Dynamic time Warping using MATLAB & PRAAT

Mrs. R. B. Shinde, Dr. V. P. Pawar

Abstract— The Voice is a signal of infinite information. Digital processing of speech signal is very important for high and precise automatic voice recognition technology. Speech recognition has found its application on various aspects of our daily lives from automatic phone answering service to dictating text and issuing voice commands to computers. We focus mainly on two points 1. The pre-processing stage that extracts salient features of a speech signal and a technique called Dynamic Time Warping commonly used to compare the feature vectors of speech signals. These techniques are applied for recognition of isolated as well as connected spoken words. 2. In this case the experiment is conducted in MATLAB & PRAAT to verify these techniques. Using the Matlab statistical difference between two wave signals are calculated. & using PRAAT two wave signals are aligned. Statistical difference obtained by these two ways is noticeable i.e. using MATLAB average distance is calculated which is 3.2843 & using PRAAT average distance is calculated which is 0.002235.

Index Terms— *SRS- Speech Recognition System, LPC- Linear Predictive Coding, DTW- Dynamic time warping, FFT- Fast Fourier transform, DCT-Discrete cosine transform.*

1 INTRODUCTION

One of the earliest approaches to isolated word speech recognition was to store a prototypical version of each word (called a template) in the vocabulary and compare incoming speech with each word, taking the closest match. This presents two problems: what form do the templates take and how are they compared to incoming signals. The simplest form for a template is a sequence of feature vectors that is the same form as the incoming speech. We will assume this kind of template for the remainder of this discussion. The template is a single utterance of the word selected to be typical by some process; for example, by choosing the template which best matches a cohort of training utterances. Comparing the template with incoming speech might be achieved via a pair wise comparison of the feature vectors in each.

The total distance between the sequences would be the sum or the mean of the individual distances between feature vectors. The problem with this approach is that if a constant window spacing is used, the lengths of the input and stored sequences is unlikely to be the same. Moreover, within a word, there will be variation in the length of individual [1]. The Dynamic Time Warping algorithm achieves this goal; it finds an optimal match between two sequences of feature vectors which allows for stretched and compressed sections of the sequence. The paper [2] gives a detailed description of the algorithm.

The presented paper is deals with 6 sections. Section 1 Introduction, 2 Speech Processing, 3 Dynamic time warping, 4 Experimental work, 5Praat sound Processing., 6. Formants & feature comparison ,7. Results, 8. Conclusion.

2 SPEECH PROCESSING

Speech processing follows the steps like pre-emphasis, framing , windowing, DFT etc.After speech processing features are extracted for further processing

2.1 Pre-emphasis wave:

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. This step process the passing of signal through a filter which emphasizes higher frequencies. This process will increase the energy of signal at higher frequency.

2.2 Sampling:

Sampling is a process of converting a continuous-time signal into a discrete-time signal. It is convenient to represent the sampling operation by a fictitious switch. The switch closes for a very short interval of time T, during which the signal presents at the output. The time interval between successive samples is T seconds and the sampling frequency if given by

$$f = \frac{1}{t} Hz \qquad (1)$$

2.3 Framing:

The process of segmenting the speech samples obtained from an ADC into a small frame with the length within the range of 20to 10 msec. The voice signal is divided into frames of N samples. Adjacent frames are being separated by M .To avoiding the frame overlapping problem the frame is shifted every 10 samples. The used values for N & M are 200ms & 10ms when the sampling rate of speech is 11025 Hz.

2.4 Windowing:

Next process is to apply window to each individual frame so

- *Ms. R. B. Shinde,Computer Science dept. College Of Computer Science & Information Technology,Latur, (Maharashtra- INDIA)PH-9822797930 rvmali007@gmail.com*
- *Dr. V. P. Pawar ,Associate Professor in Computer Science Dept,Swami Ramanand Teerth Marathwada University. Nanded, (Maharashtra- INDIA) PH-9604134298*

as to minimize the signal discontinuities at the beginning and end of each frame. A Hamming window is used for autocorrelation method in LPC. Hamming window has the form as given below.

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

$$0 \le n \le N-1$$

2.5 Fourier Transform:

Next step is to perform FFT. The FFT is based on decomposition and breaking the transform into smaller transform and combining them to get the total transform. FFT reduces the computation time required to compute a discrete Fourier Transform and improves the performance by a factor 100 or more over direct evolution of the DFT. implement the transform and inverse transform pair given for vectors of length by:

$$X\,(k) = \sum_{j=1}^{N} X\,(\,j)\,w_N^{(\,j-1)(k-1)} \qquad (3)$$

$$x(\,j) = (1/N)\sum_{k=1}^{N} X\,(k)\,w_N^{-(\,j-1)(k-1)} \qquad (4)$$

Where $w_N = e^{(-2\pi i)/N}$

2.6 Linear Predictive Coefficient:

Linear predicted Co-efficient: LPC determines the coefficients of a forward linear predictor by minimizing the prediction error in the least squares sense. It finds the coefficients of a nth-order linear predictor that predicts the current value of the real valued time series s(n) based on past samples.[4]

$$\tilde{s}(n) = -A(2) * X\,(n-1) - A(3) * X\,(n-2) -$$

$$A(N+1) * X\,(n-N)$$

(5)

n is the order of the prediction filter polynomial, a = [1 a(2) ... a(p+1)]. If n is unspecified, LPC uses as a default n = length(x)-1. If x is a matrix containing a separate signal in each column, LPC returns a model estimate for each column in the rows of matrix and a column vector of prediction error variances. The n is the order of the prediction filter polynomial, a = [1 a(2*) ... a(p+1)].* If n is unspecified, LPC uses as a default *n = length(x)-1.* If x is a matrix containing a
separate signal in each column, LPC returns a model estimate for each column in the rows of matrix and a column vector of prediction error variances. The length of n must be less than or equal to the length of x.

2.7 Discrete Cosine Transform:

DCT can be used to achieve the coefficients. DCT reconstruct a sequence very accurately from only a few DCT coefficients, a useful property for applications requiring data reduction. DCT returns the discrete cosine transform of X. The vector Y is the same size as X and contains the discrete cosine transform coefficients[15]

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

k=1,……………, N. (6)

Where $wk = \begin{cases} \dfrac{1}{\sqrt{N}} & k=1 \\ \dfrac{\sqrt{2}}{N} & 2 \le k \le N \end{cases}$

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.

2.8 Cepstral Analysis:

Cepstral analysis is a nonlinear signal processing technique that is applied most commonly in speech processing and homomorphic filtering returns the complex cepstrum of the real data sequence x using the Fourier transform. The input is altered, by the application of a linear phase term, to have no phase discontinuity at ±π radians. That is, it is circularly shifted (after zero padding) by some samples, if necessary, to have zero phase at π radians.

3 Dynamic Time Warping

DTW algorithm is based on Dynamic programming. This algorithm is used 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 time series if one time series may be wrapped non-linearly by stretching or shrinking it along its time axis. This wrapping between two time series can then be used to find corresponding regions between the two time series to determine similarity between the two time series. DTW provides a procedure to align in the test and reference pattern to give the average distance associated with the optimal wrapping path [9].

Dynamic time warping (DTW) is a time series alignment algorithm developed originally for speech recognition [8]. It aims at aligning two sequences of feature vectors by warping the time axis iteratively until an optimal match between the two sequences is found.

Consider two sequences of feature vectors:

A=a1, a2, a3, ………………, an

B=b1, b2, b3, ……………….., bn

The two sequences can be arranged on the sides of a grid, with one on the top and the other up the left hand side. Both sequences start on the bottom left of the grid. Figure 1. shows the schematic representation of aligning two sequences along the grid.

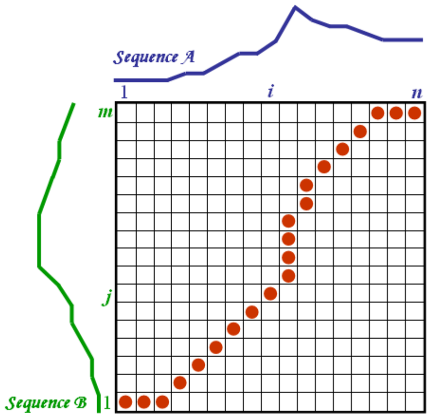


Figure 1 : schematic representation dynamic time warping of two Sequence A & sequence B along the grid.

Inside each cell a distance measure can be placed, comparing the corresponding elements of the two sequences. To find the best match or alignment between these two sequences one need to find a path through the grid which minimizes the total distance between them. The procedure for computing this overall distance involves finding all possible routes through the grid and for each one compute the overall distance. The overall distance is the minimum of the sum of the distances between the individual elements on the path divided by the sum of the weighting function. The weighting function is used to normalize for the path length. It is apparent that for any considerably long sequences the number of possible paths through the grid will be very large. The major optimizations or constraints of the DTW algorithm arise from the observations on the nature of acceptable paths through the grid: Following are the feature due to which DTW is becomes more popular.

Today's detection techniques can accurately identify the starting and ending point of a spoken word within an audio stream, based on processing signals varying with time. They evaluate the energy and average magnitude in a short time unit, and also calculate the average zero-crossing rate. Establishing the starting and ending point is a simple problem if the audio recording is performed in ideal conditions. In this case the ratio signal-noise is large because it's easy to determine the location within the stream that contains a valid signal by analyzing the samples. In real conditions things are not so simple, the background noise has a significant intensity and can disturb the isolation process of the word within the stream.

DTW is guaranteed to find the lowest distance path through the matrix, while minimizing the amount of computation. The DTW algorithm operates in a time-synchronous manner: each column of the time-time matrix is considered in succession (equivalent to processing the input frame-by-frame) so that, for a template of length N, the maximum number of paths being considered at any time is N. If $D(i,j)$ is the global distance up to (i,j) and the local distance at (i,j) is given by $d(i,j)$

$$d(i,j)=min[d(i-1,j-1),d(i-1,j),d(i,j-1)]+d(i,j) \quad –(7)$$

Given that D(1,1) = d(1,1) (this is the initial condition), we have the basis for an efficient recursive algorithm for computing D(i,j). The final global distance D(n,N) gives us the overall matching score of the template with the input. The input word is then recognized as the word corresponding to the template with the lowest matching score.

A matching path technique basically follows the next steps to calculate the distance between two sequences:[10]

1. A matching path needs to be created. A matching path is a list of combinations of the points of the first sequence and the points of the second sequence. The technique used for the creation of this list is what distinguishes different matching-path-using methods, of which DTW is one
2. For each of the combinations of points i, j in the matching path, the distance D(i,j) between them is calculated. Various methods for calculating the distance between two points exist. This DTW-implementation uses Euclidean distance Eq. 8
3. The distances calculated in the previous step are summed and this total distance is normalized by divided it by the number of combinations in the matching path. The resulting value is the distance between the sequences.
   The Euclidean distance between two points A= (a1, b1, c1) and B= (a2, b2, c2) is calculated as

$$|AB| = \sqrt{(a_1 - a_2)^2 + (b_1 - b_2)^2} \quad \text{--(8)}$$

4. Experimental Work

For the experiment purpose isolated Uttered words are taken & analyzed. Experiment is done using the Matlab for experiment connected words are used which uttered by 6 different persons. Using Matlab Statistical data is obtained for the connected word

Table1 . Result obtained by DTW for 5 speakers for word "BEED" using Matlab

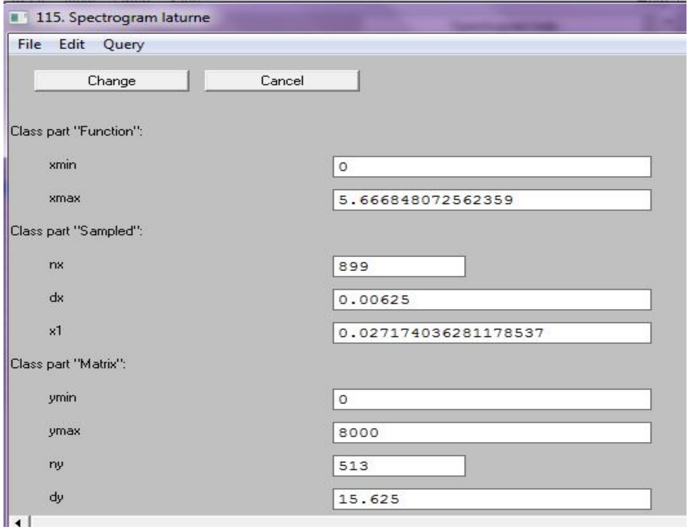
		P 1	P 2	P 3	P 4	P 5
		BEED	BEED	BEED	BEED	BEED
P 1	BEED	0	3.9986	2.6523	3.9635	2.3056
P 2	BEED	3.9986	0	3.4335	3.0589	3.7746
P 3	BEED	2.6523	3.4335	0	2.9833	3.4902
P 4	BEED	3.9635	3.0589	2.9833	0	4.2635
P 5	BEED	2.3056	3.7746	3.4902	4.2635	0

5. PRAAT Sound processing

After the statistical analysis using Matlab same wave files are used in PRAAT for aligning the wav files. This is a freeware program for the analysis and reconstruction of acoustic speech signals. PRAAT is a very flexible tool to do speech analysis. It offers a wide range of standard and non-standard procedures, including spectrographic analysis, articulatory synthesis, and neural networks [6, 7]. For the experiment five connected uttered words are used & these spoken word wav files are treated using spectrogram object.

5.1 SPECTROGRAM

A spectrogram object represents an acoustic time-frequency representation of a sound: the power spectral density *PSD (f, t),* expressed in Pa² / Hz. It is sampled into a number of points centered around equally spaced times t_i and frequencies f_i.



Class part "Function":

xmin 0

xmax 5.666848072562359

Class part "Sampled":

nx 899

dx 0.00625

x1 0.027174036281178537

Class part "Matrix":

ymin 0

ymax 8000

ny 513

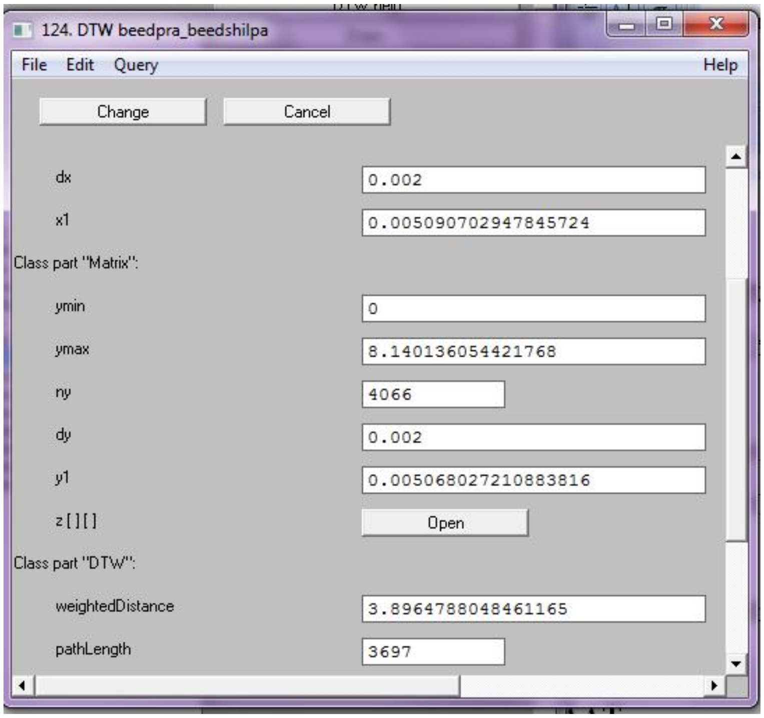
dy 15.625

Figure 2: Screen shot for spectrogram object specifications for a wav file

Figure 2 will be displayed by Inspect option. It gives following attributes:

- *xmin*: Start time of wav in seconds.
- *xmax*: end time of wav in seconds.
- *nx*: the number of times (≥1).
- *dx*: time step in seconds.
- *xl*: the time associated with the first column in seconds. This will usually be in the range [*xmin, xmax*]. The time associated with the last column (*i.e., xl* + (*nx-1*) *dx*))will also usually be in that range.
- *ymin*: lowest frequency, in hertz. Normally 0.
- *ymax*: highest frequency, in hertz.
- *ny*: the number of frequencies (≥1).
- *dy*: frequency step in hertz.
- *y1*: the frequency associated with the first row, in hertz. Usually *dy/2.* The frequency associated with the last row, (i.e., *yl* + (*ny*-1) *dy*)) will often be *ymax* – *dy /2.*
- zij,i= 1….ny, j=1…………nx: the power spectral density in Pa2 / Hz.

5.2 DTW USING PRAAT



dx 0.002

x1 0.005090702947845724

Class part "Matrix":

ymin 0

ymax 8.140136054421768

ny 4066

dy 0.002

y1 0.005068027210883816

z[][]

Class part "DTW":

weightedDistance 3.8964788048461165

pathLength 3697

Figure 3: Depicts the DTW in PRAAT for two wav signals**.**

Using the PRAAT, DTW is calculated & path is draw using DRAW option. It gives following drawing for the connected word uttered by five different persons. Figure 3 shows the screenshot for DTW in PRAAT for two wav signals.
For comparing the wave files to find out DTW we use following technique shown in Figure 4

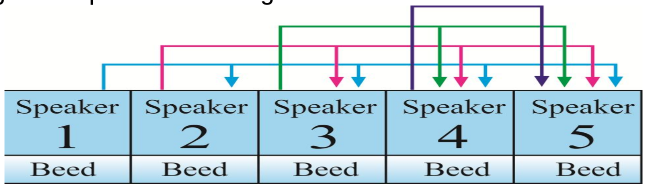


Figure 4: Illustration of sample comparison**.**

In comparison, samples of speaker 1 are compared with samples of speaker 2, 3, 4, 5. It is indicated with blue colored arrow in figure 2. Similarly to obtain more alignment comparison is done for samples of speaker 2 with samples of speaker 3, 4, 5. Comparison is done for samples of speaker 3 with samples of speaker 4, 5. Comparison is done for samples of speaker 4 with samples of speaker 5. In this way more possible combinations are taken & obtained 10 alignment for DTW. Figure 5 shows the alignment path for comparing wave signals.

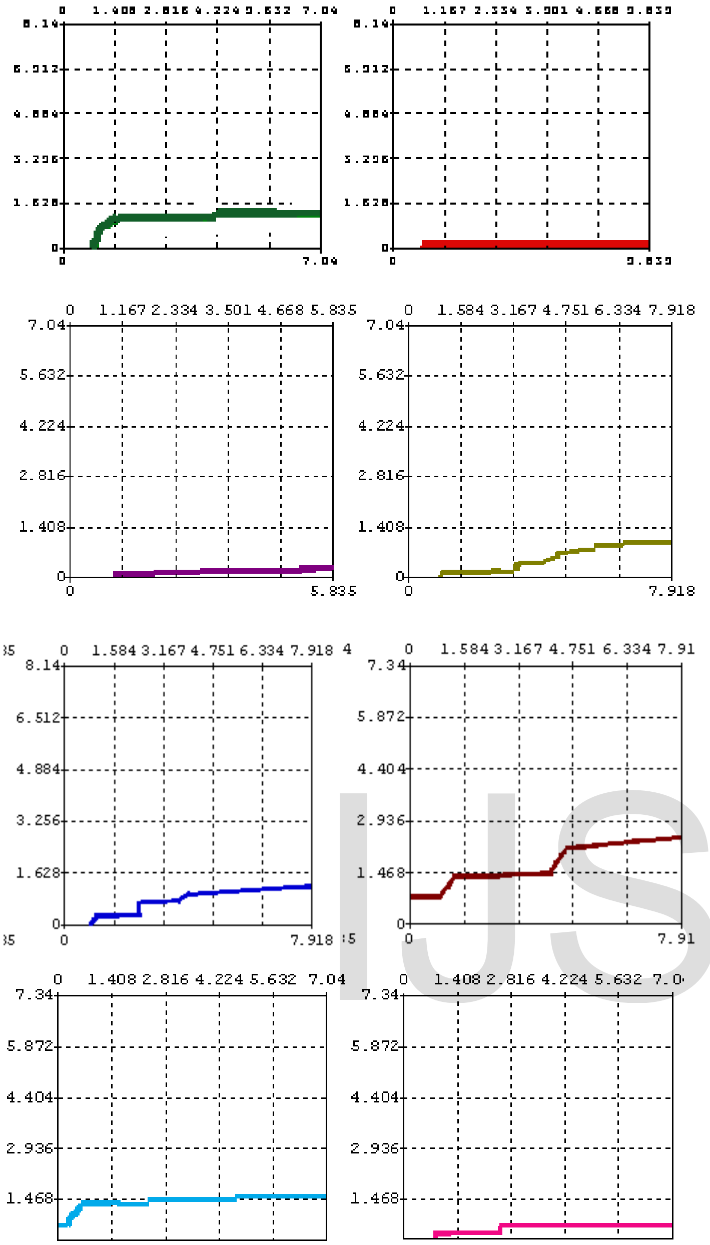


Figure 5: Result obtained by DTW for 5 speakers for word

"BEED" using PRAAT

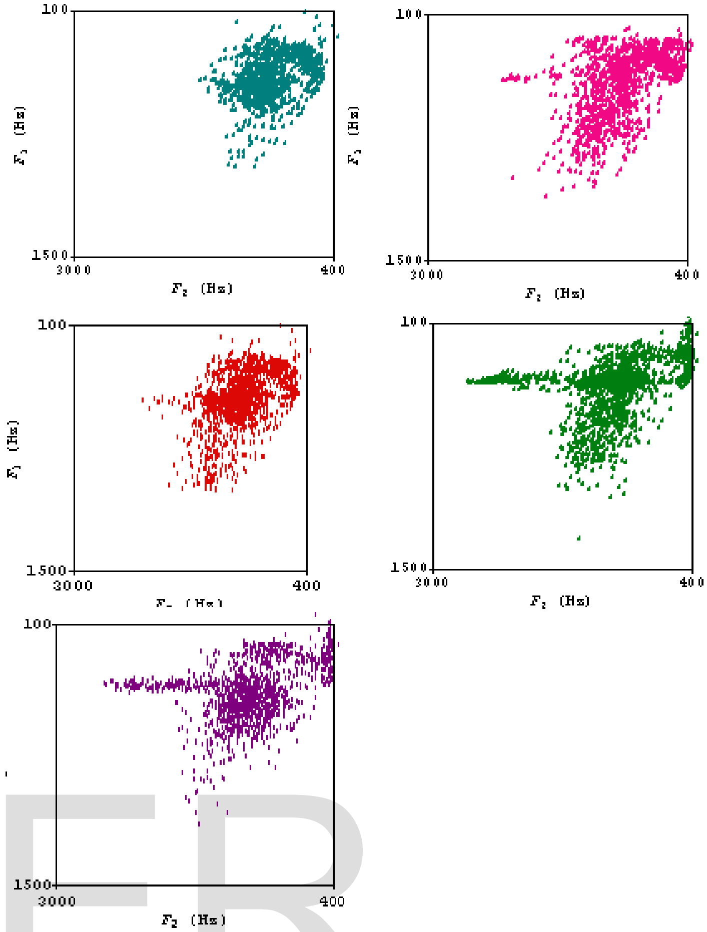


Figure 6: Scattered plots for the words for 5 speakers.

6. Formants Feature Comparison

It has been known for many years that formant frequencies are important in determining the phonetic content of speech sounds. Several authors have therefore investigated formant frequencies as speech recognition features, using various methods for basic analysis, such as linear prediction [11], [12], analysis by synthesis with Fourier spectra [13], and peak picking on cepstrally smoothed spectra [14].For the experiment we select some vowels & words for extracting the features & lateron these features are clustered for comparisons using praat.

Figure 6 depicts the scattered plots of 5 words uttered by 5 speakers. Words uttered by 5diiferent speakers obtains similar formants that's why the patterns of points fall on scatter plot are similar for the same word uttered by dif speaker.

For scatter ploting the features are extracted using praat. Each signal is processed using LPC LPC represents filter coefficients as a function of time. The coefficients are represented in frames with constant sampling period.After that formants are applied to find the formants feature. A formant object represents spectral structure as a function of time. A formant contour. Nlike the time stamped formant grid object, it is samples into a number of frames centered around equally spaced times, each frame contains frequency and bandwidth information about several formants

7 Results

The input voice signals of different and same speakers have been taken and compared. The results obtained are given in the table 1. Aim if this research work is to compare the performance of LPC & DTW using MATLAB with DTW in PRAAT. The speech data used in this experiment are connected words specifically we take the names of cities. The test pattern is compared with the reference pattern to get the best match. Form the analysis of result obtained from MATLAB we get two results.

1. Patterns of the speaker 1 are compared with speaker 1 then perfect match found.
2. Patterns of speaker 1 are compared with patterns speaker 2, 3, 4, & 5 for same word then it gives the difference as shown in Table 1. The differences of different speaker pattern are fall in same range i.e. 2.3056 to 4.263 Average difference is 3.2843

From the analysis of PRAAT results obtained in alignment format Figure 6 depicts that 8 samples are compared & all these 8 samples are aligned in maximum distance i.e. *2.347* & minimum distance i.e. 0.002. Using PRAAT path alignment is obtained very easily. Combined alignment of compared wave signal is given in the figure 7.

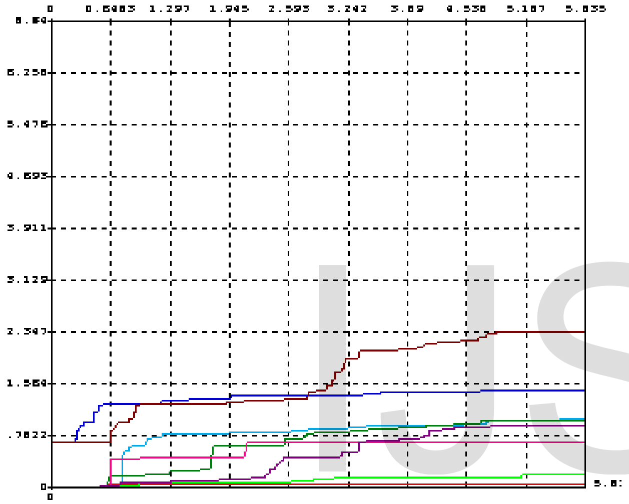


Figure 7: Result obtained by DTW for 5 speakers for word “BEED” using PRAAT

8. CONCLUSION

From the result we can conclude that PRAAT gives minimum distance while comparing wave signals as compared to MATLAB function. The differences of different speaker pattern are fall in same range i.e. 2.3056 to 4.263 Average difference is *3.2843* it is obtained using MATLAB. The difference of different speaker samples are aligned in maximum distance i.e. *2.347* & minimum distance i.e. 0.002. & average distance is *0.0022347* which is very less than Matlab function. PRAAT gives the better result as compared to MATLAB.

PRAAT gives flexibility while working on 1) Finding information in the Manual. 2) Create a speech object 3) Process a signal 4) Label a waveform 5) General analysis (waveform, intensity, sonogram, pitch, duration) 6) Spectrographic analysis 7) Intensity analysis 8) Pitch analysis 9) Using Long Sound files.

9 REFERENCES

[1] Steve Cassidy “Speech Recognition” Department of Computing, Macquarie University Sydney Australia (2002)

[2] H. Sakoe & S. Chiba “Dynamic programming algorithm optimization for spoken word recognition.” IEEE Trans. Acoustics, Speech and Signal Proc., Vol. ASSP-26, No. 1, page no. 43-49, Feb. 1978.

[3] F. Jelinek. L. R. Bahl and R. L. Mercer, "Design of a linguistics Statistical decoder for the recognition of continuous speech”, IEEE Trans. Informat. Theory, Vol IT-21 PP. 250-250, 1975.

[4] Michael Grinm, Kristian Kroschel and Shrikanth Narayanan, “The vera AM Mittag German Audio-Visual Emotional Speech Database.

[5] H. Ney. “The use of a one-stage dynamic programming algorithm for connected word recognition.” IEEE Transactions on Acoustics, Speech and Signal Processing, ASSP-32(2):263–271, April 1984.

[6] Pascal van Lieshout “ PRAAT Short Tutorial A basic introduction” V. 4.2.1, October 7, 2003 (PRAAT 4.1.x)

[7] Pauline welby and Kiwako Ito “PRAAT Tutorial” January 13, 2002.

[8] C. Myers, L. Rabiner, and A. Rosenberg, \Performance tradeoffs in dynamic time warping algorithms for isolated word recognition," Acoustics, Speech, and Signal Processing [see also IEEE Transactions on Signal Processing], IEEE Transactions on, vol. 28, no. 6, pp. 623-635, 1980.

[9] Palden Lama and Mounika Namburu “ Speech Recognition with Dynamic Time Warping using MATLAB” PROJECT REPORT 1. SPRING 2010

[10]Bharti W. Gawali, Santosh Gaikwad, “Marathi Isolated Word Recognition System using MFCC and DTW Features” Proc. of Int. Conf. on Advances in Computer Science ACEEE 2010 Page no-143-146.

[11] M.J. Hunt, "Delayed Decisions in Speech Recognition - The Case of Formants", Pattern Recognition Letters, Vol. 6, pp. 121-137, July 1987.

[12] P. Schmid and E. Barnard, "Robust, N-Best Formant Tracking", Proc. EUROSPEECH'95, pp. 737-740, Madrid, 1995.

[13] L. Welling and H. Ney, "A Model for Efficient Formant Estimation", Proc. IEEE ICASSP, pp. 797-800, Atlanta, 1996.

[14] Y. Laprie and M.-O. Berger, "Active Models for Regularizing Formant Trajectories", Proc. ICSLP, pp. 815-818, Banff, 1992.

Ms. R. B. Shinde received the M.Sc. (CS) degree from Dr. B. A. M. University, Aurangabad, in the year 2001. She is currently working as lecturer in the College of Computer Science and Information Technology, Latur, Maharashtra. She is leading to Ph. D degree in S.R.T.M. University, Nanded.

Dr. Vrushsen V. Pawar received MS, Ph.D. (Computer) Degree from Dept .CS & IT, Dr. B. A. M. University & PDF from ES, University of Cambridge, UK. Also Received MCA (SMU), MBA (VMU) degrees respectively. He has received prestigious fellowship from DST, UGRF (UGC), Sakaal foundation, ES London, ABC (USA) etc. He has published 90 and more research papers in reputed national international Journals & conferences. He has recognized Ph. D Guide from University of Pune, S. R. T. M. University & Singhaniya University (India). He is senior IEEE member and other reputed society member. Currently working as a Associate Professor in CS Dept of SRTMU, Nanded.