Isolated Word Recognition System based on LPC and DTW Technique

Rohini B. Shinde,
College of Computer Science and Information Technology, Latur,
Maharashtra. India

V. P. Pawar, PhD.
Associate Professor in Computer Science Dept of SRTM University, Nanded, Maharashtra. India

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. In this paper, we present the steps involved in the design of a speaker-independent speech recognition system for Marathi Language. We focus mainly on 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 words spoken. In his case the experiment is conducted in MATLAB to verify these techniques.

Keywords

Feature extraction; feature matching, Linear predicted coefficient (LPC) , Dynamic Time Wrapping(DTW). Speech Recognition System (SRS), Fast Fourier transform, Marathi Language Speech Recognition System (MLSRS)

1. INTRODUCTION

Speech recognition allows you to provide input to an application with your speech. Just like clicking with mouse, typing on the keyboard, or pressing a key on the phone keypad provides an input to an application, Speech recognition system provides input by talking. In the desktop world microphones allows you to use SRS. When the user say's something, it known as utterance. An utterance may be a single word or a continuous speech. These utterances are recorded & converted to digital signal form. Digitized forms of signals are later on processed using FFT & LPC to produce speech features. Produced speech features later on goes through the DTW to select pattern matches. The LPC & DTW features techniques can be implemented using MATLAB.

The proposed technique may useful for developing the Marathi Language Speech Recognition System. Marathi is one of the recognized regional languages in India. It is an Indo-Aryan language spoken by 90 million people all over the world & mainly used in Maharashtra state in India.[1] There is a lot of scope to develop system using Indian languages of different aspects and variations; some of the works are done in the direction of Isolated words in languages like Bengali, Tamil, Telugu, Marathi, and Hindi. The method used for speech recognition system in this research, focuses 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.[2,3]

The paper is divided in to five sections, Section 1, gives Introduction, Section 2 deals with Marathi Speech Database, section 3 focuses on feature extraction techniques, section 4 focuses on DTW Pattern Matching technique ,Section 5 covers Results & conclusion followed by Section 6 gives the valuable References.

2. Marathi Speech Database.

Proposed work needs a collection of utterance for MLSRS for training & Testing. The collection of required & proper recorded speech is the database for this work. For the generation of database the continuous speech & isolated words are recorded from the different age group (from 16 to 35 age). The total number of speaker is 21. Vocabulary size of database consist 2730 tokens.

Sentences=105 samples

Isolated words=125 samples per person total (2625).

To achieve the good quality of audio-results, recording of the sentences by speakers is done in college computer laboratory, after completion of college work to avoid noisy surrounding. The speaker's were relaxed in the chair having equal distance of mice from their mouth. The sampling frequency for all recording was 11025 Hz. We collected speech data with the help of sound recording software. Sound files are recorded in the PCM format and saved with the extension .wav. For more accuracy we use the trimmed mean for sorting data. Trimmed Mean finds the average waves for the work & eliminates the unwanted waves, so 5% trimmed mean is used. Using this technique first & last signal waves are eliminated. First signal wave is eliminate because while recording speech speaker enthusiasm is high so obviously frequency we get is high & the last signal wave is eliminate because in continuous speech whenever a sentence is stop at the ending of sentences tone of speaker goes down due to which we get low frequency. The files edited are labeled properly and these files are stored in memory for further processing.

3. FEATURE EXTRACTION FOR MLSRS

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. There are many techniques used to parametrically represent a voice signal for speaker recognition tasks. These techniques include Linear Predictor Coefficients (LPC), Auditory Spectrum-Based Speech Feature (ASSF), and the Mel- Frequency Cepstrum Coefficients (MFCC).[4],[5],[6] The LPC technique is used in this paper. Following are the steps involved in the feature extraction.[7]

3.1 Speech Signal

When producing speech sounds, the air flow from a speakers lung first passes the glottis and then throat and mouth. The spectral shape of the speech signal is determined by the shape of the vocal tract (the pipe formed by your throat, tongue, teeth and lips). By changing the shape of the pipe (and in addition opening and closing the air flow through your nose) you change the spectral shape of the speech signal, thus articulating different speech sounds.

3.2 Speech Processing

Speech processing follows the steps like pre-emphasis, framing , windowing, DFT etc…

3.2.1 *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.

3.2.2 *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

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

3.2.3 *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.

3.2.4 *Windowing:*

Next process is to apply window to each individual frame so 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.

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

3.2.5 *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

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

3.2.6 *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.[3]

$$s(n) = -A(2) * X(n-1) - A(3) * X(n-2) - \qquad (5)$$
$$A(N+1) * X(n-N)$$

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.

3.2.7 *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]

$$y(k) = wk \sum_{n=1}^{n} x(n) \cos \frac{\pi(2n-1)(k-1)}{2N}$$
$$\text{k=1,……………, N.} \qquad (6)$$

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

3.2.8 *Discrete Cosine Transform:*

DCT returns the unitary discrete cosine transform of x

$$y(k) = w(k) \sum_{n=1}^{N} x(n) \cos \frac{\pi(2n-1)(k-1)}{2N} \qquad (7)$$
$$k = 1, \ldots.., N$$

Where

$$w(k) = \begin{cases} \frac{1}{\sqrt{N}} \\ \sqrt{\frac{2}{N}} \end{cases}$$

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.

3.2.9 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.

4. DYNAMIC TIME WRAPPING

A speech signal is represented by a series of feature vectors which are computed every 10ms. A whole word will comprise dozens of those vectors, and we know that the number of vectors (the duration) of a word will depend on how fast a person is speaking. In speech recognition, we have to classify not only single vectors, but sequences of vectors. Let's assume we would want to recognize a few isolated words from Marathi. For an utterance of a word w which is TX vectors long, we will get a sequence of vectors X= {x0, x1, . . . , xTX−1} from the acoustic preprocessing stage. What we need here is a way to compute a "distance" between this unknown sequence of vectors X and known sequences of vectors W = {w0,w1, . . . ,wTW} which are prototypes for the words we want to recognize.

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. The result of DTW are given in table 1

5. Result

The input voice signals of different and same speakers have been taken and compared. The results obtained are as follows in the table 1, 2, 3. Aim if this research work is to compare the performance of LPC & DTW. The speech data used in this experiment are isolated 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 we get three 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. 2.031 to 3.427. Average difference is 2.156
3. When patterns of word 1(Latur) are compared with patterns of word 2(BEED, Aurangabad) then it gives the result as shown in Table 4. Average difference between Latur & Aurangabad is 4.923 & average difference between Latur & BEED is 5.112

Table1 Result obtained by DTW for 5 speakers for word "Latur"

		P 1	P 2	P 3	P 4	P 5
		Latur	Latur	Latur	Latur	Latur
P 1	Latur	0	2.4238	2.5448	2.7688	3.4271
P 2	Latur	2.4238	0	2.4597	2.879	2.865
P 3	Latur	2.5448	2.4597	0	2.8295	2.4409
P 4	Latur	2.7688	2.879	2.8295	0	2.031
P 5	Latur	3.4271	2.8652	2.4409	2.031	0

Table 2. Result obtained by DTW for 5 speakers for word "Aurangabad"

		P 1	P 2	P 3	P 4	P 5
		Aurang abad	Aurang abad	Aurang abad	Aurang abad	Aurang abad
P 1	Aurang abad	0	3.4856	3.0014	2.4613	3.0449
P 2	Aurang Abad	3.4856	0	5.0641	3.6159	3.7744
P 3	Aurang Abad	3.0014	5.0641	0	2.8253	2.6113
P 4	Aurang Abad	2.4613	3.6159	2.8253	0	3.0088
P 5	Aurang abad	3.0449	3.7744	2.6113	3.0088	0

Table3. Result obtained by DTW for 5 speakers for word "BEED"

		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

Table4. Results obtained after comparing the two different words

	Latur	Latur	Latur	Latur	Latur
Aurangaba d	5.8079	5.3521	4.5379	4.1721	4.7483
BEED	4.9608	4.9609	5.3521	4.9611	5.3521

6. Conclusion

In this paper, discussion is done on the issues relevant to the development of unit selection speech systems for MLSRS. From the result conclusion is stated that the proposed technique is useful for Speaker Dependent & Speaker Independent Speech Recognition system for Marathi Language. Observations of Table 1, 2, 3, 4 states that the differences between the same word is less as compared to differences between the two different words. Paper is focuses on the acoustic pre-processing technique used to extract salient features of a speech signal and a Dynamic Time Warping technique used to efficiently compare the feature vectors of speech signals. Implementation and verification of these technique is done using MATLAB.

7. REFERENCES

[1] Gopalakrishna Anumanchipalli, Rahul Chitturi, “Development of Indian Language Speech Databases for Large Vocabulary Speech Recognition Systems”

[2] 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.

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

[4] “Isolated Word, Speech Recognition using Dynamic Time Warping.” Dynamic Time Warping. 14 June 2005.

[5] Jamel Price and Ali Eydgahi, “Design of Matlab®-Based Automatic Speaker Recognition Systems,” 9th International Conference on Engineering Education T4J-1, July 23 – 28, 2006.

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

[7] Jiehua Dai Zhengzhe Wei, “Study and Implementation of Feature Extraction and Comparison In Voice Recognition”

[8] 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.

[9] H. Ney. Modeling and search in continuous speech recognition. Proceedings of EUROSPEECH 1993, pages 491–498, 1993.

[10] H. Sakoe and S. Chiba, Dynamic Programming Algorithm Quantization for Spoken Word Recognition, IEEE Trans. Acoustics, Speech and Signal Proc., Vol. ASSP-26, No. 1, pp. 43-49, Feb. 1978.

[11] Palden Lama and Mounika Namburu “Speech Recognition with Dynamic Time Warping using MATLAB” CS 525, SPRING 2010 – PROJECT REPORT