International Journal of Current Engineering and Technology
ISSN 2277 - 4106

Available at http://inpressco.com/category/ijcet

Research Article

Fusion of MFCC & LPC Feature Sets for Accurate Speaker Identification

R. B. Shinde[Á*] and V. P. Pawar[Ḃ]

[Á]College Of Computer Science & Information Technology, Latur, (Maharashtra- India)
[Ḃ]Computer Science Dept. Swami Ramanand Teerth Marathwada University, Nanded, Maharashtra. India.

Accepted 10 November 2013, Available online 01 December 2013, **Vol.3, No.5 (December 2013)**

Abstract

Several feature extraction techniques are proposed for SRS. All these techniques achieve good recognition rate using ANN. To achieve good recognition rate with best performance is the objective of this paper. Using the two feature extraction techniques, features can concatenate & this technique will achieve 100% recognition rate with best performance using SCG algorithm. For this purpose MFCC & LPC techniques are used for extracting features.

Keywords: *SRS- Speaker Recognition System, ANN- Artificial Neural Network, SCG-Scaled conjugate gradient, MFCC-Mel Frequency Cepstral Coefficients, LPC- Linear Predictive Coding.*

1. Introduction

Today language identification and speech recognition systems are used for hand-free interaction with digital devices in conference hall by speaker or may be used by a handicapped candidate to use the computer skills for their purpose. Another application of this is speaker identification for biometric identification & verification. It has also applications in automatic language translation and routing incoming telephone calls to a human switchboard operator fluent in the corresponding language. Speech signal contains many levels of information i.e. a message is conveyed via the spoken words. At other levels, speech conveys the information about the language being spoken, the emotion, gender, and the speaker identity.

The automatic speaker recognition and speech recognition are very closely related. This paper deals with speaker recognition concept. Speaker recognition is commonly used in biometric system that has taken place for control of access to information services or user accounts on computers. Speaker recognition offers the ability to replace or augment the personal identification numbers and passwords with something that cannot be stolen or lost. For the proposed work a path way is followed as shown in the fig1.1.

The presented paper is deals with 6 sections. Section 1.gives Introduction , section 2. Deals with Feature extraction a) LPC b) MFCC, section 3.include Fusion of extracted features, section 4. Introduces Classification and section 5.gives Experimental results and conclusion.

2. Feature Extraction

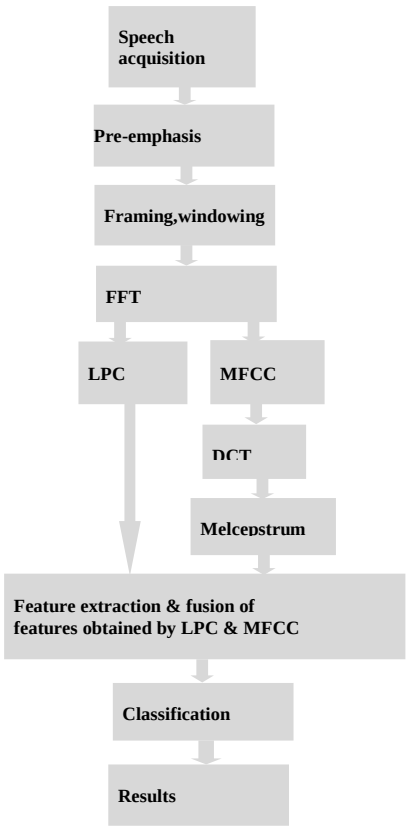

Fig.1: Path way for feature extraction & classification

Feature extraction is the most important phase in the speech processing. Speaker recognition is the process of automatically recognizing who is speaking based on unique characteristics contained in speech waves. This technique makes it possible to use the speaker's voice to verify their identity and control access to services such as

*Corresponding author: **R. B. Shinde**

voice dialing, data base access services, information services, voice mail, and security control for confidential information areas, and remote access to computers. In speaker recognition a premium is placed on extracting features that are somewhat invariant to changes in the speaker. So feature extraction involves analysis of speech signal. For extracting the features of speech signal two techniques are followed i.e. LPC & MFCC. Front end of analysis is the same for both the techniques using these technique features are extracted.

1. Pre-emphasis: The digitized speech signal, s(n), is put through a low order digital system, to spectrally flatten the signal and to make it less susceptible to finite precision effects later in the signal processing. The output of the pre-emphasizer network, is related to the input to the network, s(n) , by difference equation:

$$\tilde{s}(n) = s(n) - \tilde{a}s(n-1) \quad (1)$$

2. Frame Blocking: The output of preemphasis step, s~ (n) is blocked into frames of N samples, with adjacent frames being separated by M samples. If xl (n) is the l th frame of speech, and there are L frames within entire speech signal,then

$$x_1(n) = \tilde{s}(Ml + n) \quad (2)$$

where n = 0,1,…,N – 1
and *l* = 0,1,…,L – 1

3. Windowing: After frame blocking, the next step is to window each individual frame so as to minimize the signal discontinuities at the beginning and end of each frame. If we define the window as w(n), 0 ≤ n ≤ N– 1, then the result of windowing is the signal:

$$\tilde{x}_l(n) = x_l(n)w(n) \quad (3)$$

Where 0 ≤ n ≤ N – 1
Typical window is the Hamming window, which has the form

$$w(n) = 0.54 - 0.46\cos\left\{\frac{2\pi n}{N-1}\right\} \quad (4)$$

4. FFT: Fast Fourier Transformation (FFT) is calculated for each frame to extract frequency components of a signal in the time-domain. FFT is used to speed up the processing. FFT is extremely important in the area of frequency (spectrum) analysis because it takes a discrete signal in the time domain and transforms that signal into its discrete frequency domain representation. FFT is used to convert the convolution of the glottal pulse and the vocal tract impulse response in the time domain. FFT obtains by applying following equation

$$x_i(n) = e^{j\left[\frac{2\pi}{N}\right]in}\left[\sum_{K=0}^{N-1} u_n(k)e^{-j\left(\frac{2\pi}{N}\right)ik}\right] \quad (5)$$

LPC Feature Extraction Technique

LPC Analysis: The next processing step is the LPC analysis, which converts each frame of *p* + 1 autocorrelations into LPC parameter set by using Durbin's method. This can formally be given as the following algorithm:

$$E^{(0)} = r(0) \quad (6)$$

$$k_i = \frac{r(i) - \sum_{j=1}^{l-1}\alpha_j^{l-1} r(|i-j|)}{E^{i-1}} \quad 1 \le i \le p \quad (7)$$

$$\alpha_i^{(i)} = k_i \quad (8)$$

$$\alpha_i^{(i)} = \alpha_j^{(i-1)} - k_i\alpha_{i-j}^{(i-1)} \quad 1 \le j \le i-1 \quad (9)$$

$$E^{(i)} = (1-k_i^2)E^{i-1} \quad (10)$$

By solving (6) to (10) recursively for *i* = 1, 2,…,*p*, the LPC coefficient, *m a* , is given as

$$a_m = \alpha_m^{(p)} \quad (11)$$

Features extracted by the LPC are again statistically analyzed. Proposed work already has been done using LPC technique is discussed in.

MFCC Feature Extraction Technique

The most prevalent and dominant method used to extract spectral features is calculating Mel-Frequency Cepstral Coefficients (MFCC). MFCCs are one of the most popular feature extraction techniques used in speech recognition based on frequency domain using the Mel scale which is based on the human ear scale. MFCCs being considered as frequency domain features are much more accurate than time domain features.

Mel filter :The low-frequency components of the magnitude spectrum are ignored. The useful frequency band lies between 64Hz and half of the actual sampling frequency. This band is divided into 23 channels equidistant in mel frequency domain. Each channel has triangular-shaped frequency window. Consecutive channels are half overlapping. The choice of the starting frequency of the filter bank, fstart = 64Hz , roughly corresponds to the case where the full frequency band is divided into 24 channels and the first channel is discarded using any of the three possible sampling frequencies. We know that human ears, for frequencies lower than 1 kHz, hears tones with a linear scale instead of logarithmic scale for the frequencies higher that 1 kHz.

The mel-frequency scale is a linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz. The voice signals have most of their energy in the low frequencies. It is also very natural to use a mel-spaced filter bank showing the above characteristics.

mel1=sign(X_i).*log(1+af1/700)*k; (12)

where k=1.5 & X_i is the speech signal

Discrete cosine transform: DCT returns the unitary discrete cosine transform of x

$$y(k) = w(k)\sum_{n=1}^{N} x(n)\cos\frac{\pi(2n-1)(k-1)}{2N} \quad (13)$$

$$k = 1, \ldots\ldots\ldots, N$$

$$\text{Where } w(k) = \begin{cases} \frac{1}{\sqrt{N}} & k = 1 \\ \sqrt{\frac{2}{N}} & 2 \le k \le N \end{cases} \quad (14)$$

N is the length of x, and x and y are the same size. If x is a matrix, DCT transforms its columns. The series is indexed from $n = 1$ and $k = 1$ instead of the usual $n = 0$ and $k = 0$.

Mel spectrum:

The *Mel spectrum* is computed by multiplying the *Power Spectrum* by each of the *Triangular Mel Weighting filters* and integrating the result.

$$\tilde{s}[l] = \sum_{k=0}^{N/2} S[k] M_l[k] \qquad l = 0,1,\ldots\ldots,L-1 \tag{15}$$

S[k] is the power spectrum

N is the length of the Discrete Fourier Transform

L is total number of Triangular Mel weighting filters.

Statistical Analysis: For this work we use a simple statistical parameter Standard deviation. After applying the DCT on the speech signal a matrix is produced and generally it's very difficult to operate on such a large data. So we reduce that data by using standard deviation & we extract the 12 features from MFCC & 10 features from LPC.

$$s = \left(\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \right)^{\frac{1}{2}}$$

$$where$$

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{16}$$

3. Fusion of Features

Features extracted by LPC & MFCC technique are combined. Now we have total 15 samples from the 3 different speakers. In this way we get total 330 features from 3speakers. These features are given to ANN for classification

4. Artificial Neural Network

A Neural Networks are composed of simple elements operating in parallel. These elements are inspired by biological nervous systems. As in nature, the network function is determined largely by the connections between elements. We can train a neural network to perform a particular function by adjusting the values of the connections (weights) between elements. Commonly neural networks are adjusted, or trained, so that a particular input leads to a specific target output.

An Artificial Neural Network is used as recognition and identification method. The network has varying neurons input n, which receive input of LPC or MFCC or Both. Number of neurons in hidden layer varies from 5 to 20 neurons. In this paper we used a two-layer feed-forward network, with sigmoid hidden and output neurons, can classify vectors arbitrarily well, given enough neurons in its hidden layer. The network will be trained with scaled conjugate gradient back propagation. In this algorithm, input vectors and the corresponding target vectors are used to train a network until it can approximate a function, associate input vectors with a specific output vectors, or classify that vectors in an approximate way.

5. Experimental Results

The speech recognition system consists of MFCC and LPC-based two subsystems. These subsystems are trained by neural networks with MFCC and LPC features, respectively. The recognition process is realized by two stages: 1. In MFCC and LPC–based recognition subsystems recognition processes are realized in parallel. 2. The recognition results of MFCC and xLPC–based recognition subsystems are compared and the speech recognition system confirms the result, which confirmed by the both subsystems.

For the experiment purpose 5 samples from 3 speakers are taken so here we get total 15 samples. As per mentioned in section III features extracted by LPC are 10 & features extracted by MFCC are12for each sample. Fig 2. Depict that confusion matrix obtained after training and testing on LPC feature.

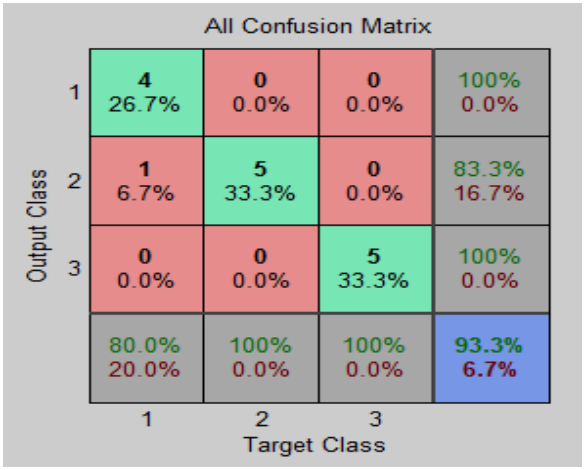

Fig. 2: Confusion matrix showing 93.3% recognition rate using LPC feature

In this case 93.3% recognition rate is obtained & 6.7 % is error rate is observed. While using LPC features for training & testing purpose 20 neurons are used in hidden layer. Table 1 depict the performance status for LPC feature extraction while classification.

Table 1 Performance Analysis of LPC feature extraction

Epoch	**30**
Time	0.00
Performance	0.430
Gradient	0.000106
Validation Checks	6
Neurons	20
Recognition Rate	93.3%
Error rate	6.7%
Feature Extraction Technique	LPC

When the training is given to MFCC features using SCG 100% recognition rate is obtained. Table 2 depicts the performance analysis of MFCC features.

Table2 Performance Analysis of MFCC feature extraction

Epoch	**50**
Time	0.00.26
Performance	0.527
Gradient	8.26
Validation Checks	6
Neurons	10
Recognition Rate	100%
Error rate	0.0%
Feature Extraction Technique	MFCC

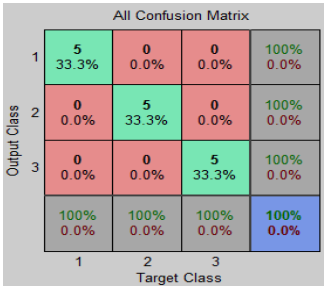

Fig. 3: Confusion matrix showing 100% recognition rate using MFCC feature

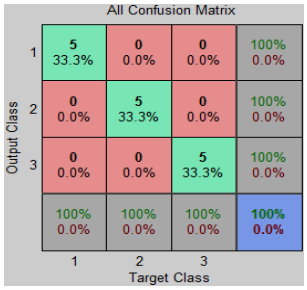

Fig. 4: Confusion matrix showing 100% recognition rate using LPC & MFCC features

These features are concatenated by obtaining 22 features per sample. Total 330 features are obtained for training & testing. In the next step these combined features are given to artificial neural network for classification purpose. Table 3 depicts the performance analysis of concatenated LPC+MFCC feature extraction. Fig 4 shows the obtained results after classification in confusion matrix format.

Table 3 Performance Analysis of LPC & MFCC Feature.

Epoch	**12**
Time	**0.00**
Performance	0.0356
Gradient	0.00481
Validation Checks	6
Neurons	**5**
Recognition Rate	100%
Error rate	0.0%
Feature Extraction Technique	MFCC + LPC

Table 4 Comparatives Analysis of LPC, MFCC & concatenated LPC+MFCC Performance.

Features	**LPC**	**MFCC**	**MFCC+LPC**
Epoch	**30**	**50**	**12**
Time	**0.00**	**0.00.26**	**0.00**
Performance	0.430	0.527	0.0356
Gradient	0.000106	8.26	0.00481
Neurons	**20**	**10**	**5**
Recognition Rate	**93.3%**	**100%**	**100%**
Error rate	6.7%	0.0%	0.0%

Fig 3 & 4 indicate that, MFCC Features & Concatenated MFCC+LPC features gives the 100% recognition rate. Where LPC features gives 93% recognition rate using SCG training algorithm. In case of both the feature extraction technique when we give the input vectors to the neural network first 20 neurons are used in case of LPC Features. 10 Neurons are used in case of MFCC features. When these two features are concatenated and given to neural network 100% recognition rate. Table 4 gives the comparative numerical information.

Conclusion

From experimental results, it can be concluded that features extracted by Linear Predictive Coding (LPC) obtains 93.3% recognition rate and features extracted using Mel Frequency Cepstral Coefficient (MFCC) obtain 100% recognition rate. In both the techniques SCG is used for training purpose. While comparing LPC & MFCC SCG can identify and recognize the speech signal better than using LPC by achieving 100% recognition .

When both the features are concatenated the highest identification and recognition rate i.e. 100% can be achieved. As depict in table 4 comparative analysis gives following some observations

1. LPC obtains 93.3% recognition rate in 30 epochs & required 20 neurons for better performance
2. MFCC obtains 100% recognition rate in 50 epochs & required 10 neurons for better performance
3. Concatenated features obtains 100% recognition rate in 12 epochs & required 5 neurons for better performance.

From these observations it can be conclude that concatenated features gives best performance with 100% recognition rate.

References

Lawrence r. Rabiner (1997), Applications Of Speech Recognition In The Area Of Telecommunicati0ns, 0-7803-3698-4/97, IEEE.

Transform Dr. H B Kekre1, Vaishali Kulkarni (Mar 2011), Speaker Identification using Row Mean of DCT and Walsh Hadamard *International Journal on Computer Science and Engineering (IJCSE)*, ISSN : 0975-3397 Vol. 3 No. 3.

L. Rabiner and B. H. Jung (1993), Fundamentals of Speech Recognition, *Prentice Hall, New Jersey.*

D. O. Shaughnessy (2001) , Speech Communication: Human and Machine, India, *University Press*.

Rohini B. Shinde & Dr. V. P. Pawar (May 2012) Combination of LPC & ANN for Speaker Recognition, *Journal Of Computing*, ISSN: 2151-9617 Vol. 4, No.5.

Alina Nica, Alexandru Caruntu, Gavril Toderean, Ovidiu Buza (May 2006), Analysis and Synthesis of Vowels Using Matlab, *IEEE Conference on Automation, Quality and Testing, Robotics*, Vol. 2, pp. 371-374.