Speech processing for isolated Marathi word recognition using MFCC and DTW features

Mayur Babaji Shinde
Department of Electronics and Communication Engineering
Sandip Institute of Technology & Research Center, Nasik, Maharashtra, India

Dr. S. T. Gandhe
Department of Electronics and Communication Engineering
Sandip Institute of Technology & Research Center, Nasik, Maharashtra, India

Abstract- Even though speech recognition is a broad subject, the commercial and personal use implementations are rare. Although some promising solutions are available for speech synthesis and recognition, most of them are tuned to English. The acoustic and language model for these systems are for English language. Most of them require a lot of configuration before they can be used. If we go to the rural area across the India then still there are people who don't understand English as well as can't speak proper English. So this available speech recognition system is of no use for these people. There are so many regional languages in India, but being Maharashtrian we got inspired to think about Speech Recognition system for Marathi. There are several problems that need to be solved before speech recognition can become more useful. The amount of pattern matching and feature extraction techniques is large and the decision on which ones to use is debatable. After studying the history of speech recognition we found that the very popular feature extraction technique Mel frequency cepstral coefficients (MFCC) is used in many speech recognition applications and one of the most popular pattern matching techniques in speaker dependent speech recognition is Dynamic time warping (DTW). The signal processing techniques, MFCC and DTW are explained and discussed in detail and these techniques have been implemented in MATLAB.

Keywords – MFCC, DTW, FFT, FIR, DCT, HMM, Neural networks

I. INTRODUCTION

Keyboard, although a popular medium to input user data is not very convenient as it requires a certain amount of skill for effective usage. A mouse on the other hand requires a good hand-eye co-ordination. It is also cumbersome for entering non-trivial amount of text data and hence requires use of an additional media such as keyboard. Physically challenged people find computers difficult to use. Partially blind people find reading from a monitor difficult.

Current computer interfaces also assume a certain level of literacy from the user. It also expects the user to have certain level of proficiency in English. In India where the literacy level is as low as 50% in some states, if information technology has to reach the grass root level; these constraints have to be eliminated. Speech interface can help us tackle these problems. Speech synthesis and speech recognition together form a speech interface. A speech synthesizer converts text into speech. Thus it can read out the textual contents from the screen. Speech recognizer had the ability to understand the spoken words and convert it into text. We would need such softwares to be present for Indian languages.

The Speech is the most prominent and natural form of communication between humans. There are various spoken Languages throughout the world. Marathi is an Indo-Aryan Language, spoken in western and central India. There are 90 million of fluent speakers all over world. However; there is lot of scope to develop systems using Indian languages which are of different variations. Some work is done in this direction in isolated Bengali words, Hindi and Telugu. The amount of work in Indian regional languages has not yet reached to a critical level to be used it as real communication tool, as already done in other languages in developed countries [1]. Thus, this work was taken to focus on Marathi language. It is important to see that whether Speech Recognition System for Marathi can be carried out similar pathways of research as carried out in English. In this research we are presenting the work which consists of the creation of Marathi speech database and its speech recognition system for isolated words. This paper is divided into five sections, Section 1 gives Introduction, Section 2 deals with System description and

proposed algorithm, section 3 discusses experiment and result, Section 4 draws conclusion followed by section 5 with the references.

In this research work, we have developed an Automatic Marathi speech recognition system in MATLAB. System is developed for recognition of isolated Marathi words. Three databases of three different speakers are created and performance of system is verified on these three databases.

First the feature extraction from the speech signal is done by a parameterization of the wave formed signal into relevant feature vectors (MFCC). This parametric form is then used by the recognition system both in training the models and testing the same. Mel frequency Cepstral Coefficients (MFCCs) of Input speech samples from database are obtained first. These coefficients represent the features of each speech sample. After obtaining MFCCs of every speech sample, feature matching for new recorded sample is done using Dynamic Time Warping (DTW).

II. System Description and Proposed Algorithm

A. System block diagram –

Figure 1 shows the block diagram of proposed system. After acquiring an input voice signal from microphone it is processed in three steps. First preprocessing on signal is done. Then some useful and unique features are extracted from signal using MFCC and then these features are matched against the features of database using DTW algorithm. These three techniques are explained in following sections

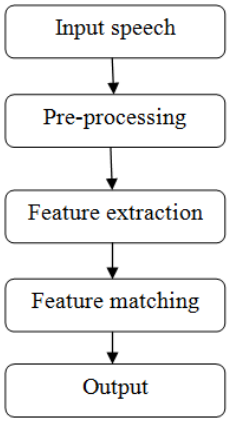

Figure 1. Block diagram of proposed system

B. Pre-processing –

In pre-processing different steps used is A/D conversion, pre-emphasis, use of noise gate and Alignment.

The acoustic sound must be A/D-converted to be digitally processed in the verification program. It is important to have a sufficient sampling rate to avoid aliasing [2]. Due to the characteristics of the human speech production the higher frequencies get dampened while the lower frequencies are boosted. To avoid that lower frequencies dominate the signal it is common to apply a high-pass FIR-filter before the feature extraction phase in speaker verification and speech recognition systems [3]. This process is called as pre-emphasis. The following first order FIR-filter can be applied to flatten the spectrum of the human voice-

$$F(z) = 1 - kz^{-1} \quad , 0 < k < 1 \qquad (1)$$

A noise-gate is used to remove background noise from a voice sample and thereby remove its influence on the feature vectors. After a noise-gate is applied the voice sample can be aligned to start from zero on the time axis. This can reduce some of the workload for the pattern matching process later in the program since the voice samples are much closer to each other than they otherwise would be [4].

C. Feature Extraction algorithm –

One of the most popular and effective ways is to use MFCC. The idea behind using MFCC is the assumption that the human hearing is an optimal speaker recognizer, although this has not yet been confirmed by studies [2]. The

feature extracting technique is decided on by considering several techniques. The use of Mel Frequency Cepstrum Coefficients (MFCC) is one of the most used feature extraction techniques in speaker recognition. [5][6][7] Studies show that MFCC parameters appear to be more effective then power spectrum based features when representing speech. In the MFCC feature extraction function there was a need to decide on parameters. Many of them had default values and these were confirmed to be reasonable and thus were used. The MFCC feature extraction process consists of six major steps. These steps are Framing, Windowing, Discrete Fourier Transforming, Mel-Frequency Warping, Log Compression, Discrete Cosine Transforming and Calculation of Delta and Delta-Delta Coefficients. Figure 2 shows the block diagram of MFCC

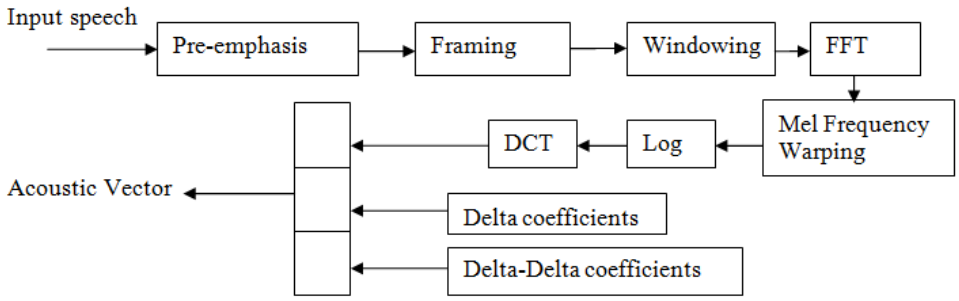

Figure 2. Block diagram of MFCC

Framing–

The first step is framing. The speech signal is split up into frames typically with the length of 10 to 30 milliseconds. The frame length is important due to the tradeoff between time and frequency resolution. If it is too long it will not be able to capture local spectral properties and if too short the frequency resolution would degrade. The frames overlap each other typically by 25% to 70% of their own length. Overlapping is used to make sure that each speech sound is approximately centered at some frame.

Windowing-

After the signal is split up into frames each frame is multiplied by a window function. A good window function has a narrow main lobe and a low side lobe. A smooth tapering at the edges is desired to minimize discontinuities. There are three options for windowing: Rectangular, Hanning and Hamming. The choice of windowing is made by verifying that the default windowing choice, Hamming windowing, is reasonable. Comparison between rectangular and Hanning windowing quickly ruled out rectangular windowing [8]. A comparison of triangular, rectangular and hamming window combined with either a linear scale and a mel scale shows that the combination of Hamming windows and the mel scale gives the best performance. Therefore the decision was made to keep the default choice of the function. The most common window used in speech processing is the Hamming window [2].
The Hamming window is defined as

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

Discrete Fourier Transforming-

The third step is to apply the discrete Fourier transform on each frame.

$$X_k = \sum_{n=0}^{N-1} x_n * e^{\frac{-2\pi ikn}{N}} \qquad ,k = 0, \dots \dots, N-1 \qquad (3)$$

The fastest way to calculate the DFT is to use FFT which is an algorithm that can speed up DFT calculations by hundred-folds [13].

Mel-Frequency warping-

The mel scale is based on how the human hearing perceives frequencies. It was defined by setting 1000 mels equal

to 1000 Hz as a reference point. Then listeners were asked to adjust the physical pitch until they perceived it as two-fold ten-fold and half, and those frequencies were then labeled as 2000 mel, 10000 mel and 500 mel respectively. The resulting scale was called the mel scale and is approximately linear below frequencies of 1000hz and logarithmic above. [2] The mel frequency can be approximated by the following equation-

$$mel(f) = 2595 * \log_{10}\left(1 + \frac{f}{700}\right) \quad (4)$$

where f is the actual frequency and mel(f) is the perceived one.

Log compression and discrete cosine transforming-

Step five is to apply compression by using a logarithm on the filter outputs Y(i) and then to apply the discrete cosine transform which yields the MFCCs c[n] according to the following formula [9]-

$$c[n] = \sum_{i=1}^{M} \log\big(Y(i)\big) * \cos\left(\frac{\pi n}{M}\left(i - \frac{1}{2}\right)\right) \quad (5)$$

Calculation of delta and delta-delta coefficients-

Delta and delta-delta coefficients are used to add time evolution information. The first order derivative is called delta coefficient and the second order derivative is called delta-delta coefficient [3].
The simplest way is to differentiate. The n^{th} delta feature is then defined by

$$\Delta f_k[n] = f_{k+M}[n] - f_{k-M}[n] \quad (6)$$

and the n^{th} delta-delta feature is accordingly defined by

$$\Delta^2 f_k[n] = \Delta f_{k+M}[n] - \Delta f_{k-M}[n] \quad (7)$$

where M typically is 2-3 frames. The differentiation is done for each feature vector separately [2].

D. Feature/pattern matching algorithm-

There are several popular classification techniques (pattern matching). Most popular are HMM, Neural networks, and DTW. Comparing HMM, GMM and DTW indicated that DTW was the more suitable classification technique. DTW works well even with small amount of enrollment data [12] and is also better at modeling temporal aspects of speech where HMM system is limited. DTW is also better than HMM for short duration phrases and word-spotting [11]. Also, when compared to neural networks, it takes long time to train neural network and choosing right structure of network, proper learning algorithm is time consuming process. So, studying all these aspects it was decided that DTW was a good choice for our program. When using DTW though there exists a tradeoff between recognition accuracy and computational efficiency. The reason for this is that creating several templates for a person's pass-phrase will increase the number of DTW paths that need to be computed every time before a decision is made. [10] Hence we chose DTW.

Dynamic Time Warping (DTW)-

The simplest way to recognize an isolated word sample is to compare it against a number of stored word templates and determine the best match. DTW is an instance of the general class of algorithms and known as dynamic programming. Its time and space complexity is merely linear in duration of speech sample and the vocabulary size. The algorithm makes a single pass though a matrix of frame scores while computing locally optimized segment of the global alignment path. The dynamic time warping algorithm provides a procedure to align in the test and reference patterns to give the average distance associated with the optimal warping path. [1]
This algorithm is for measuring similarity between two time series which may vary in time or speed. This technique also used to find the optimal alignment between two times series if one time series may be "warped" non-linearly by stretching or shrinking it along its time axis. This warping between two time series can then be used to find corresponding regions between the two time series or to determine the similarity between the two time series. Figure 4 shows the example of how one times series is 'warped' to another. [4]

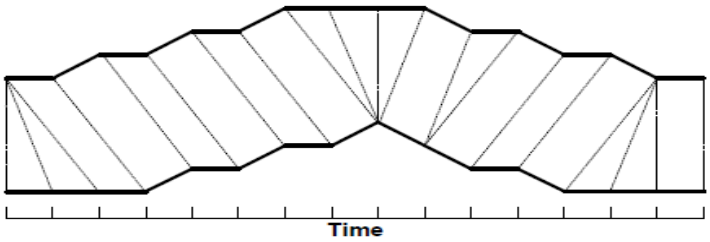

Figure 3. A Warping between two time series

the absolute distance between the values of two sequences is calculated using the Euclidean distance computation:

$$d(qi, cj) = (qi - cj)^2 \quad (8)$$

III. Experiment and Result

Three databases are created with three different male speakers of age between 20-25 years. Each database contains 72 marathi words. Each database is recorded in the silent room and microphone was held at the distance of 10-12 cm from the speaker. Experiment is conducted for offline as well as online (live) recognition of isolated mararthi words. Files are recorded in the format 'number.wav' (e.g. 1.wav, 2.wav, 3.wav). While recording all files sampling rate was kept at 8000 Hz which simply satisfies Nyquist criteria to avoid the effect of aliasing. Some waveforms belonging to database 1 are displayed below. First figure in figure 4 shows the nature of speech signal stored in 4.wav file i.e nature of Marathi word प्रसाद while second figure gives the discrete fourier transform of the same signal.

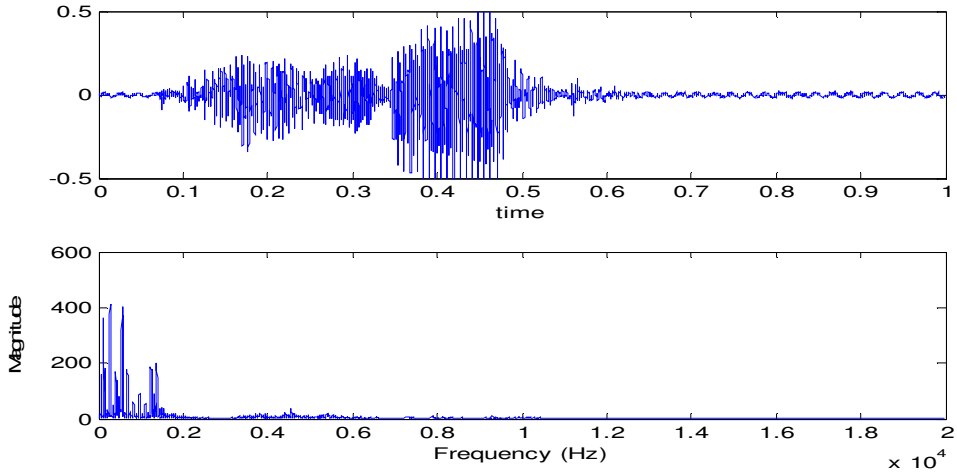

Figure 4. Speech signal & magnitude response for Marathi word प्रसाद

In offline speaker dependent Marathi word recognition, the system is trained on one of the three databases and tested on the same database used for training. While in offline speaker independent word recognition, the system is trained by one database and tested on other two databases. Table 1 shows recognition rates for above stated modes of speech recognition.

Trained by	Tested by		
	Database1	Database 2	Database 3
Database 1	100	26.38	27.77
Database 2	23.61	98.61	37.50

Database 3	**27.77**	**25**	**100**

Table 1. Recognition rate for offline speaker dependent and independent isolated Marathi word recognition

In online (live) speaker dependent word recognition, the system is trained on one of the three databases and tested on live voice of the speaker whose database is used for training.Table 2 shows the performance of speaker dependent online (live) isolated Marathi word recognition.

Trained by database 3	**Recognition Rate/ % Accuracy**
	Words
Live voice	**72.22**

Table 2. Recognition rate for live (online) speaker dependent Marathi word recognition

IV.CONCLUSION

While discussing above results, we tested the developed system for offline speaker dependent and speaker independent mode as well as online (live) speaker dependent mode. The system has been tested for Marathi words. Recognition rate for each mode has been calculated which is discussed in results. After observing these recognition rates it can be concluded that the combination of MFCC and DTW shows good performance for offline as well as online (live) speaker dependent speech recognition, but performance is poor for speaker independent mode. So, from this conclusion following observations can be made. In our experiment we found that the combination of MFCC and DTW gives good performance for speaker dependent live Marathi word recognition. Together with smaller adjustments and improvements of the weak spots of these two techniques, it can be concluded that a fully operational Marathi speech recognition program can be developed in a Matlab environment. Even though feature extraction and pattern matching are core functions of a speech recognition system there was a surprising amount of lateral problems to understand and overcome. The signal processing part of speech recognition for example is not to be underestimated.

V. REFERENCE

[1] Bharti W. Gawali, Santosh Gaikwad, Pravin Yannawar, Suresh C.Mehrotra4 “Marathi Isolated Word Recognition System using MFCC and DTW Features” Proc. of Int. Conf. on Advances in Computer Science 2010

[2] T. Kinnunen, "Spectral Features for Automatic Text-independent Speaker Recognition", 2003

[3] Ling Feng, "Speaker Recognition", 2004

[4] Lindasalwa Muda, Mumtaj Begam and I. Elamvazuthi “Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques”, Journal of computing, volume 2, issue 3, march 2010, ISSN 2151-9617

[5] Shi-Huang Chen, Yu Ren-Luo, "Speaker Verification Using MFCC and Support Vector Machine", 2009

[6] Ben J. Shannon, Kuldip K. Paliwal, "A Comperative Study of Filter Bank Spacing for Speech Recognition", 2003

[7] Carlos S. Lima, Adriano C. Tavares, Carlos A. Silva, Jorge F. Oliveira, "spectral Normalization MFCC Derived Features for Robust Speech Recognition", 2004

[8] Vibha Tiwari, "MFCC and its Applications in Speaker Recognition", 2010

[9] Huang X., Acero A., Hon H.-W. Prentice-Hall, "Spoken Language Processing: a Guide to Theory, Algorithm, and System Development", 2001

[10] Talal Bin Amin, Iftekhar Mahmood, "Speech Recognition Using Dynamic Time Warping", 2008

[11] Jean-François Bonastre, Philippe Morin, Jean-Claude Jonqua, "Gaussian Dynamic Warping Method applied to Text-Dependent Speaker Detection and Verification", 2003

[12] Atanas Ouzounov, "An Evaluation of DTW, AA and AVRM for Fixed-Text Speaker Identification", 2003

[13] Siddheshwar S. Gangonda, Dr. Prachi Mukherji “ Speech Processing for Marathi Numeral Recognition using MFCC and DTW Features” National Conference on Emerging Trends in Engineering & Technology (VNCET-30 Mar’12)