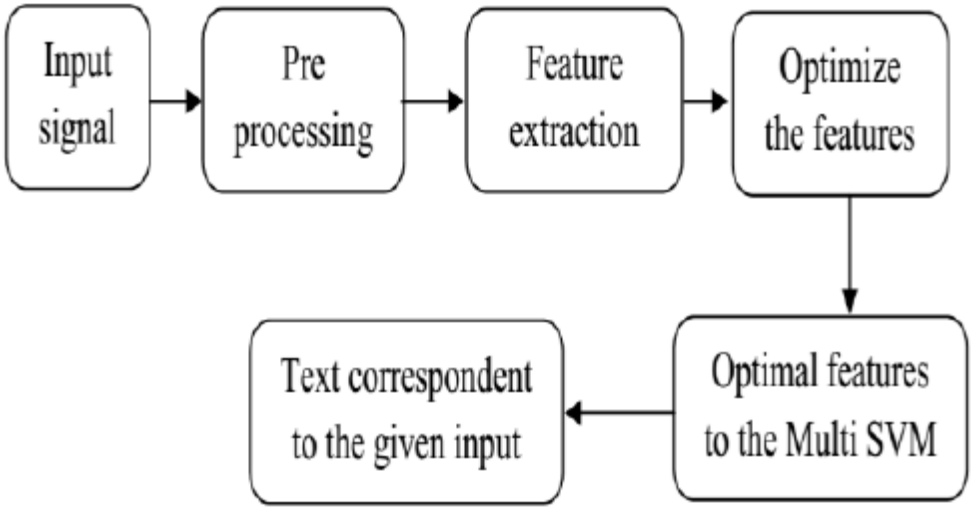

Figure.1 Block diagram for proposed work

3.1 Input Signal

The various voices are captured by the individuals and the same words are spoken, such as dove, rose etc. The input signal S(i) is used.

3.2 Pre Processing

The input signal is vector modified and the variance is discovered . A gauze filtration , defined as P (k) where y represents the noise and¨indicate a deviation, is typically employed for the phase of noise removal.

$$P(k) = \exp\left(\frac{-y^2}{2y\sigma^2} \right)$$

3.3 Feature Extractions

In principle, expression directly from the digital waveform should be identifiable. Nonetheless, it is best to exhaust any feature that would reduce the variability due to the wide variability of the speech signal. Particularly removing different information sources, such as whether the sound is spoken or not, eliminates the effect of periodism or pitch, excitement signal amplitude and simple frequency, etc. when it is spoken.
That is because the human ear cochlevy carries out a near-frequency examination. The short-term range is measured. A non-linear frequency scale (called the bark scale or mel scale) is used for the study of the cochlea. It is essentially linear to about 1000 Hz and is then approximately logarithmic. Thus, after spectral estimation, the frequency distortion of the frequency axis is very normal when extracting features.

Optimal feature selection using APSO algorithm

A particle swarm optimizer is a population-based stochastic optimization algorithm modeled on the simulation of the social behavior of bird flocks. PSO is a population-based quest process, in which people initially clustered into a swarm with a population of random alternatives known as particles.

1. Initialize the solution (F_{ei})

$$F_{ei} = \{F_{e1}, F_{e2}, \ldots\ldots F_{en}\}$$

2. Find the fitness value (F_i)

$$F_i = \max(A)$$

3. Initialize the P_{best} *and* G_{best} value

4. Compute the acceleration factor $nc1$ *and* $nc2$

$$nc_1 = \frac{2}{3}(c_{1\max} - c_{1\min})\left(\frac{f_{\min}}{f_{avg}} + \frac{f_{\min}}{2f_{\max}}\right) + c_{1\min}$$

$$nc_2 = \frac{2}{3}(c_{2\max} - c_{2\min})\left(\frac{f_{\min}}{f_{avg}} + \frac{f_{\min}}{2f_{\max}}\right) + c_{2\min}$$

5. Calculate the velocity and update the position

$$V_i^d = w^d V_i^d + nc_1 r_1 (bp_i^d - p_i^d) + nc_2 r_2 (gp^d - p_i^d)$$

6. Find the fitness for updating solution

$$if(F_{enew}) > f(F_e)$$

7. Store the best solution so far attained

Iteration=Iteration+1

8. Stop until optimal solution attained

Pseudo code for APSO

3.4 Emotional Speech Databases

Feature Extraction for Speech Emotion Recognition

The drawing of the most detailed features to accurately describe various emotions is still an open problem when developing a speech emotion reconnaissance program. Short-term functions, called system by application review, have been widely used by researchers. Regardless of the various emotional speakers' signals, all recordings of emotional words have been conveniently sampled at 8 kHz. The unvoiced portions of the words were extracted from the reported emotional speech signals by segmenting the down-samples of emotive speech signals into non-overlapping 32 ms (256 samples) long images based on the energy of the frames. Low energy frames were disposed of and the majority of the frames (voiced portions) were combined and used for extraction of features. Emotional voice signals (only articulated sections) are then transmitted through a low-pass filter first to flatten the signal spectrally and make later the signal processing less sensitive to finite accuracy. The first pre-filter order is described as

$$H(z) = 1 - a * z^{-1} 0.9 \leq a \leq 1.0$$

PSO clustering for Feature Enhancement

The clustering techniques used for the grouping of related objects / instances in large numbers have been commonly used in various applications, such as mathematics, information engineering, biology, psychology, and social sciences. The increase of inter-class variance and the decrease intra-class variance of attributes or characteristics are key issues in any pattern recognition application for the improvement of classification / accuracy. The efficiency of classifiers can be influenced by high intracategory variance and low interclass variance, which contributes to a poor perception rate.

PSO-based clustering was proposed in this research to decrease the variance intra-class and increase the inter-class variance between the features in order to boost the discriminatory ability of the features extracted. In 1995, Eberhart RC and Kennedy J initially suggested the so-called PSO approach to stochastic optimization. The PSO's principal challenge is that particles can be stored optimally at

local level. For data clusterning Van der Merwe D and Engelbrecht AP proposed the PSO and obtained promising results. Inspired by human social interaction in a global community, Cohen SC and de Castro LN have suggested PSO based clusters to group the data points into clusters based on each individual particle's interdependence. In 2010, Szabo proposed a revamped PSO clustering that did not require speed and inertia during the update operation. Yuwono et al. Mitchell.They suggested a simple amendment by reducing the frequency of the upjustment of the distance matrix to reduce time complexity.

4. EXPECTED RESULTS

Statistical measures of the performance of different texts

Different texts are considered to categorize the text for optimizing functionality in the SVM process. The sample input text with a performance analysis method is shown under the table below.

In Table 1, the input speech signal 'dove' is evaluated by TP, TN, FP, FN and the output from this includes the following: Sensitivity (Se), specificities (Sp), accuracy (A), false positive rate (FPR), positive prediction value (PPV), negative predictive value (NPV), false discovery rate (FDR).

Table 1. Statistical measures of Dove

Persons	Se	Sp	A	PPV	NPV	FPR	FDR	MCC
1	0	1	0.9	-	0.9	0	-	-
2	1	1	1	1	1	0	0	1
3	1	1	1	1	1	0	0	1
4	1	1	1	1	1	0	0	1
5	1	1	1	1	1	0	0	1

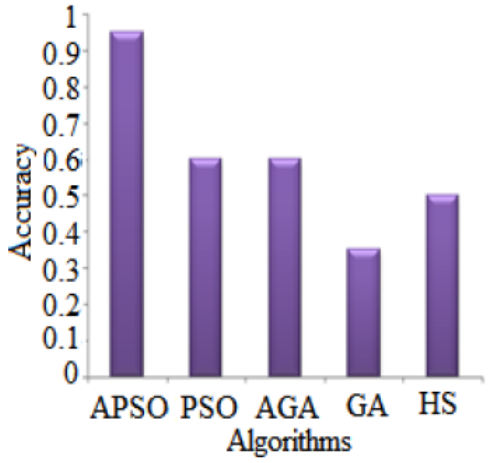

Figure.2 Comparison graph for different algorithm

Figure 2shows that all terms in GA, PSO, AGA and APSO various optimization algorithms have the highest accuracy in the APSO technique. APSO algorithm is the highest accuracy in this optimization method, as 97.8% compared to PSO, and the accuracy is reduced by 3.4%.

Then APSO has been reduced to 3.56% compared with the AGA, the accuracy of the AGA has been reduced to 94%, and the accuracy is reduced to 5.8% compared with the GA. The discrepancy between HS methodology and proposed work is 0.56 percent.

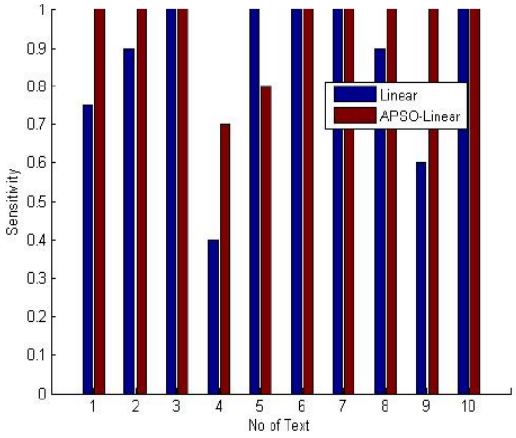

Figure.3Comparsion graph for sensitivity

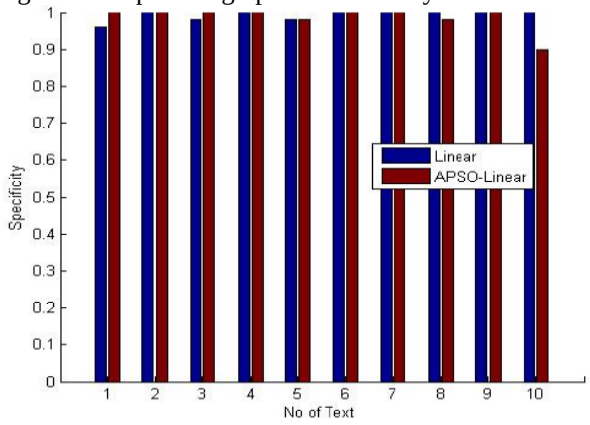

Figure.4 comparison graph for specificity

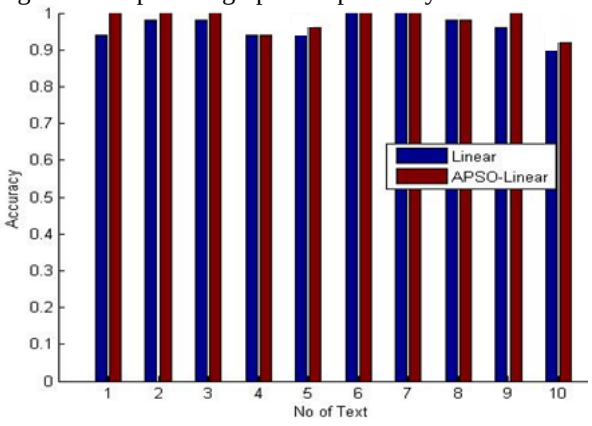

Figure.5 comparison graph for accuracy

CONCLUSION

Improved multi-class identification of the speaker-independent emotions will provide improved contact between human and machine. Throughout this research, we tested the efficacy of the PSO classification and feature selection algorithms to increase the speaking characteristics extracted and also to enhance the precision of the emotional accuracy of the multi-class speaker. This paper categorizes the text in which the Multi Support vector machine with a linear kernel function is able to execute the speech signal on the basis of the various optimized features. The exactness of each person is 97%, 99%, 98%, 99%, 98%, and 97% in APSO technology. The above results clearly demonstrate that our methodology suggested is better than the linear kernel function provided. The benefit of the function is that the complexity of estimation is minimized and the results are reasonable recognition together with a minimized use of time.

REFERENCES

[1] C. Ittichaichareon, S. Suksri and T. Yingthawornsuk, "Speech Recognition using MFCC", *International Conference on Computer Graphics*, pp. 135-138, 2012.

[2] K. Kumar, Aggarwal and A. Jain, "An Analysis of Speech Recognition Performance Based Upon Network Layers and Transfer Functions", *Journal of Computer Science, Engineering and Applications*, Vol. 1, No. 3, pp. 11-20, 2011.

[3] K. Daqrouqa and T. A. Tutunjiba, "Speaker identification using vowels features through a combinedmethod of formants, wavelets, and neural network classifiers", *Journal of Applied Soft Computing*, Vol. 27, pp. 231-239, 2015.

[4] M. Seltzer, D. Yu and Y. Wang, "An Investigation of Deep Neural Networks for Noise Robust Speech Recognition", *Journal of microsoft research*, pp. 7398-7402, 2013.

[5] A. Biswas, Sahu, A. Bhowmick and M. Chandra, "Articulation based admissible wavelet packet feature based on human cochlear frequency response for TIMIT speech recognition", *Journal of Ain shams Engineering*, Vol. 5, No. 4, pp. 1189-1198, 2014.

[6] M. ElAyadi, M. Kamel and F. Karray, "Survey on speech emotion recognition:Features, classification schemes, and databases", *Journal of Pattern Recognition*, Vol. 44, No. 3, pp. 572–587, 2011.

[7] G. Heigold, H. Ney, R. Schlüter and S. Wiesler, “Discriminative training for automatic speech recognition”, *Journal of signal processing*, pp. 58-69, 2012.

[8] O. A. Hamid, L. Deng and D. Yu, "Exploring Convolutional Neural Network Structures and Optimization Techniques for Speech Recognition", *Journal of computer science and engineering*, pp. 3366-3370, 2013.

[9] A. Black, T. Bunnell, Y. Dou, P. K. Muthukumar, F. Metze, D. Perry, T. Polzehl, K. Prahallad, S. Steidl and C. Vaughn, "Articulatory Features For Expressive Speech Synthesis", *Journal of audio speech and language processing*, pp. 4005-4008, 2012.

[10] M. A. Nematollahi, A. Haddad and F. Zarafshan, "Blind digital speech watermarking based on Eigen-value quantization in DWT", Journal of King Saud University Computer and Information Sciences, Vol.27, No. 1, pp.58-67, 2015.