Speech Dependent Speaker Classification Based Application Using DSPc6713

Maeira Imtiaz and Kamran A. Bhatti

Abstract—**Artificial intelligence advancements increased speech usage as an interface for different processes. The idea of Speaker based speech Recognition is basically to recognize automatically who is speaking what; on the basis of the individual information included in speech waves. That can only be accomplished using speech recognition technique along with speaker recognition algorithm. Among biometric recognition techniques, speech recognition is one of the most useful tools. Many banks, institutions, industries etc. are currently using speaker based speech recognition for providing greater security to their equipment. The idea is to implement this algorithm on voice controlled robot incorporated with security features by virtue of speech recognition algorithms. This robot will be capable of recognizing "What is being said and who has said it?" The algorithm is developed on MATLAB and implemented on a TMS320C6713 DSP kit which will interface with a robot – A voice controlled robot (VCR).**

Index Terms—**Mahalanobis distance, mel frequency cepstral coefficients, speech recognition, vector quantization.**

I. Introduction

For speech recognition, it is necessary to covert voice signals into features that can be processed digitally and that can differentiate efficiently between different words and speakers. There are many algorithms and techniques available to us whom we can use for voice signal analysis and speech recognition [1].

Saying technologically, we can distinguish between various speech signals using two different types of Automatic Speech Recognition:

1) Direct voice input
2) Continuous speech recognition

Direct voice input or DVI systems can recognize small to medium size vocabulary and immediate response to voice command is required whereas vocabularies of Continuous speech recognition or CSR systems are comprised of hundreds rather thousands of words and are not being used in real time processing until now [2]-[4].

ASR benefits include the provision of an extra communication channel in human-machine interaction (HMI), or in simple words it can be said that talking can be faster than typing.

Information provided by human voice contains identity of speaker, distinction among genders and emotions. In speaker based speech recognition, first of all, the speaker is identified and then his command is transferred for further processing for the purpose of speech recognition. Also there are two phases of ASR which are testing and training phase. Different techniques are used to reduce mismatch in testing and training phase which generally use spectral or cepstral domain.

Front-end and Back-end are two major parts of speech recognizer. Front-end extracts the features out of speech, while the back-end has a well-trained database and a classifier used for decision making purposes. The general layout of recognizer is represented in Fig. 1.

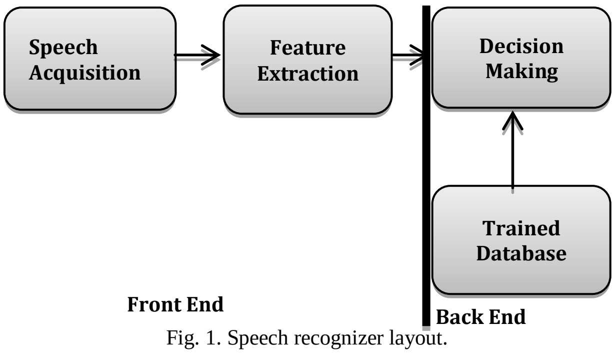

Fig. 1. Speech recognizer layout.

Basically front end involves the calculation of feature vectors which are extracted by the division the speech signal into short intervals called segments or frames. After division, a particular algorithm is applied on each segment to calculate the features in form of some coefficients.

After successful extraction of features they are passed to the back end classifier for further processing and decision making. Back end matches the upcoming feature vector with the stored trained database and makes its decision upon finding the best match among all of them [5].

II. Voice Recognition Algorithms

Two critical steps of voice recognition algorithms are:

1) Feature Extraction
2) Feature Matching

A. Feature Extraction

Some feature extraction techniques are discussed below:

1) *Linear Prediction Cepstral Coefficients (LPCC):* The idea of using Linear Prediction Cepstral Coefficients (LPCC) for speech recognition applications is common for many years. In this technique, a digital all-pole filter is used to model the human vocal tract [6]. The extraction of LPCC feature vectors from human speech is illustrated by the block diagram in Fig. 2.
2) *Perceptual Linear Prediction Coefficients (PLP):* Perceptual Linear Prediction (PLP) coefficient is based

Manuscript received August 8, 2015; revised November 27, 2015.

Maeira Imtiaz is with National University of Sciences and Technology, Pakistan. She is also with MOL Pakistan Oil and Gas Co. B.V., Pakistan (e-mail: maeiraimtiaz@yahoo.com).

Kamran A. Bhatti is with National University of Sciences and Technology, Pakistan (e-mail: kambhatti2k@gmail.com).

doi: 10.18178/ijke.2016.2.1.047

on emulating human auditory system. The critical concepts behind PLP to be understood are:

- Critical band frequency selectivity
- Equal-loudness curve
- Intensity-loudness power law.

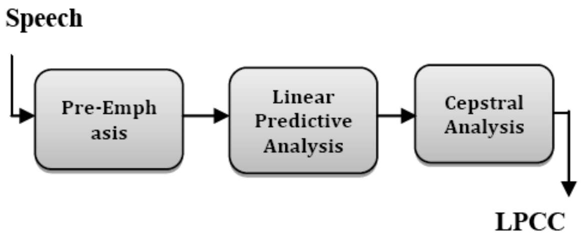

Fig. 2. LPCC extraction technique.

This technique consists obtaining the FFT-spectrum (First Fourier Transform) of a windowed speech frame. The obtained spectrum is passed through Bark-scale filter bank, through which the feature of “Critical Band Frequency Selectivity” of human cochlea is modelled. After frequency selection, to approximate the human sensitivity of, we require an equal-loudness curve to outputs of filter. Once weighted, the intensity-loudness power law (which states the relationship between the signal intensity and the perceived loudness) is used and filter outputs are compressed following it. Linear Predictive Analysis is performed after taking Inverse Fourier Transform of the compressed filter outputs. Finally, after application of cepstral analysis, set of PLP coefficients is obtained [7].

3) *Mel Frequency Cepstral Coefficients (MFCC):*The most common and robust feature extraction front-ends in speech recognition systems are Mel Frequency Cepstral Coefficients (MFCC). This technique is also based on Fast Fourier Transform, in which acquisition of feature vectors is obtained through frequency spectrum of the windowed speech frames [8], [9]. Fig. 3 illustrates the steps of MFCC extraction.

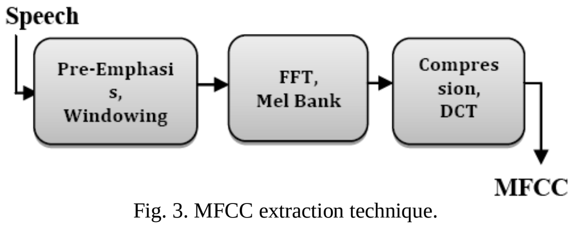

Fig. 3. MFCC extraction technique.

1) Comparison between different techniques

TABLE I: Recognition Correctness (C) Relative to Different Front-Ends V/S SNR Environments

Correctness C %	SNR (dB)				
	0	5	10	15	Clean
LPCC	9.97	12.48	15.77	18.02	58.64
MFCC	13.95	18.02	22.19	25.28	67.16
PLP	12.59	16.03	20.70	24.82	66.84

Results from Table I show that MFCC is most efficient among all of them.

2) MFCC extraction

1) *Pre-emphasis, Hamming windowing and FFT*: Pre-Emphasis is the first step of the algorithm to extract MFCC. Pre-emphasis is basically done to flatten the speech signal spectrally and to equalize the inherent spectral tilt of speech. Implemented by a first order FIR digital filter [10], [11], the transfer function of which is given in (1):

$$H_p(z) = 1 - az^{-1} \tag{1}$$

where a = 0.95

Comparison of speech signal before and after pre-emphasis is represented by Fig. 4 and Fig. 5.

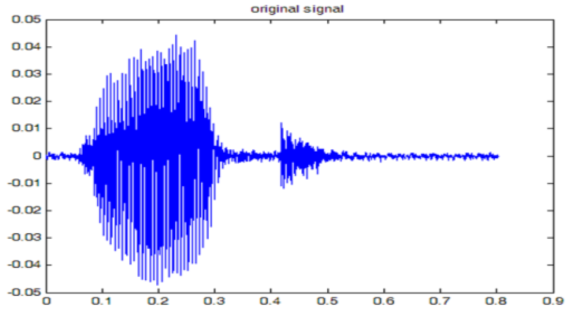

Fig. 4. Recorded signal by spoken word “HELLO”.

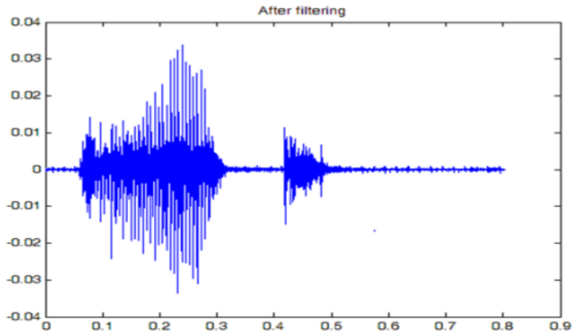

Fig. 5. Recorded signal after pre-emphasis.

After pre-emphasis, the division of speech signal in frames is carried out. In this process the entire speech sequence by a windowing function expressed in (2).

$$s_m[n] = s[n]w[n-m] \tag{2}$$

where:

$s[n]$: is the original speech sequence

$s_m[n]$: is mth windowed speech frame

w[n]: is the windowing function.

Generally the length of a frame is about 20-30ms. The shape of the windowing function is important. Rectangular window is not recommended since it causes severe spectral distortion (leakage) to the speech frames. Other types of windowing function, which minimize the spectral distortion, should be used. Hamming window is most commonly used given in (3):

$$w[n] = \begin{cases} 0.54 - 0.46cos\left(\frac{2\pi n}{N-1}\right), & 0 \le n \le N \\ 0, & otherwise \end{cases} \tag{3}$$

where:

N: is the length of the windowing function.

After Hamming windowing, the speech frame is passed to the next stage for further processing.

The next processing step is to convert the frame of N samples from the time to frequency domain by applying the Fast Fourier Transform [12]. The FFT of set of N samples is defined as (4):

$$X_n = \sum_{k=0}^{N-1} x_k e^{-2\pi jkn/N} \; n = 0 \ldots N-1 \tag{4}$$

The output of this step is called **signal's spectrum** or **period gram.**

2) *Mel scale filter bank:* The series of triangular band pass filters is called Mel Frequency Filter Bank. Mel frequency filter mimics the human auditory system. The filter bank is based on Mel scale which is basically a non-linear frequency scale. Below 1000Hz, the Mel scale is approximately linear to the linear frequency scale. Above the 1000Hz reference point, listener percept the same pitch increments with longer and longer frequency intervals. Hence the relationship between the Mel scale and the linear frequency scale is non-linear and approximately logarithmic above 1000Hz [13]. The mathematical relationship between the Mel scale and the linear frequency scale is described by relation given in (5):

$$f_{Mel} = 1127.01 \, ln\left(\frac{f}{700} + 1\right) \tag{5}$$

where

f: is the linear frequency (Hz)

f_{Mel}: is the Mel frequency (Mels).

3) *Discrete Cosine Transform (DCT):* The final step of the algorithm is to de-correlate the filter outputs. The first few coefficients obtained after the application of Discrete Cosine Transform (DCT) to the filter outputs, are grouped together to form a feature vector for a particular speech frame.

Consider the order of the Mel scale cepstrum is "*p*". To obtain the feature vector we take the first "*p*" *DCT* coefficients. Mathematically, the k^{th} MFCC coefficient can be expressed by (6):

$$MFCC_k = \sqrt{\frac{2}{M} \sum_{m=1}^{M} X_{m(n)} cos\left(\frac{\pi k\,(m-0.5)}{M}\right)} \quad 1 \le k \le p \tag{6}$$

Either log of energy component or a zero order coefficient (or both) are appended to the static feature vector. The zero order DCT coefficient of the filter output is basically the MFCC coefficient. The expression of the zero th order MFCC coefficient is described by (7):

$$MFCC_k = \sqrt{\frac{1}{M} \sum_{m=1}^{M} X_{m(ln)}} \tag{7}$$

B. Feature Matching

1) *Vector Quantization:* To map the vectors from a large vector space to a finite number of regions known as "Cluster" in that space is called Vector Quantization. Each cluster can be represented by its centre called a centroid. All code words collection constitutes a codebook.

The distance of the upcoming vector to the closest codeword of a codebookis called VQ-distortion and it is computed in the recognition phase. The speaker corresponding to the VQ codebook with smallest total distortion is identified.

2) *VQ Distortion Measuring Techniques:* To measure VQ Distortion we can have different type of Distance calculation Techniques given as follow :

Euclidean Distance: The distance calculation is carried out in following steps:

- Matching matrix dimensions
- Difference Calculation

Example:

Let's take Spoken word feature vector:

$$T = [T_1 T_2 T_3 T_4 \;\; T_5]$$

And Training Database:

$$U = \begin{bmatrix} a_1 a_2 a_3 a_4 & a_5 \\ b_1 b_2 b_3 b_4 & b_5 \\ c_1 c_2 c_3 c_4 & c_5 \end{bmatrix}$$

$$\boldsymbol{D(x)} = \sqrt{((\boldsymbol{T_1} - \boldsymbol{U(x)_1})^2 + (\boldsymbol{T_2} - \boldsymbol{U(x)_2})^2 + \cdots \, (\boldsymbol{T_n} - \boldsymbol{U(x)_n})^2} \tag{8}$$

$$D(x) = \sqrt{\sum_{i=1}^{n} (T_i - U(x)_i)^2} \tag{9}$$

where:

x: Row number of U (Rows of U represent different words or speakers)

Finally the absolute value of this distance is taken.Later the results for all the words are compared to find which word had the minimum distance and that word is then passed on for decision making [14].

- It is Easy to calculate
- Suitable for speaker recognition

But not robust for speech recognition because it is more complicated and required a more robust method for classification.

Mahalanobis Distance: Different patterns of variables can be identified or analyzed by correlation between variables and Mahalanobis Distance is calculated through these correlations [15]. The distance is calculated as follow:

- Matching the matrix dimensions
- Difference Calculation

Example:

Spoken word feature vector:

$$T = [T_1 T_2 T_3 T_4 \;\; T_5]$$

Training Database:

$$U = \begin{bmatrix} a_1 a_2 a_3 a_4 & a_5 \\ b_1 b_2 b_3 b_4 & b_5 \\ c_1 c_2 c_3 c_4 & c_5 \end{bmatrix}$$

$$D(x) = \sqrt{(T - U(x))^T * S^{-1} * \left(T - U(x)\right)} \tag{10}$$

where:

x: Row number of U (Rows of U represent different words or speakers)

S: Covariance matrix

Finally the absolute value of this distance is taken.Later the results for all the words are compared to find which word had the minimum distance. This word is then passed on for decision making.

III. Conclusions

In this paper voice recognition algorithms are discussed which are important in improving the voice recognition performance. MFCC provide a better recognition percentage as compare to PLP and LPCC as it is based on logarithmically spaced filter banks. A new technique of Mahalanobis Distance has been introduced in speech recognition process that takes both mean and variance of cluster in account to make results of feature matching phase more robust as compare to the outcome by usage of Euclidean Distance which provides results on the basis of mean of cluster only.

References

[1] L. Rabiner and B. H. Juang, "Fundamentals of speech recognition," - *Digital Signal Processing and Applications with the C6713 and C6416 DSK*.

[2] D. Rocchesso, "Introduction to sound processing," *Creative Commons Attribution-Share Alike License*, 2003.

[3] S. K. Mitra, *Digital Signal Processing — A Computer-Based Approach*, Second Edition, McGraw-Hill.

[4] S. Deketelaere, O. Deroo, and T. Dutoit, *Speech Processing for Communications: What's New?*

[5] E. Chandra, K. Manikandan, and M. Sivasankar, *A Proportional Study on Feature Extraction Method in Automatic Speech Recognition System*

[6] W. H. Abdullah, "Auditory based feature vectors for speech recognition systems," Electrical and Electronic Engineering Department, The University of Auckland.

[7] B. Logan, *Mel Frequency Cepstral Coefficients for Music Modeling*, Cambridge Research Laboratory, Compaq Computer Corporation.

[8] L. Muda, M. Begam, and I. Elamvazuthi, *Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques*.

[9] S. K. Hasnain and P. Akhter, "Digital signal processing: Theory & worked examples," *Discrete Time Signal Processing*, Second Edition, Prentice Hall.

[10] W. Han, C. F. Chan, C. S. Choy, and K. P. Pun, "An efficient MFCC extraction method in speech recognition," Department of Electronic Engineering, The Chinese University of Hong Kong, Hong, ISCAS, 2006.

[11] P. Kumar and P. Rao, "A study of frequency-scale warping for speaker recognition," Dept. of Electrical Engineering, IIT- Bombay, National Conference on Communications, NCC 2004, IISc Bangalore, 2004

[12] A. Alexander and A. Drygajlo, "Speaker identification: A demonstration using matlab," Swiss Federal Institute of Technology, Lausanne, Signal Processing Institute.

[13] D. C. Nguyen and H.-Y. Chung, *Performance Improvement of Microphone Array Speech Recognition using Features Weighted Mahalanobis Distance*, Department of Information and Communication, Yeungnam University, 2009.

Maeira Imtiaz did her graduation in the field of electrical engineering from NUST College of Electrical and Mechanical Engineering, Pakistan in the Year 2013.

She has been working with MOL Pakistan Oil and Gas Co. B.V., Pakistan, as an assistant electrical engineer (productions) for 2 years. Her research interests include electronics, signal processing, automation, embedded and communication systems. This research was a part of her final year project at College of Electrical and Mechanical Engineering, NUST, Pakistan.

Kamran A. Bhatti received the M.Sc. and MPhil degrees from Quaid-e-Azam University, Islamabad, Pakistan in 2000 and 2004, respectively.

He is the undergraduate coordinator and an assistant professor in Electrical Engineering Department of NUST College of Electrical and Mechanical Engineering, Rawalpindi, Pakistan. His research interests include speech/signal processing and methods for EEG-analysis, rehabilitation technology, brain–computer interface.