Recognition of Marathi Isolated Spoken Words Using Interpolation and DTW Techniques

Ganesh B. Janvale[1], Vishal Waghmare[2], Vijay Kale[3], and Ajit Ghodke[4]

[1] Symbiosis International University, Symbiosis Centre for Information Technology,
Pune-411 057(MS), India
[2,3] Dept. of CS&IT, Dr. Babasaheb Ambedkar Marathwada University,
Aurangabad-411004, India
[4] Sinhgad Institute of Business Administration & Computer Application (SIBACA),
Lonavala, India
ganesh@scit.edu, {vishal.pri12,vijaykale1685}@gmail.com,
ajit_1974_in@yahoo.com

Abstract. This paper contains a Marathi speech database and isolated Marathi spoken words recognition system based on Mel-frequency cepstral coefficient (MFCC), optimal alignment using interpolation and dynamic time warping. Initially, Marathi speech database was designed and developed though Computerized Speech Laboratory. The database contained Marathi isolated words spoken by the 50 speakers including males and females. Mel-frequency Cepstral Coefficients were extracted and used for the recognition purpose. The 100% recognition rate for the isolated words have been achieved for both interpolation and dynamic time warping techniques.

Keywords: Speech Data base, CSL, MFCC, Speech Recognition and statistical method formatting.

1 Introduction

Currently, a lot of research is going on speech recognition and synthesis. The speech recognition understands basically what some speak to computer, asking a computer to translate speech into corresponding textual message, where as in speech synthesis, a computer generate artificial spoken dialogs. The speech is one of the natural forms of communication among the humans. The Marathi is an Indo-Aryan language, spoken in western and central India. There are 90 million of fluent speakers all over world [1][2]. 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. 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 [3]. Present work consists of the Marathi speech database and speech recognition system. The first part describes technical details of the database and second words recognition system. This paper is also split into four parts i.e. Introduction, Database, Words recognition System and Result and Conclusion.

S.C. Satapathy et al. (eds.), *ICT and Critical Infrastructure: Proceedings of the 48th Annual Convention of CSI - Volume I*, Advances in Intelligent Systems and Computing 248,
DOI: 10.1007/978-3-319-03107-1_3, © Springer International Publishing Switzerland 2014

2 Spoken Marathi Isolated Works Database

2.1 Features of Prosody in Marathi Words

Standard Marathi is based on dialects used by academic and printed media. There are 42 dialects of Marathi some of these are Ahirani, Khandeshi, Varhadi, Zadiboli, Vadvali, Samavedi and Are Marathi. The phonetics inventory of Marathi (Devnagar) along with International Phonetic Alphabets (IPA) is shown in Figure 1.a and b [4].

Devnagary	अ	आ	इ	ई	उ	ऊ	ऋ	ए	ऐ	ओ	औ	अं	अः
Transliterated	A	Āa	i	ī	u	ū	ṛ	e	Ai	o	au	aṃ	aḥ
IPA	/ə/	/a/	/i/		/u/		/ru/	/e/	/əi/	/o/	/əu/	/əⁿ/	/əh/

Fig. 1a. Vowels in Marathi Language along with Transliteration and IPA

क	ka	/kə/
ख	kha	/kʰə/
ग	ga	/gə/
घ	gha	/gʱə/
ङ	ṅa	/ŋə/

च	ca	/tsə/
छ	cha	/tsʰə/
ज	ja	/zə/
झ	jha	/zʱə/
ञ	ṅa	/ŋə/

ट	ṭa	/ʈə/
ठ	ṭha	/ʈʰə/
ड	ḍa	/ɖə/
ढ	ḍha	/ɖʱə/
ण	ṇa	/ɳə/

प	pa	/pə/
फ	pha	/fə/
ब	ba	/bə/
भ	bha	/bʱə/
म	ma	/mə/

त	ta	/t̪ə/
थ	tha	/t̪ʰə/
द	da	/d̪ə/
ध	dha	/d̪ʱə/
न	na	/n̪ə/

य	ya	/jə/
र	ra	/rə/
ऱ	r̲a	/ɽə/
ल	la	/lə/
व	va	/wə/
श	śa	/ʃə/

ष	ṣa	/ʂə/
स	sa	/sə/
ह	ha	/hə/
ळ	ḷa	/ɭə/
क्ष	kṣa	/kʃə/
ज्ञ	jña	/ɟʝɲə /

Fig. 1b. Consonants in Marathi Language along with Transliteration and IPA

Marathi words are formed with the combination of vowels and consonants. e.g. 'Aai' means 'Mother'. This is pronounced as '/a i /' according to International phonetics alphabet (IPA) and composed of two vowels. The collection of utterances in digital format is called computerized speech database and is required for recognition purpose. The isolated spoken words [5] were collected in 35 different sessions from 50 speakers. Initially all the speakers were asked to pronounce of Marathi words, the speaking skill has been examined by examiner; speakers have been selected on the bases of test. After selection, they have been trained how to speak the given words.

2.2 Acquisition Setup

The samples were recorded in 15 X 15 X 15 feet room with sampling frequency 11 KHz in normal temperature and humidity. The microphone was kept 5 – 7 cm from speakers. Whole speech database was collected with the help of Computerized Speech Laboratory (CSL). It is an input/output recording device for a PC, which has special features for reliable acoustic measurements. CSL offers input signal-to-noise performance typically 20-30dB. Analog Inputs with 4 channels two XLR and two phono-type, 5mV-10.5V peak-to-peak, channels 3 and 4, switchable AC or DC coupling, calibrated input, adjustable gain range >38dB, 24-bit A/D, Sampling rates::<-90dB F.S., 8000-200,000Hz, THD + NFrequency Response (AC coupled): 20 to 22kHz +.05dB at 44.1kHz [6][7].

2.3 Isolated Marathi Words Corpus

There are 12 vowels and 36 consonants in Marathi alphabets. Spoken words from the selected speakers were recorded and stored in different files and folders with respective to speakers' initial names in the 'wav' format. The database was included mostly the words starting from each vowels, some of the words with phonetics.

3 Words Recognition System

The recognition system was developed using Mel – Frequency Cepstral Coefficient (MFCC) [8]. The following are the step to find the MFCC features.

3.1 Speech Signal

The waveform of spoken word 'Ati' along with pronounces of vowels is shown in figure 2. The words of exciting signal and impulse response of vocal tract is called speech signal as shown in Equation (1).

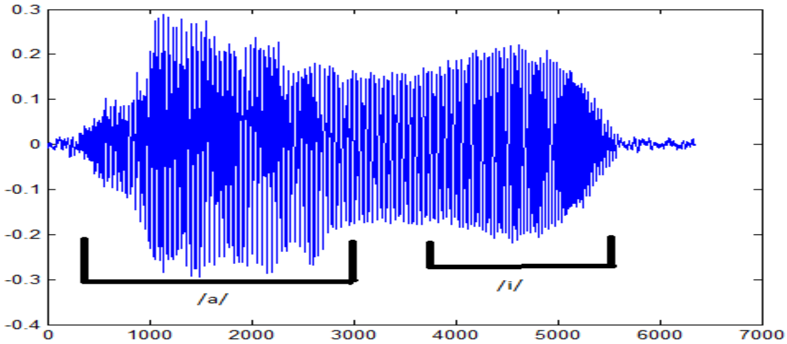

Fig. 2. Waveform of Spoken word 'Ati'

$$S[n] = e[n] * \theta[n] \tag{1}$$

Where, S[n], e[n] and θ[n] are speech signal, exciting signals and impulse response of vocal tract respectively.

3.2 Pre-emphasizing

Speech is emphasized with filter 1-az-1 where, “a” is between 0.9 and 1. In the time domain, the relationship between the output S’n and input Sn of the pre-emphasis by the default value of a ia 0.97 as shown in equation (2).

$$S_n^{'} = S_n * aS_{n-1} \tag{2}$$

3.3 Framing and Windowing

The signal is remained stationary at 20ms. Calculation of number of frame is done by multiplying the signal, consisting of N samples, with a rectangular window function of a finite length as shown in Equations (3), (4) and (5).

$$S_{frames}[n] = S[n].W[n] \tag{3}$$

$$W[n] = \begin{cases} 1 \\ 2 \end{cases} \tag{4}$$

1 - N ·r ≤ n < N ·(r+1), r = 0,1,2,3, ..., M−1
0- otherwise

$$N = fs\,.tframe \tag{5}$$

Where ‘M’ is the number of frames, ‘fs’ is the sampling frequency and tframe length of frame measured in time. The frame is shifted 10 ms so that the overlapping between two adjacent frames is to avoid the risk of losing the information from the speech signal. Each frame that contains nearly stationary signal blocks the windowing function is applied. There are a number of different windows functions to minimize the discontinuities. The Hamming window has a very good relative side lobe amplitude and good frequency resolution compared to the rectangular window. The window function is described by Equation (6).

$$x[n] = 0.54 - 0.46 \cdot \left(\frac{2\pi \cdot n}{N-1} \right) \tag{6}$$

3.4 Fourier Transform

A 512 point Fast Fourier Transform was applied and calculated using equation 7, for good frequency resolution. The resolution of the frequency spectrum is primarily decided by the main lobe, while the degree of leakage depends on the amplitude of the side lobes.

$$X(k) = \sum_{n=1}^{n} x_n * e^{(-j*2*\pi*(k-1)*(n-1)/N)} \tag{7}$$

Where $1 <= n <= N$

3.5 Mel-Frequency Filter Bank

A 24 triangular shaped band-pass filter banks are created by calculating a number of peaks uniform spaced in the mel – scale and then transforming them back to the normal frequency scale. The transmission from linear frequency to mel-frequency is shown in equation (8).

$$mel = 2595 \cdot \log(1 + \frac{frequency}{700}) \tag{8}$$

The mel frequency spectrum furthermore reduces the amount of data without losing vital information in a speech signal. The resolution as a function of the frequency is logarithmic.

3.6 Discrete Cosine Transform

The DCT is widely used in the area of speech processing and is often used when working with cepstrum coefficients. In a frame, there are 24 mel cepstral coefficients obtained, out of 24 only 13 coefficient has been selected for the recognition system as shown in figure 3.

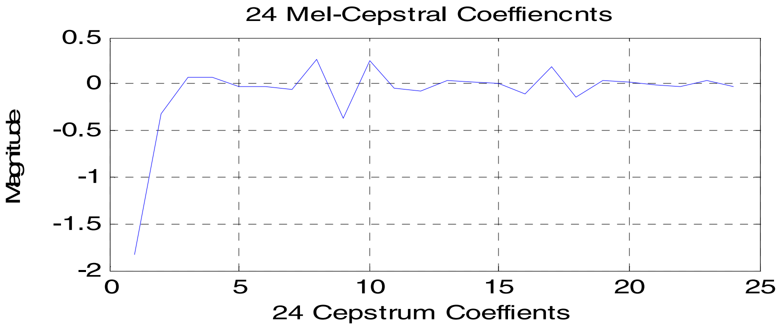

Fig. 3. 24 Mel-Cepstral Coefficients

For the word recognition system, we have selected 13 Mel-Cepstral Coefficients frame wise. Figure 4 shows the mean of the 13 Mel Cepstral coefficients of all the frames of spoken for the word.

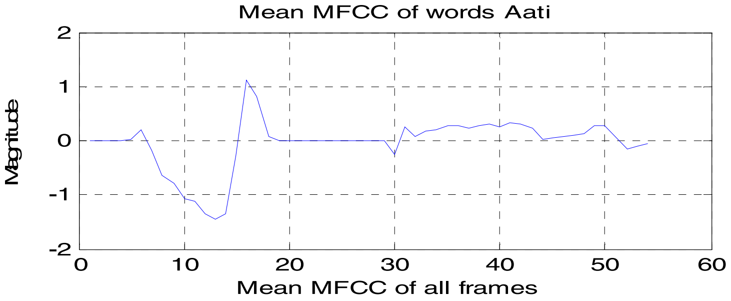

Fig. 4. Mean of MFCC of Marathi word i.e. Aati

3.7 Distance Measures

There are some commonly used distance measures i.e. Euclidean Distance, City Block, Weighted Euclidean Distance and Mahalanobis distance. A Euclidean Distance measure is the "standard" distance measures between the two vectors in feature space. Euclidean distance of two vectors x and p is measured using the equation (9).

$$d^{2}\ Euclid\ (\vec{x}\cdot\vec{p}) = \sum_{n=1}^{N}\left(x_i - p_i\right)^2 \tag{9}$$

3.7.1 Optimal Alignment between Two Time Series by Simple Interpolation

As the Euclidean distance measure the standard distance of two same in length vectors, but speech signals are not same in size. To make the vector same in size, for this purpose position (i' prime) of prime value is calculated by equation (10).Simple interpolation has been calculated by using the equation (11) for example two speech signals having different frame numbers i.e. m and n as shown in figure 5.

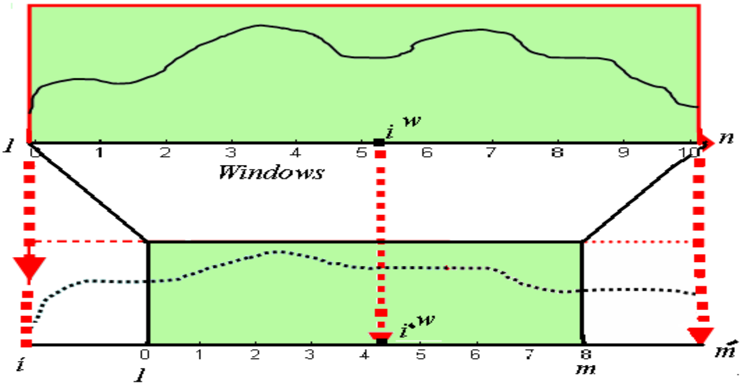

Fig. 5. Speech signals having different frames

$$I' = \left(\frac{m-1}{n-1}\right) i + \left(\frac{n-m}{n-1}\right) \tag{10}$$

Wk' is values of a frame at a position k' which can be defined as k' is not integer. So that integer part is separated from the fraction.

$$I'' = \text{int}\,(k'), f = k' - \text{int}\,(k')$$

Where 'int' is integer, so that the interpolated values Wk' at the position k' is calculated as shown in equation (11)

$$Wk' = Wk''(1-f) + Wk'' \cdot f \tag{11}$$

Now, both the number of frames is same, is applied the Euclidean distance on the different signals.

3.7.2 Optimal Alignment between Two Time Series by DTW Techniques

DTW algorithm is used to determine time series [9] [10]. The minimum distance warp path is calculated using the equation (12).

$$Dist\,(w) = \sum_{k=1}^{k=K} Dist\,(W_{ki} - W_{kj}) \tag{12}$$

Dist (w) is typically Euclidean distance of warp path 'w', and Dist (Wki, Wkj) is the distance between the two data point indexes in the 'kth' of the warp path.

Table 1. Euclidean distance matrix Marathi words calculated by Interpolation Method

Average distant of spoken words by all 50 subjects						
	Aabhar	Aabhas	Aadhar	Aai	Anand	Apali
Aabhar	**0.04**	0.13	0.14	0.16	0.07	0.14
Aabhas	0.13	**0.02**	0.15	0.11	0.14	0.10
Aadhar	0.20	0.23	**0.02**	0.20	0.23	0.13
Aai	0.19	0.15	0.18	**0.03**	0.13	0.21
Anand	0.10	0.21	0.23	0.15	**0.03**	0.25
Apali	0.17	0.13	0.11	0.20	0.21	**0.02**

Table 2. Euclidean Distance Matrix Marathi Words calculated by DWT method

Average distant of spoken words by all 50 subjects						
	Aabhar	Aabhas	Aadhar	Aai	Anand	Apali
Aabhar	83	283	391	328	475	341
Aabhas	283	24	265	409	212	137
Aadhar	391	265	27	379	316	347
Aai	328	409	379	34	471	555
Anand	475	212	316	471	47	272
Apali	341	137	347	555	272	59

4 Conclusion

We have computed the distance matrixes for 750 corpuses, including males and females. During the experiment, we experienced the effectiveness of MFCC in feature extraction. For this experiment, we have used a limited number of samples. Increasing the number of samples may give the complete recognition. To compare two vectors of different in size, two methods have been used i. e. interpolation and DTW. Table 1 and 2 shows the average values of spoken words from 50 speakers. It is also found the values of diagonal elements are smaller than rest of the elements shown in all the tables. It can be seen from the tables that 100 % recognition is achieved for all subjects in both comparative approaches. The recognition by DTW is more correct than simple interpolation methods. Marathi Speech recognition system based on MFCC features seems to be successful.

References

1. Ghadge, S.A., Janvale, G.B., Deshmukh, R.R.: Speech Feature Extraction using Mel – Frequency Cepstral Coefficient (MFCC). In: International Conference on Emerging Trends in Computer Science Communication and Information, January 9-11 (2010)
2. Rbiner, L., Juary, B.-H.: Fundamental of speech Recognition, 1st edn. Pearson Education, Inc., Copyright 1993 AT & T (1993) ISBN 9780130151575
3. Samudravijaya, K., Rao, P.V.S., Agrawal, S.S.: Hindi Speech database. In: Proc. Int. Conf. Spoken Language Processing, ICSLP 2000, Beijing (October 2000)
4. http://en.wikipedia.org/wiki/Marathi_language
5. Lipeika, A., Lipeikience, J., Telkonys, L.: Development of Isolated word speech Recognition system. Information 13(1), 37–46 (2002)
6. KayPENTAX.: Multi-speech and Software. Software instrumentation manual, issue G, A division of PENTAX – medical company, Bridgewater line, Lincoln Park, NJ 07035 -1488 USA (February 2007)
7. Janvale, G.B., Gawali, B.W., Deshmukh, S.N., Deshmukh, R.R., Mehrotra, S.C.: Speech Analysis through CSL. on 8th -10th at Inter – University Research Festival. Sant Gadge Baba Amravati University, Amravati (MS) India

8. Sato, N., Obuchi, Y.: Emotion Recognition using Mel – Frequency Cepstral Coefficients. Journal of Natural Language processing 14(4), 83–96 (2007)

9. Keogh, E., Pazzani, M.: Derivative Dynamic Time Warping. In: International Processing of First Intl SIAM International Conference on Data Mining, Chicago, Illinoisi (2001)

10. Kruskall, J., Liberman, M.: The Symmetric Time Warping Problem: From continuous to Discrete. In: Time Warps String Edits and Macromolecules. The Theory and Practice of Sequence Comparison, pp. 125–161. Addison – Wesley Publication Co., Reading Massuchusetts (1983)