This full-text paper was peer-reviewed and accepted to be presented at the IEEE ICCSP 2015 conference

Keyword Spotting System for Tamil Isolated Words using Multidimensional MFCC and DTW Algorithm

Senthildevi K. A, Chandra E

***Abstract*—Audio mining is a speaker independent speech processing technique and is related to data mining. Keyword spotting plays an important role in audio mining. Keyword spotting is retrieval of all instances of a given keyword in spoken utterances. It is well suited to data mining tasks that process large amount of speech such as telephone routing and to audio document indexing. Feature extraction is the first step for all speech processing tasks. This Paper presents an approach for keyword spotting in isolated Tamil utterances using Multidimensional Mel Frequency Cepstral Coefficient feature vectors and DTW algorithm. The accuracy of keyword spotting is measured with 12D, 26D and 39D MFCC feature vectors for month names in Tamil language and the performances of the mu1tidimensional MFCCs are compared. The code is developed in the MATLAB environment and performs the identification satisfactorily.**

***Index Terms*— Audio mining, Keyword spotting, Speech processing, DTW algorithm, MFCC Feature vectors.**

I. Introduction

AUDIO mining is a speaker-independent, speech recognition technique used to search audio signals for occurrences of spoken words or phrases[1]. Audio Mining has gained interest due to availability of voluminous audio content. For example newscasts, sporting events, telephone conversations, recordings of meetings, Web casts, documentary archives etc. Speech technology is used to recognize the words are spoken in an audio file but audio mining searches can be carried out to locate specific words and phrases within the audio. Keyword spotting (KWS) is a technologically relevant problem, playing an important role in audio indexing and speech data mining applications [2]. KWS is also used for locating occurrences of keyword in speech signal [3]. The proposed problem is similar to speech recognition, although ignoring the additional signal information around the words of interest [4].

Keyword spotting approaches are based on Dynamic Time Warping (DTW), Hidden Markov Models (HMM), Neural Network and other techniques. In recent years, mostly DTW based KWS techniques are developeddue to its speed and efficiency in detecting similar patterns. In the existing papers [5], [6], [7], [8], [9], variant DTW algorithms are used for keyword spotting with different feature vectors.

Mel Frequency Cepsturm Coefficient features are suitable form for speech processing. Isolated word recognition system are developed in various languages like Marathi [10], Malayalayam [11], Bangla[12] and in Arabic[13] using MFCC features and DTW algorithm.

In this proposed work, keyword spotting system is developed for Tamil isolated utterances using multiple Mel-frequency cepstral coefficients (MFCC) and DTW pattern matching algorithm. Tamil month names are recorded in noisy environment by a female speaker. Each month name in Tamil has been repeated 25 times so that totally 300 words are recorded using AUDACITY speech software. A preprocessing is made for not only noise reduction, but also normalization. Then 12D, 26D and 39D MFCCs are calculated for all utterances and are stored as templates. Finally, DTW is used as a pattern matching algorithm due to its speed and efficiency in detecting similar patterns. Experiments have been conducted with multiple MFCCs to achieve best word identifier. The system is designed for Tamil language but all its modules except the spoken word references developed are language independent.

The organization of this paper is as follows. In Section II, the Audio mining method is explained. In section III, various processes in the Keyword spotting system are discussed. In section IV, the developed Keyword spotting system for Tamil isolated words is presented. In Section V, the experimental results with three different MFCCs are given.

II. Audio Mining

Audio mining also called audio searching technique is used to search audio files for occurrences of spoken words or phrases [14]. The major advantage of audio mining is the ability to process and search audio data thousands of times faster for large files than a human could - because the human would have to listen to each word of the audio [15].

Senthildevi K.A is with the Department of Computer Science, Gobi Arts & Science College, Gobichettipalayam (e-mail: senthildevigasc@gmail.com).

Chandra E is with the Department of Computer Applications, Dr. SNS Rajalakshmi College of Arts and Science, Coimbatore 641 049 (e-mail: crcspeech@gmail.com).

978-1-4799-8081-9/15/$31.00 © 2015 IEEE

There are two main phases in audio mining system: training and template matching [16]. During the training phase, a training vector is generated from the speech signal of each word spoken by the user. The training vectors extract the spectral features for distinguishing different classes of words. Each training vector can serve as a template for a single word or a word class. These feature vectors are stored in a database for subsequent use in the template matching phase. During the keyword matching phase, the user speaks any word that is to be identified in the trained templates. Feature vector is generated for that word and compared with the trained templates using pattern matching algorithm. The audio mining system phases block diagrams are shown in Fig. 1 and Fig. 2.

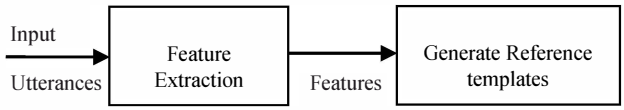

Fig. 1. Feature Extraction Phase

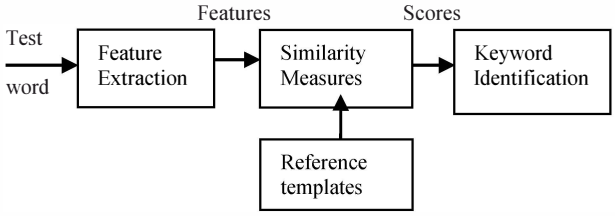

Fig. 2. Feature Matching Phase

In phase I, most commonly used feature extraction techniques linear predictive coding (LPC) and Mel frequency cepstral coefficients (MFCC) are used to extract speech features. LPC is a time domain technique and suffers from variations in the amplitude of the speech signal due to noise. The preferred technique for feature extraction is MFCC wherein the features are generated by transforming the signal into frequency domain. In general, cepstral features are more compact, discriminable, and most importantly, nearly decorrelated and therefore, they can provide higher baseline performance over filter bank features.

In phase II, several techniques including analysis methods based on Bayesian discrimination, Hidden Markov Models (HMM), Dynamic Time Warping (DTW) based on dynamic programming, Support Vector Machines, Vector Quantization and Neural Networks are used.

III. KEYWORD SPOTTING

Keyword spotting has become an important branch of audio mining. Keyword spotting is well suited to data mining tasks, process large amount of speech such as real time monitoring and to audio document indexing [5]. The task of locating the occurrences of a given keywords K in a speech utterance is termed as keyword spotting (KWS) [6]. It plays an important role in audio indexing and speech data mining applications. The method is applied mainly for a large amount of spoken documents must be searched to learn whether they contain some specific words [1]. The fast detection of the words and information about the exact location eliminate a lot of human work in such tasks like audio data mining [6]. Hence, Keyword spotting is well suited to data mining tasks for process large amount of speech because keyword spotting requires significantly less processing power than transcription, and can therefore run at considerably faster speeds.

The keyword spotting process carried out in the proposed work is divided into three stages as follows: firstly, the input sample utterances are preprocessed to reduce noise effects. Secondly, Multidimensional (13D, 26D, 39D) Mel-frequency cepstral coefficients (MFCC) features are extracted from the speech signal. Lastly, each keyword is compared to the template references of input utterances with multidimensional MFCC features using DTW algorithm.

A. Feature Extraction TEchnique

The speech signal is represented by a sequence of feature vectors. The selection of appropriate features along with methods to estimate (extract or measure) the vectors are known as feature selection and feature extraction. Various feature extraction techniques exist like Principal Component Analysis (PCA), Linear Discriminate Analysis (LDA), Independent Component Analysis (ICA), Linear Predictive Coding (LPC), Mel-Frequency Cepstral Coefficients (MFCC) etc. In this paper, we have calculated 13D, 26D, 39D MFCCs and the performances of keyword spotting system are compared with these various MFCCs.

Features extraction is usually performed in three main stages. The first stage is called the speech analysis or the acoustic front-end, which performs spectra-temporal analysis of the speech signal and generates raw features describing the envelope of the power spectrum of short speech intervals. The second stage compiles an extended feature vector composed of static and dynamic features. Finally, the last stage transforms these extended feature vectors into more compact and robust vectors that are then supplied for processing.

B. MFCC Feature Extraction Process

Mel Frequency Cepstral Coefficients (MFCCs) are a feature widely used in automatic speech recognition. Overall process of the MFCC calculation is shown in Fig. 3. The MFCC calculation consists of seven computational steps [17].

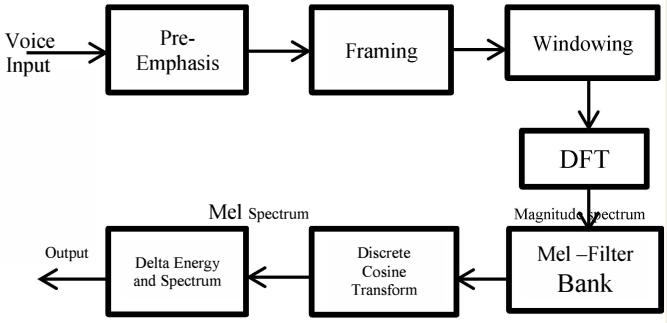

Fig . 3. MFCC Block Diagram

most important algorithms in data mining and speech recognition techniques. DTW algorithm is for measuring similarity between two time series which may vary in time or speed [18]. Unlike Linear Time Warping (LTW) compares two time series based on linear mapping of the two temporal dimensions, Dynamic Time Warping (DTW) allows a nonlinear warping alignment of one signal to another by minimizing the distance between the two as shown in Fig.4 [19].

Dynamic time warping is an efficient algorithm optimizes to find the nonlinear alignment path between two sequences their distance is obtained by time scaling one of the signals nonlinearly so that it aligns with the other. For keyword spotting, a prototype of the keyword is stored as template and compared to each of word in the incoming speech utterance.

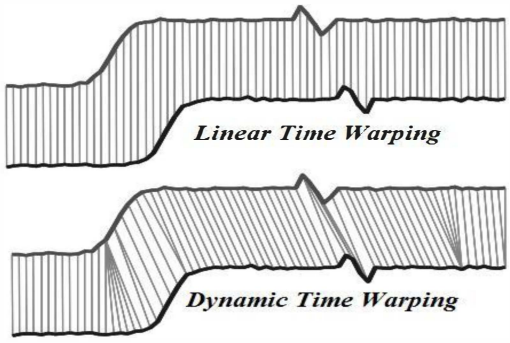

Fig. 4. DTW alignment of Two Time Series

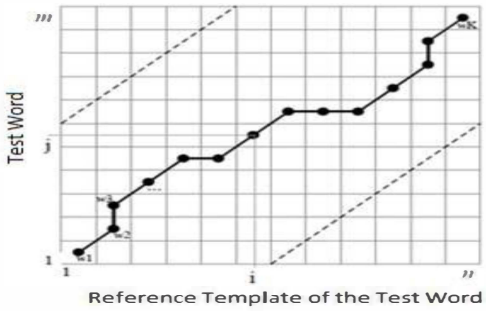

Fig. 5. Warping between tow time series

In the Fig. 5, the feature vector for the reference keyword and input keyword are arranged along the two sides of the grid. In this case, the reference template of length n is arranged along the vertical axis. Each block in the grid is the distance between corresponding feature vectors. The best match between these two sequences can be computed from the path through the grid which minimizes the total cumulative distance between them as shown by the dark in the figure [18].

IV. Keyword Spotting System for Tamil Month Names

This Paper presents an approach for keyword spotting in Tamil Isolated month name utterances using multidimensional mel frequency cepstral coefficient feature vectors and DTW algorithm. Feature of test word is compared with trained reference templates and distance similarities are computed with 12, 26 and 39 MFCC features coefficients using Dynamic Time Warping (DTW).

A. Speech Corpus

Test templates were recorded with a female speaker of 35-40yrs age. The speaker was asked to speak the 12 month names of Tamil language and each word is uttered 25 times. Totally 300 utterances of the words were recorded. The isolated words were recorded using built in microphone of laptop using the AUDACITY speech software. The data were recorded in a closed room where background noise was present. The recorded words are shown in the Table I. The waveform of utterance 'CHITHIRAI' is shown in Fig. 6.

TABLE I
Speech Training Dataset

SERIALNO	SPOKEN WORD
1.	CHITHIRAI
2.	VAIKASI
3.	AANI
4.	AADI
5.	AAVANI
6.	PURATTASI
7.	AIPPASI
8.	KARTHIKAI
9.	MARKAZHI
10.	THAI
11.	MAASI
12.	PANGUNI

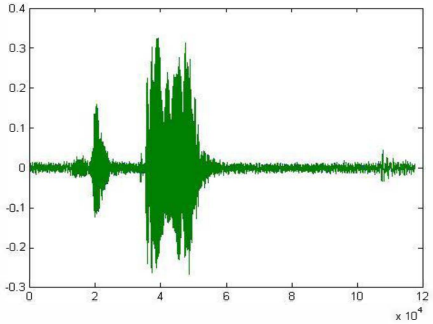

Fig 6. Waveform of word "CHITHRAI"

B. Multidimensional MFCC Calculation

For all the recorded Tamil utterances, 12D, 26D and 39D MFCC vectors are calculated and stored as templates with the following process. Frequency cepstral coefficients (MFCCs) computed from speech signal are given in vector S and sampled. The speech signal is first pre-emphasised using a first order FIR filter with pre-emphasis coefficient. The pre-emphasised speech signal is subjected to the short-time Fourier transform analysis with frame durations of TW (25ms), frame shifts of TS (10ms) and analysis window function given as a function handle in WINDOW. This is

followed by magnitude spectrum computation followed by filterbank design with triangular filters uniformly spaced on the mel scale between lower and upper frequency limits given in R [300 3700] Hz. The filterbank is applied to the magnitude spectrum values to produce filterbank energies (FBEs) (M per frame). Log-compressed FBEs are then decorrelated using the discrete cosine transform to produce cepstral coefficients. Final step applies sinusoidal lifter to produce liftered MFCCs.

C. Similarity Distance Measure

The keyword spotting is implemented in MATLAB using DTW where the distance calculation is done between the tested speech and the reference templates. After running the pattern matching algorithm with three different MFCCS, the similarity distances are obtained.

D. Keyword Detection

Once the similarity distances are computed for the entire utterances, the detection decision can be made based on a threshold value. The distortion score of templates which satisfies the threshold value are counted as occurrences of the given keyword. The accuracy rates of keyword detection for every month with the three different MFCCs are calculated and compared.

V. Experimental Results

TABLE II

KEYWORD SPOTTING RESULTS WITH MULTIDIMENSIONAL MFCC

S.No.	Input Utterances	Accuracy in %		
		With 12D MFCC	With 26D MFCC	With 36D MFCC
1.	CHITHIRAI	80	84	92
2.	VAIKASI	80	84	88
3.	AANI	72	76	92
4.	AADI	80	88	92
5.	AAVANI	76	80	88
6.	BHURATASI	72	76	88
7.	IYPPASI	84	88	92
8.	KARTHIKAI	68	80	84
9.	MARKAZHI	64	76	88
10.	THAI	76	84	92
11.	MASI	80	88	92
12.	PANGUNI	64	80	88
	AVERAGE ACCURACY	74.6	82	89.7

The system requires the user to record month names in the Tamil language. After that the system saves the recorded voice as reference templates in a directory. Then, the user is required to record any single month name to test. Features vectors for reference utterances and test utterance are calculated in the above method.

The keyword spotting is implemented using DTW where the distance calculation is done between the tested speech and the reference templates. After running the pattern matching algorithm with three different MFCCS, the results are obtained. The accuracy rates of every month with the three different MFCCs are given in the Table II. Average accuracy of all months for 12D, 26D and 39D are found.

A. Comparison of Different Implementations of MFCCs

The accuracies of MFCCs with all months and average accuracies of them are compared in the charts as shown in the Fig. 7 and Fig. 8. The spectrogram of 12D, 26D, 39D MFCCs are shown in the Fig. 9, 10, and 11.

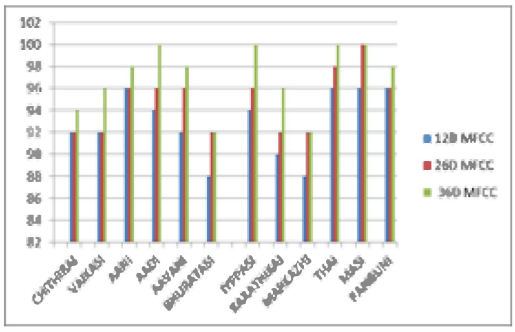

Fig. 7. Comparison of KWS performance for Tamil months with Multidimensional MFCC

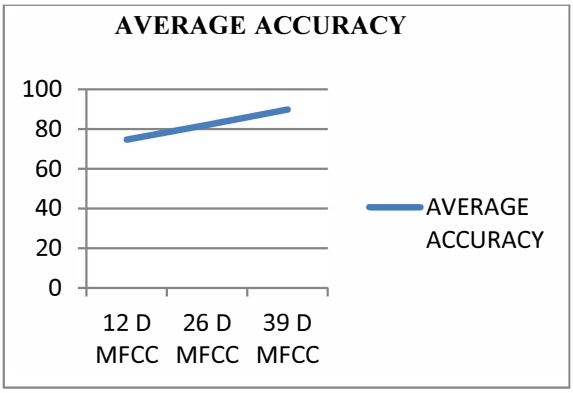

Fig 8. Comparison of Accuracy of 12D,26D, 39d MFCC in KWS

VI. Conclusion

In this paper, Keyword spotting in isolated utterances for Tamil language are computed with three different dimensional MFCCs and DTW algorithm. The number of occurrences of given keywords are identified and maximum of 89.7% average accuracy of keyword spotting is obtained with 39 dimensional MFCC. The 12, 26 and 39 dimensional MFCC are used for keyword spotting and it is found that 39 dimensional MFCC are more efficient and has greater accuracy than 13 and 26 dimensional MFCCs.

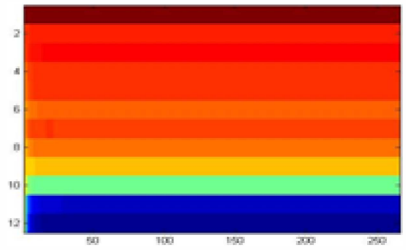

Fig. 9. Image of 12D MFCC of the word "CHITHIRAI"

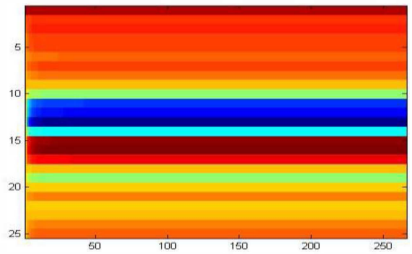

Fig. 10. Image of 26D MFCC of the word "CHITHRAI"

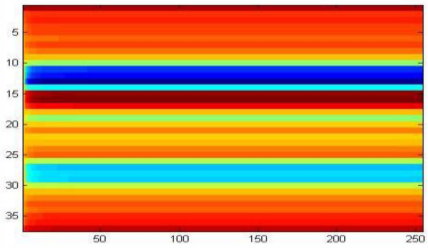

Fig. 11. Image of 39D MFCC of the word "CHITHIRAI"

REFERENCES

[1] S.Shetty, and K.K. Achary, "Audio Data Mining Using Multi-perceptron Artificial Neural Network," *International Journal of Computer Science and Network Security*, vol.8, pp.224-229, Oct. 2008.

[2] Jansen, A., Niyogi, P.: Point process models for spotting keywords in continuous speech. Audio, Speech, and Language Processing, IEEE Transactions on 17/8/2009, 1457-1470 IEEE

[3] Shao, J., Zhao, Q., Zhang, P., Liu, Z., Yan, Y.: A fast fuzzy keyword spotting algorithm based on syllable confusion network. In: Eighth Annual Conference of the International Speech Communication Association. (2007) 2405-2408, The IEEE website. [Online]. Available: http://www.ieee.org

[4] Ramachandran, R.P., Mammone, R.J.: Modern methods of speech processing. Volume 327. Springer (1995)

[5] Halima Bahi, Naida Benati, A New Keyword Spotting Approach in IEEE transaction on speech and audio processing,2009.

[6] Z. Yaodong and J. R. Glass, "Unsupervised spoken keyword spotting via segmental DTW on the Gaussian posteriorgrams," in Automatic Speech Recognition and Understanding, 2009. ASRU 2009. IEEE Workshop on, 2009, pp. 398-403

[7] M. S. Barakat, C. H. Ritz, D. A. Striling "Keyword Spotting Based on Analysis of Template Matching Distances", 978-1-4577-11800/11/$26.00 2011 IEEE..

[8] J. Keshet, D. Grangier, and S. Bengio, "Discriminative keyword spotting," Speech Communication, vol. number 51, pp. 317-329, 2009

[9] John Sahaya Rani Alex, Nithya Venkatesan, "Spoken Utterance Detection Using Dynamic Time Warping Method Along With a Hashing Technique", InternationalJournal of Engineering and Technology (IJET), Vol 6 No 2 Apr-May 2014

[10] Siddheshwar S. Gangonda, Dr. Prachi Mukherji, "Speech Processing for Marathi Numeral Recognitionusing MFCC and DTW Features ", International Journal of Engineering research and Applications, ISSN 2248-9622,2012.

[11] Sreejith c, Reghurai PC, "Isolated Spoken Word Identification in Malayalam using Mel-frequency Cepstral Coefficients and K-means clustering", International Journal of Science and Research(IJSR),ISSN:2319-7064, Vol.1 Issue 3,2012.

[12] Md. Akkas Ali, Manwar Hossain and Mohammad Nuruzzaman Bhuiyan, "Automatic Speech Recognition Technique for Bangla Words", International Journal of Advanced Science and Technology, Vol. 50, 2013.

[13] MarutiLimkar, RamaRao, VidyaSagvekar, "Isolated Digit Recognition Using MFCC AND DTW", International journal on Advanced Electrical and Electronics Engineering(IJAEE), ISSN(Print):2278-8948, Vol.1, Issue-1, 2012.

[14] Manpreet Kaur Mand, Diana Nagpal, Gunjan, "An Analytical Approach for Mining Audio Signals", *International Journal of Advanced Research in Computer and Communication Engineering* Vol. 2, Issue 9, September 2013,ISSN 2319-5940.

[15] Atkinson Baker , " Audio Mining ", Court Reporting.htm

[16] *Vibha Tiwari, "* MFCC and its applications in speaker recognition**",** *International Journal on Emerging Technologies***1**(1): 19-22(2010) , ISSN : 0975-8364.

[17] Ravi Kumar K.S., Ganesan S, "Comparison of Multidimensional MFCC Feature Vectors for Objective Assessment of Stuttered Disfluencies", Int. Journal Advanced Networking and Applications, Volume: 02, Issue: 05, Pages: 854-860 (2011).

[18] A.J. Kishon Thambiratnam, "Acoustic Keyword spotting in speech with application to data mining", PhD Thesis

[19] E.Keogh and M.Pazzani. "Scaling up Dynamic Time Warping for Data Mining Applications in *KDD"*, pages 285-289, 2000.