Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Engineering 38 (2012) 952 – 958

Procedia Engineering

www.elsevier.com/locate/procedia

INTERNATIONAL CONFERENCE ON MODELLING OPTIMISATION AND COMPUTING-(ICMOC-2012)

Speech and Non-speech identification and classification using KNN algorithm

T.Lakshmi Priya*, N.R.Raajan, N.Raju, P.Preethi, S.Mathini.

Dept. of ECE,School of Electrical and Electronics Engineering
SASTRA University, Thanjavur, TamilNadu, India

Abstract

Speech and non-speech identification along with its classification method that need to be improved in the endpoint detection for speech in noisy environments. The proposed method uses few features to increase the robustness in various noisy environments, and the classification used here KNN technique is applied to effectively combine these multiple features for classification of each speech signal. We evaluate the performance of the proposed method by conducting speech and non-speech classification experiments on noisy speech. We also investigate the importance of various features on speech and non-speech classification in noisy environments and by using this KNN algorithm to obtaining 80% accuracy.

Keword: Speech, Non-speech, noisy environment, KNN, accuracy.

1. Introduction

Speech is a natural way of essential communication between human beings.Speech is an acoustic signal produced from a system of speech production. The speech production organs consist of lungs, trachea, glottis, pharynx, larynx, oral cavity and nasal cavity. The glottis consists of two membranes called vocal chords.They is the elastic band of muscles that open and closes during speech production. During speech production, the shape of the vocal tract varies due to the movement of articulators namely tongue, jaw, lips, velum and their process is known as articulation [1]. If any one of the speech organs is altered it leads to impairment in vocal communication.

* Corresponding author. Tel.: +0-890-343-1693; fax: +0-000-000-0000 .

doi:10.1016/j.proeng.2012.06.120

E-mail address: rmtlpriya1208@gmail.com

Diagnosis of vocal disorders is an important criterion in the field of speech communication. Speech specialist often uses subjective techniques to improve the vocal problems [2].But it provides a tough process for the specialist and may irritates the patients.So, objective method is proposed for an early detection of vocal disorder. The vocal tract change their shape continuously with time and thus creating an acoustic filter with time varying frequency response.When air from the lungs travels through the tract,it can acts as a filter that shapes the spectrum of sound source to produce speech.The resonance frequencies of vocal tract tube are called formant frequencies.

Our goal is to classify normal and disorderd voices with good classification results for various extracted features.The database consists of dysarthric voice,producing unintelligible speech.Dysarthria is a neurological disorder causing damages to motor speech [3] systems.Normally persons with dysarthric speech is not clear.Dysarthric speaker has pitch breaks,excessive fluctuation of pitch,excessive loudness variations,speech rate,rate fluctuations,prolonged intervals. Using speech processing tools, the above features are nullified.These features are classified into two groups normal and dysarthria and decision is made.K-NN classifier is used which is the simplest classifier and easy to run.This dysarthric speech is given to speech transformation system that transforms dysarthric speech into intelligible speech.

2.1. Feature Extraction

During voiced speech, exhalation of air from wave is created that excites [4] the speech production system. Voiced sound is characterized by strong periodicity present in the signal, with the fundamental frequency called pitch. Normally pitch ranges from 50 to 250 Hz for male and 120 to 500 Hz for female.Males have lower pitch than female due to the difference in length of the vocal folds. Each short time frame of the speech is converted into feature vector.The feature vector contains information that is helpful to identify and differentiate speech sounds.

2.1.1. Autocorrelation method

This method is used in signal processing for analyzing time domain signals.Basically pitch detection algorithms use this method.Fig.1 shows the flow diagram of autocorrelation method.

2.1.1.1. Pre-processing

Pre-processing of speech signal is important in speech processing applications.It includes noise removal, silence period removal, end point detection, framing, windowing, echo cancellation, etc. Silence period removal normally speech signal consist of three regions namely voiced, unvoiced [5] and silence. The speech characters are present only in the voiced part.So, segregating voiced regions from unvoiced/silence regions.Finally removal of silence/unvoiced regions leads to reduction in computational complexity.

2.1.1.2. Framing

Framing is done after removing the noise from the speech signal.Because of slowly varying nature of the speech signal(nonstationary),it is necessary to process the speech into blocks called frames.By framing

the speech signal ,the signal characteristics are uniform in that region to achieve stationarity.Frame duration ranges between 10-25 ms.

2.1.1.3. Windowing

The next important step in speech signal processing is to estimate the spectral characteristics of the speech signal.The data limitations in the speech signal will cause an error in resulting spectrum called frequency or spectral leakage.To reduce this error and to suppress the sidelobe power levels,windowing technique is used.

2.1.2. Set clipping level

Compute the clipping threshold level for the first 30 ms frame of the speech signal.The clipping level is set at a value which is 64 percent of the smaller peak absolute samples in the first and last 10 ms portion of the first frame.The resulting signal produces three possible values -1,0,+1.

2.1.3. Autocorrelation computation

The autocorrelation function is find for its maximum value(normalized).If maximum exceeds 0.3 then the frame is classified as voiced and the maximum point is pitch period [6] otherwise it is unvoiced. The autocorrelation $R(\tau)$ is obtained from the below equation

$$R(\tau) = \int_{-\infty}^{+\infty} f(t) + f(t+\tau)dt \tag{1}$$

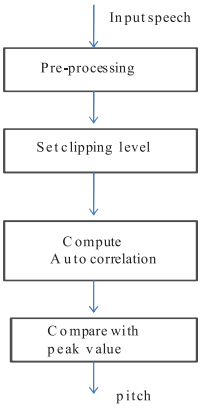

fig.1 flow diagram for pitch detection algorithm using auto correlation method

2.1.4. Compare with peak value

If the peak signal level within the frame is below a given threshold then the frame is classified as unvoiced and further no pitch computation is required.

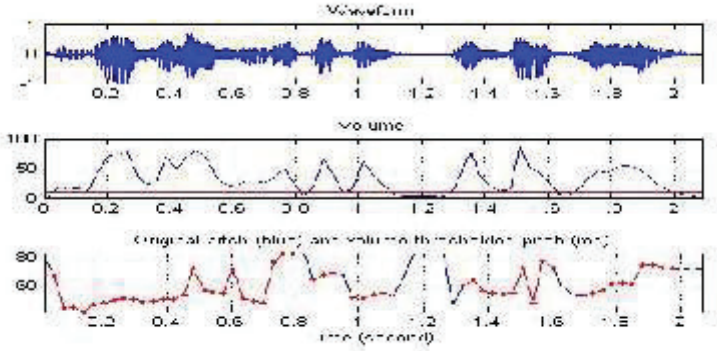

Fig.2 pitch computation for normal speech signal using autocorrelation method

Fig.2 shows that for normal speaker there are no fluctuations in pitch and no pitch breaks occurs.This implies that the speaker is normal and disease free voice communication.

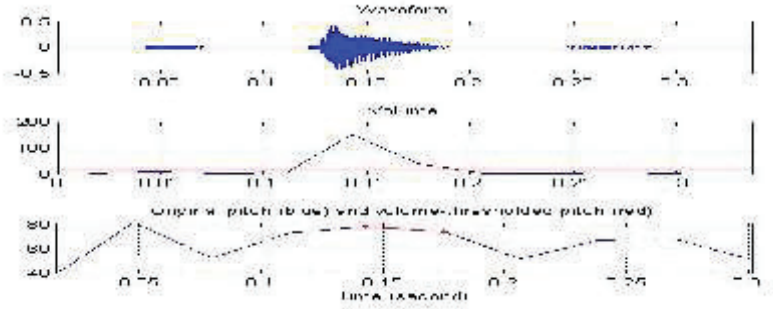

Fig.3 pitch computation for dysarthric speech signal using autocorrelation method

Fig.3 shows that there is pitch breaks and excessive fluctuations of pitch and prolonged time interval are present.This confirms that there is some vocal problem that deviates from normal pitch variations and the person is dysarthria.

2.1.2. Formant Extraction

A formant is a concentration of acoustic energy of particular frequency in the speech signal.Formants are important because they are the essential and meaningful frequency [7] component in the speech signal.Three formants are generally required to synthesize a vowel sound.

2.1.2.1. Filtering

Filtering is done by LPC analysis.The LPC system is used to determine the formant from the speech signal. As the LP spectrum provides the vocal tract characteristics, the vocal tract resonances (formants) are computed from the LP spectrum

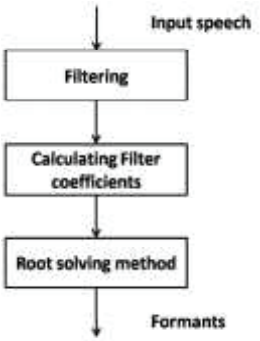

Fig.4 flow diagram of formant extraction

2.1.2.2. Calculate filter coefficients

The filter coefficients are calculated from output of the filtering.let s(n)is the output filtered signal and a_k is the filter coefficients and it is mathematically given by,

$$s(n) = \sum_{k=1}^{p} a_k s(n-k) \tag{2}$$

2.1.2.3. Root solving method

To find the formant frequencies from the filter first find the location [8] of resonance in the vocal tract.Then treating the filter coefficients as a polynomial and solving for roots of the polynomial. Various formant locations can be obtained by picking the peaks from the magnitude of LP spectrum.

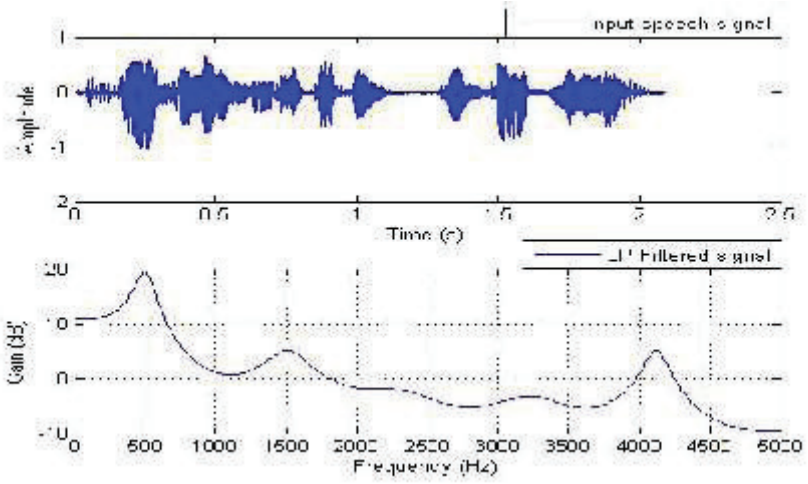

Fig.5 shows standard range of formant frequencies for normal speaker

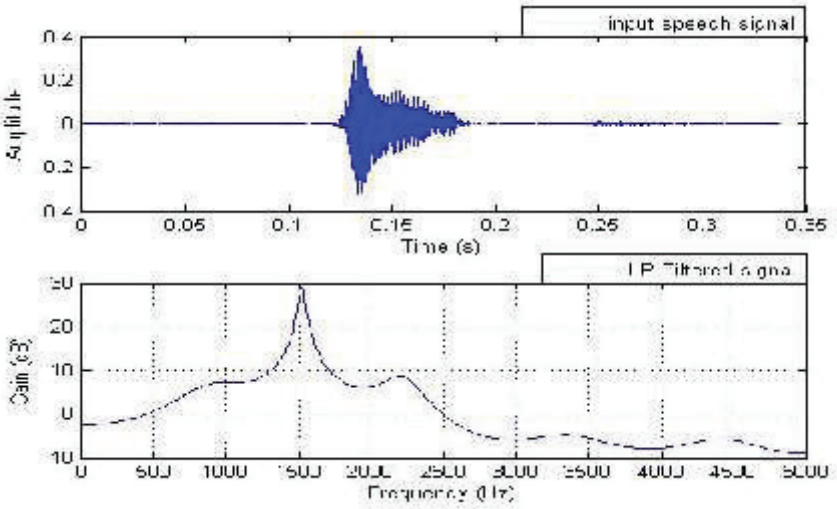

Fig.6 shows sudden and abnormal shifts in the formant frequencies and thus confirms that the speaker is affected some vocal disorders termed as dysarthria

3. Classification

Systematic placement of certain classess in certain groups.The simplest classification algorithm is a K-Nearest Neighbor (KNN) for classifying objects based on closest training examples in the feature space.K is a user defined constant.The larger value of k provides smoother decision region and gives probabilistic information.

K-NN is classified into supervised and unsupervised learning.The supervised learning comprises of labeled classes.These classes are analyzed and trained using supervised learning resulting in desired output.

There are two phases training and testing shows in fig.7

Step 1

In learning phase a computer system is said to learn from the examples to perform certain task.After learning the system performance is measured.

Step 2

In testing phase, trained data is tested and find its accuracy level.Accuracy is defined by the number of correct classifications to the total number of test cases.

3.1. Unsupervised learning

The problem of finding the hidden structure in unlabeled data.Clustering is one approach of unsupervised learning.The goal of clustering to separate [9] the data based on similarties between varieties of classes.Each cluster has a cluster centre called centroid.

Steps involved in clustering

- Determine centroid coordinate
- Determine the distance of each objects to the centroids
- Group the objects based on minimum distance

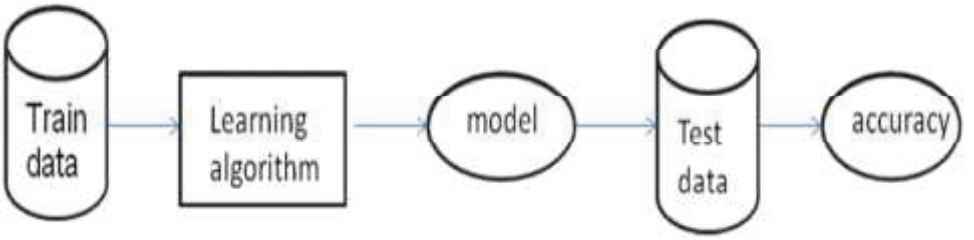

Fig.7 process of classifying data sets

4. Conclusion

From the extracted features it is confirmed that the abnormal variations in pitch and formant results in dysarthria.For Further confirmation the normal and dysarthric voice is classified using K-NN classifier that shows 80% accuracy level.

References

[1] Rabiner, L.R.; Schafer, R.W. "*Digital Processing of Speech Signals*",Prentice Hall: Englewood Cliffs, NJ, 1978.
[2] Carlos Hernindez-Espinosa, Mercedes Fernindez-Redondo, Pedro G6mez-Vilda,Juan **I.** Godino-Llorente, Santiago Aguilera-Navarro," Diagnosis of Vocal and Voice Disorders by the Speech Signal",Proc,IEEE 2000, PP 253-257.
[3] A.M.liberman and I.G.Mattingly,"The Motor Theory of Speech Perception Revised",cognition,Vol.21,(1985),PP.1-36.
[4] V.S.Balaji, N.R.Raajan, Har Narayan Upadhayay, "Identification of Predominent Frequencies in a Speech Signal Using Modeling of Vocal Chord",Proc.IEEE 2011,PP.478-481
[5] Meisam KhalilArjmandi, MohammadPooyan," Biomedical Signal Processing and Control," journal homepage: www.elsevier.com/locate/bspc
[6] S. E. Bou-Ghazale and J. H. L. Hansen, "A comparative study of traditionaland newly proposed features for recognition of speech understress," IEEE Trans. Speech Audio Process., vol. 8, no. 4, pp. 429–442,Jul. 2000
[7] Patricia Henríquez, Jesús B. Alonso, Miguel A. Ferrer, Carlos M. Travieso, Juan I. Godino-Llorente, and Fernando Díaz-de-María,"Characterization of Healthy and Pathological Voice
Through Measures Based on Nonlinear Dynamics",Proc.IEEE 2009,PP.1186-1195.
[8] B. Boyanov and S. Hadjitodorov, "Acoustic analysis of pathological voices: A voice analysis system for screening of laryngeal diseases,"IEEE Eng. Med. Biol. Mag., vol. 16, pp. 74–82, Jul./Aug. 1997.
[9] Yaug song,Jian Huang,Ding Zhou,Hongyuan Zha and C.Lee Giles,"IKNN-Informative K-Nearest Neighbor Pattern Classification",PKDD 2007,PP.248-264.