ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 3, Issue 12, June 2014

Automatic Speech Recognition of Gujarati digits using Dynamic Time Warping

Purnima Pandit, Shardav Bhatt
Department of Applied Mathematics, Faculty of Technology and Engineering,
The M. S. University of Baroda, Vadodara, India.

Abstract* — *In this article, we do Automatic Speech Recognition (ASR) for Gujarati digits using Dynamic Time Warping. For ASR, initially it is required to extract features of speech signal which is done using Mel Frequency Cepstral Coefficients (MFCC). Finally, recognition of the unknown speech signal is done with Dynamic Time Warping (DTW) algorithm.

***Index Terms* — Recognition, Gujarati Digits, Mel Frequency Cepstral Coefficients, Dynamic Time Warping.**

I. INTRODUCTION

In this era of electronic gadgets and intelligent machines, Automatic Speech Recognition (ASR) is highly applicable for our day-to-day task. It becomes boon if done in the native language. ASR is a process of automatically recognizing words, spoken by a human, using the information contained in a speech signal [1]. ASR is known as speech-to-text when the input speech signal, after being recognized by machine, produces the output as text. It is an ability of a computer to recognize in general, naturally flowing utterances from various speakers [2].

Speech-to-text processing is one of the most important applications of ASR. It is also applicable in hands-free computing, voice dialing, call routing, domestic appliance control, preparation of structured documents, health care, military, telephony, home automation, automated caller system, automated information system etc [3]. ASR is very helpful to the people with disabilities. Such people, who can't use machines due to their disabilities, can give instructions to machines by their speech using Speech Recognition technology. ASR with the output in the understandable form can be an important interface between a person who can speak and disabled one.

A speech signal, of a word spoken by human, contains important information such as gender, emotion, pitch and identity of a speaker [4]. These important features of speech signal are to be extracted in order to accurately recognize a word. The feature extraction process is done using Mel Frequency Cepstral Coefficients (MFCC). The cepstral coefficients represents speech signal, based on perception of human auditory system [5]. Not only two utterances of the same word spoken by different users differ but by a same user can also differ in time as shown in Fig.1. The matching of such speech signals in the MFCC form can be done using Dynamic Time Warping (DTW) algorithm. DTW aligns the word accurately and calculates the minimum distance between two words [6].

The data for our processing is speech signals for digits in Gujarati being spoken by ten different speakers. For these signals, computation of MFCC and then DTW matching is done using MATLAB code. In the next section we describe speech production process. Third section describes the feature extraction using MFCC and in the fourth section, DTW algorithm is introduced. The last section holds the experimental results and comparison followed by conclusion.

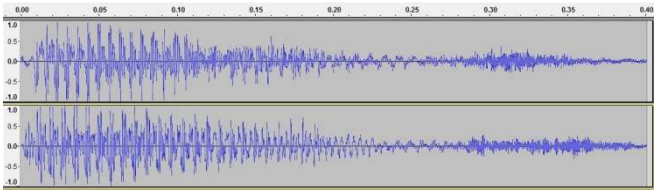

Fig.1. Speech Signal of Digit 5 ('paanch') spoken in Gujarati by two different speakers

II. SPEECH PRODUCTION

The process of Speech Production starts with a speaker formulating a message in his mind using the language that he wants to speak. When the message is converted into language code a set of phonemes (The smallest unit of speech sound) are involved with it. When a speaker is ready to execute this message, some neuro-muscular commands are produced from brain. This results in vibration of vocal cord and it gives some shape to vocal tract. The neuro-muscular commands also determine the duration of sound, loudness of sound and pitch. Based on shape of the vocal tract, a sound is produced as an acoustic signal [7].

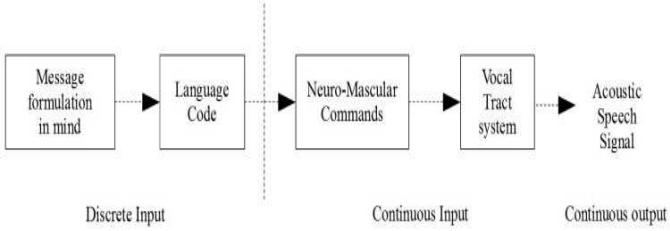

Fig.2. Speech Production Process

Vocal tract is a region starting from vocal cord and up to lips. The length of this region in average human is about 17 cm. The cross-sectional area of this region is varying from 0 to 20 cm^2. The shape of this region is determined by position of tongue, lips, jaws and velum. The combination of all this gives different shapes to vocal tract, which produces different sounds.

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 3, Issue 12, June 2014

In speech production process, along with mouth activity, nasal activity also takes place. This is represented by nasal tract which starts from velum and ends at nose. When velum is lowered, nasal tract and vocal tract works simultaneously and it produces a nasal sound. Diagrammatically, the speech production process can be summarized as shown in Fig.2.

The vocal tract which gives rise to the sound can be summarized by a schematic diagram as shown in Fig.3.

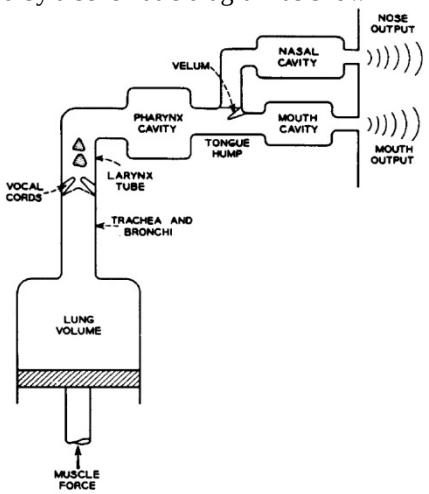

Fig.3. Schematic Diagram of Vocal Tract System [7]

A Speech signal is classified in following three states [7].

A. ***Silence State****:* A state in which no sound is produced.

B. ***Voiced State****:* A state in which vocal chords are vibrating periodically, when an air flows from lungs. A periodic triangular shaped pulse trail, called fundamental frequency, is generated. It has frequency 80 Hz to 350 Hz.

C. ***Unvoiced State****:* A state in which the vocal chords are not vibrating. So air passes through a narrow passage. This results in turbulence. A resulting acoustic wave is non-periodic and random in nature. We consider this as a noise signal.

The shape of the noise signal depends on the narrowness of the vocal tract. So any speech sound is combination of above three states of the speech signal.

After a speech is produced the next step is to recognize it. Speech recognition can be classified into different classes based on utterances. In this article, we will experiment with the isolated words (digits) spoken in Gujarati. The SR system with isolated words consists of silence in both sides of the speech sample. Single utterance is accepted at a time and speaker has to pause between each utterance. The processing of each utterance is done during a pause time. The other types of recognition involve connected words, continuous speech, spontaneous speech etc.

III. FEATURE EXTRACTION USING MEL FREQUENCY CEPSTRAL COEFFICIENTS (MFCC)

A speech signal is combination of the relevant information along with other irrelevant components like background noise, emotions etc. The process of capturing the vital information from the speech signal and discarding the rest is called feature extraction. This is done very efficiently with Cepstral Coefficients. Human ears are more sensitive to the higher frequencies. They cannot perceive frequencies higher than 1000 Hz. Therefore, for recognition purpose, the lower frequency components of the speech signal are more important than the higher frequency components [5].

There are many feature extraction techniques like Linear Predictive Coding (LPC), Linear Predictive Cepstral Coefficients (LPCC), Mel Frequency Cepstral Coefficients (MFCC), Human Factor Cepstral Coefficients (HFCC) etc. Here, we will use MFCC to extract features because it shows high accuracy result for clean speech. It is most commonly used algorithm for Speech Recognition. MFCC are best parametric representation of speech signal. They were introduced by Davis and Mermelstein in 1980 [5]. MFCC are best approximations of human ears since they are based on human hearing perception. The basic steps for calculating MFCC are as below and they are summarized in Fig.4.

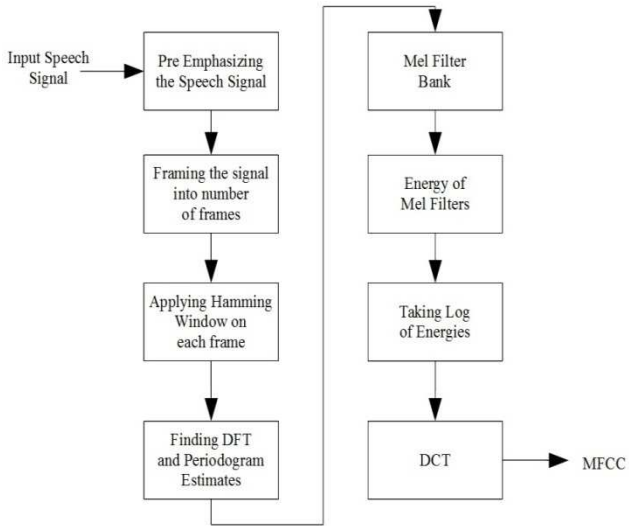

Fig.4. Steps to Calculate MFCC

Step-1. Pre-Emphasizing the speech signal

Step-2. Framing the signal into several number of frames

Step-3. Applying Hamming Window in Each Frame

Step-4. Taking Discrete Fourier Transform and calculating Periodogram Estimates of Power Spectrum

Step-5. Applying Mel Filter Bank and finding Energy of Mel Filter

Step-6. Obtaining MFCC

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 3, Issue 12, June 2014

Now we will discuss each step in detail. Let $x(n)$ be a recorded digital speech signal.

A. *Pre-Emphasizing*

The recorded digital speech signal has wide range and it also contains noise. We have to reduce this noise and make signal look spectrally flat. To do this we use first order high pass filter and from this, we get an emphasized version $s(n)$ of a recorded signal $x(n)$.

$$s(n) = x(n) - Ax(n-1);\ 0 \le A \le 1 \qquad (1)$$

Usually the value of A is taken as 0.95. It means that 95% of each sample is originating from its previous sample [7]. The aim of this step is to boost the amount of energy in the high frequencies. Due to this, the information from higher formants (which represents the frequencies that pass the most acoustic energy from source to the output [7]) is available to the acoustic model.

B. *Framing:*

Speech signal is a constantly varying signal. So it is difficult to analyze it entirely. But it does not vary much on a short time period. Therefore, we have to divide this signal into number of frames. The size of each frame is important because shorter frames will have fewer samples while the longer frames will contain a signal with large variation. The ideal frame size is 10-40 milliseconds (ms). The framing is done in such a way that each frame overlaps on its adjacent frame. So each frame has N samples out of which M samples are overlapping with the next frame. Naturally, $M < N$ and usually the values taken are $N = 256$ and $M = 100$. Hence, from a pre-emphasized signal, $s(n)$, we have $s_i(n)$, where n is the number of sample in each frame and i is the total number of frames.

C. *Windowing:*

This processed speech signal contains the unnecessary distortion. Moreover it is discontinuous in nature. So in this step of windowing the action will integrate the signal to all the closest frequency line and make a signal continuous. Usually Hamming window is used for this step [7]. The hamming window is defined as

$$w(n) = 0.54 - 0.46\ cos\left(\frac{2\pi n}{N-1}\right), 0 \le n \le N-1 \qquad (2)$$

We multiply each frame of the signal by a Hamming window as defined in (2). Therefore we get $s_i(n)w(n)$ for all i.

D. *Discrete Fourier Transform (DFT)*

The next step is to take Discrete Fourier Transform (DFT) of the windowed signal using FFT algorithm (Fast Fourier Transform, an efficient algorithm to calculate DFT). DFT of signal gives us the spectral information. Basically it represents energy level at different frequencies. To apply DFT for all i we have,

$$s_i(k) = \sum_{n=1}^{N} s_i(n)w(n)e^{\frac{-2j\pi kn}{N}} \qquad (3)$$

Here $1 \le k \le K$, where K is the length of DFT. From this we calculate energy level at different frequency using

$$p_i(k) = \frac{1}{N}\left|s_i(k)\right|^2 \qquad (4)$$

It is also called Periodogram estimates of the power spectrum.

E. *Mel Filter Bank*

According to Human ear perception experiments, the signal is perceived linearly for frequency less than 1000 Hz and for frequency greater than 1000 Hz, signal is perceived on logarithmic scale. Human ears are much better in detecting small changes in frequencies at low frequency level. So the information contained in low frequency components of the speech signal is more important compared to the high frequency component. The FFT Spectrum obtained in previous step is wide and the speech signal doesn't follow linear scale [5]. We use Mel scale to overcome this problem. The Mel scale associates this perceived frequency of the speech signal with the actual measured frequency. Any given Frequency f can be converted to Mel scale M using,

$$M(f) = 1125\ ln\left(1 + \frac{f}{700}\right)$$
$$\text{or} \qquad (5)$$
$$M(f) = 2595\ log\left(1 + \frac{f}{700}\right)$$

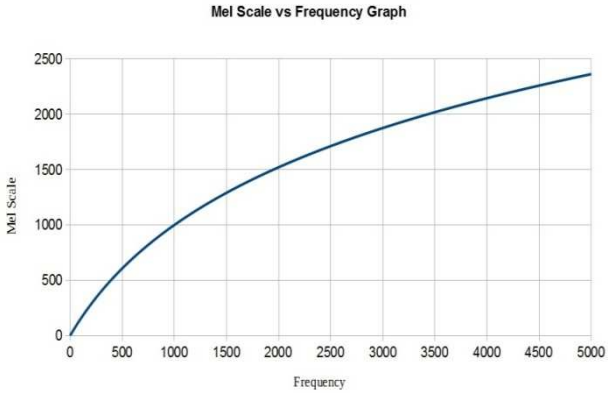

Fig.5. Mel Scale versus Frequency Graph

From Fig.5., it can be seen that for frequency less than 1000 Hz, the graph of Mel Scale versus frequency is nearly linear and beyond 1000 Hz, it is non-linear.

After converting frequencies to Mel scale, now we apply a filter bank consisting of 20-40 triangular shaped band pass filter on Mel Scale. Due to Mel scale, they are non-uniformly spaced in such a way that there are more filter in low frequency region and less filters in high frequency region. The first filter starts at first point with value zero. It takes its maximum value one at second point. Finally it comes back to zero at third point. The second filter starts at that point at which the first filter has its maximum value. Second filter

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 3, Issue 12, June 2014

takes its maximum value at third point and comes back to zero at fourth point and so on. The triangular filters used here are defined as

$$z_m(k) = \begin{cases} \frac{k-f(m-1)}{f(m)-f(m-1)} & f(m-1) \le k \le f(m) \\ \frac{f(m+1)-k}{f(m+1)-f(m)} & f(m) \le k \le f(m+1) \\ 0 & otherwise \end{cases} \quad (6)$$

Where, m is the number of filters used and $f(m)$ are the $m+2$ Mel spaced filters.

Now to obtain the filter bank energies, we multiply the periodogram estimates of power spectrum obtained in (4), with the filter defined above.

$$\tilde{p}_m = \sum_{k=0}^{K/2} p_i(k) z_m(k) \quad (7)$$

Here, K is the length of DFT and m is total number of filters, i is the frame number and $\tilde{p}_m$ is the Mel spectrum obtained from the original spectrum $p_i(k)$.

F. Calculating MFCC

This is the final step to obtain the feature vectors in the form of MFCC. To do this, first we take log of Mel filter bank energies calculated in (7). Now to come back to the time domain from the frequency domain, we have to take Discrete Cosine Transform (DCT). This results in a set of numbers for each frame and they are nothing but the Cepstral Coefficients. They are used as feature vectors containing speech signal for further processing. They are obtained using,

$$\widetilde{c_n} = \sum_{k=1}^{m} (log\ \widetilde{p_k})\ cos\left\{n\left(k-\frac{1}{2}\right)\frac{\pi}{2}\right\} \quad (8)$$

Here n is the number of Cepstral Coefficients in each frame and m is the number of filters in each frame [7].

IV. FEATURE MATCHING USING DYNAMIC TIME WARPING (DTW)

As discussed in previous section, a speech signal can be represented by a series of feature vectors for each frame, called MFCC. Now the next important process in Speech Recognition is of feature matching. In this step, we compare the feature vectors of same utterances by two different speakers. Out of these two speakers, the Cepstral Coefficients of utterance of one speaker are taken as templates. The Cepstral Coefficients of same utterance by another speaker are compared and matched with the templates and then the decision is made based on distance between the two sequences of feature vectors.

Here, the number of Cepstral Coefficients of same utterances by two different speakers need not be same because it depends on how fast or slow a speaker is speaking. So to compare these two sequences of feature vectors, their lengths have to be matched. Hence it gives rise to the problem of finding an optimal alignment between two vector sequences of unequal length.

Dynamic Time Warping (DTW) is an efficient algorithm to solve this problem. It is based on Bellman's principle of Dynamic Programming Problem. Using DTW we can find a non-linear alignment path between two vector sequences of unequal length. Here one of the vector sequences is warped non-linearly by stretching or shrinking its time axis. In Speech Recognition, this algorithm is very useful to measure similarities between two vector sequences varying with time. Fig.6. represents a non-linear alignment of the two signals of unequal length [6].

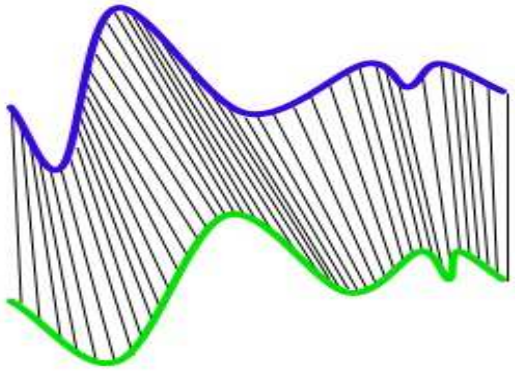

Fig.6. Non-Linear alignment using DTW

Now let us see how this algorithm works. Consider two vector sequences A and B of unequal length n and m respectively. $A = \{a_1, a_2, \dots, a_n\}$ and $B = \{b_1, b_2, \dots, b_m\}$, $m \neq n$. Using this principle, the global distance D can be iteratively calculated using,

$$D(i,j) = min\begin{cases} D(i,j-1) + d(a_i, b_j) \\ D(i,j) + 2d(a_i, b_j) \\ D(i-1,j) + d(a_i, b_j) \end{cases} \quad (9)$$

Where $d(a_i, b_j)$ is the local distance measured between a_i and b_j using Euclidean Distance formula.

V. EXPERIMENTAL RESULTS AND COMPARISON

For speech recognition of digits spoken in Gujarati, recording of digits spoken by ten different speakers is done using Audacity, a free and open source software. The recording is done with sampling rate 16,000 Hz and Mono channel. The speech signal for Gujarati digit spoken by each speaker was cut at zero-crossing using Audacity software. This gives hundred speech signals.

The speech signal files are processed for feature extraction, MFCC as described earlier using a MATLAB function file *mfcc.m*. The output of this file is the Cepstral coefficients, which are used in DTW algorithm. The MATLAB function file *dtw.m* is used to measure global distances between Cepstral coefficients of spoken digits by all speakers.

Out of ten speakers, one speaker's voice signals are taken as template. MFCC obtained for all the digits by that speaker is compared with the spoken digits of other nine speakers. The results for one such pair are as shown in the Table 1.

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 3, Issue 12, June 2014

Each entry in this table is the global distance found using DTW algorithm, between the Cepstral coefficients of the digits for a template and that spoken by other speaker. Similar tables were prepared for comparison of the spoken digits of template voice with all other speakers.

		SPEAKER									
TEMPLATE	Digits	1	2	3	4	5	6	7	8	9	10
	1	**2200**	4391	3083	4411	3082	3674	4453	3921	4293	2901
	2	3984	**2191**	2891	4489	3006	4025	3818	3538	2994	2453
	3	2645	4457	**1503**	5172	4044	3553	4551	4292	3321	3926
	4	6399	6857	3869	**3714**	4267	4294	**3128**	4301	6524	6391
	5	3356	4559	3084	3798	**2178**	3874	4227	**3065**	4185	2481
	6	3775	7193	3313	5305	5903	**3082**	5146	5925	4585	6579
	7	6932	6618	4736	4171	5277	5250	3520	4881	7302	7493
	8	4450	5364	2153	4908	4053	4111	4300	3313	4134	4038
	9	2900	5234	2748	7225	5942	4995	5975	5440	**2302**	4630
	10	3338	3100	3629	4838	2774	4569	4867	3650	4054	**1402**

Table 1. DTW distance measures

In Table 1, the column under title '1' holds the comparison of digit *one* by a speaker with the all digits of template. The cells with bold numbers represent least distances in each column. From this table we can conclude that 8 out of 10 digits by the selected speaker are matching with the template. This experiment was repeated with all speakers keeping one person as template. In our experiment, we got 84.44 % success. However, with the extra care taken during recording, the recognition success rate increased to 95.56 %.

VI. CONCLUSION

In this paper, we have done the speech recognition for Gujarati digits using DTW. For this purpose, the speech features were extracted using MFCC. In future we would continue with the phoneme based speech recognition for the sentences in Gujarati.

ACKNOWLEDGMENT

We are grateful to Mr. Priyank Makwana for his experiments in MATLAB and valuable discussion on Speech Recognition.

REFERENCES

[1] S. D. Dhingra, G. Nijhawan, and P. Pandit, "Isolated Speech Recognition using MFCC and DTW", International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, vol. 2, pp. 4085-4092, August 2013.

[2] P. Lama, M. Namburu. "Speech Recognition using Dynamic Time Warping", 2010.

[3] A. Thakur, R. Kumar, and N. Kumar, "Automatic Speech Recognition System for Hindi Utterances with Regional Indian Accents: A Review", International Journal of Electronics and Communication technology, vol. 4, pp. 38-43, June 2013.

[4] L. Muda, M. Begam, and I. Elamvazuthi, "Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques", Journal of Computing, vol. 2, pp. 138-142, March 2010.

[5] S. Davis, and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences", IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 28, 357-366, 1980.

[6] H. Sakoe, and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition", IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 26, pp. 43– 49, February 1978.

[7] L. Rabiner, B. H. Juang, and B. Yegnanarayana, "Fundamentals of Speech Recognition", Pearson Education, vol. ED-1, 2010.

AUTHOR'S PROFILE

Purnima Pandit UGC-CSIR NET (2001) qualified, PhD (2008) in the area of Control Theory and Artificial Neural Network. Life Member of ISTE, ISTAM and Gujarat Ganit Mandal. Presently she is working as Assistant Professor in the Department of Applied Mathematics, Faculty of Technology and Engg., The M .S. University of Baroda. Coordinator of M.Sc. (Financial Mathematics) Course from (2010). Her areas of interest are Dynamical Systems, Fuzzy Sets and Systems, Speech Recognition Financial Mathematics, Soft Computing, Optimization.

Shardav Bhatt Master of Science in Applied Mathematics with specialization in Industrial Mathematics (2013). Presently he is working as Teaching Assistant in the Department of Applied Mathematics, Faculty of Technology and Engg., The M .S. University of Baroda. Life Member of ISTE. His areas of interest are Speech Recognition, Artificial Neural Networks, Soft Computing, Signal Processing, Computational Fluid Dynamics.