Received: June 9, 2017

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Modified Multivariate Euclidean Dynamic Time Warping Based Spoken Keyword Detection

John Sahaya Rani Alex[1] **Nithya Venkatesan**[2*]

[1]*School of Electronics Engineering, VIT University, Chennai Campus, India*
[2]** School of Electrical Engineering, VIT University, Chennai Campus, India .*
* Corresponding author's Email: nithya.v@vit.ac.in

Abstract: Traditional Dynamic Time Warping (DTW) technique find similarities between two one-dimensional time series sequence. Initially, in earlier decades, DTW was not preferred because of its computational complexity. However, due to the evolution of computing power, this has been revisited for spoken keyword detection recently. Conventional spectral features such as Mel-Frequency Cepstral Coefficients (MFCC) and contemporary wavelet features are multi-dimensional in nature which are used in speech recognition. In this work, a new strategy of DTW is proposed to work with multi-dimensional feature vector in calculating the local distance matrix. Additionally, a faster approach is specified to find the similarity in the global distance matrix. The proposed methods are evaluated with MFCC and wavelet features on a connected TIDIGITS corpus for spoken keyword detection system. Experimental results prove that there is an improvement in reduction of computational complexity compared to traditional DTW. Also, contemporary wavelet feature based spoken keyword detection system gave better detection accuracy than MFCC based spoken keyword detection system in the noisy environment.

Keywords: DTW, Euclidean, Multi-variate, Spoken keyword, MFCC, Wavelet.

1. Introduction

Spoken keyword spotting or spoken term detection in the spoken utterance is finding the occurrence of a spoken word from an audio data. It is a subclass of Speech Recognition (SR). Keyword spotting is used as an information retrieval from audio data such as broadcast news, audio lectures, call monitoring by law enforcement, call center conversations and so on. In the real world, for example in call center application, transcribing the whole customer response is not needed, it is enough to look for particularly sensitive information from the response and drive the system based on that. So designing a spoken keyword spotting system relies upon the application of interest.

There are a lot of approaches for designing the keyword spotting system. From the literature, it is noted that the keyword spotting research has been started in 1973 by Bridle using dynamic programming algorithms[1]. Speech features are extracted from the keyword which then treated as a template, which is searched through spoken utterance by using nonlinear Dynamic Time Warping (DTW) algorithm. The complexity of the algorithm is $O(N^2)$ where N is being the maximum length of the feature vectors in the keyword and unknown spoken utterance. Because of the impediment, instead of DTW, Hidden Markov Model(HMM), based keyword spotting is implemented[2]. In HMM-based keyword spotting, three acoustic models have been created, one for a keyword, another for non-keyword or out of vocabulary (OOV) and the third for background are modeled from the training data. Even though this system is popular, the drawback of the system is that it needs a lot of annotated data and if there is the new keyword to be added to the system, the whole system has to be retrained again[3]. More study has been made in HMM based model in terms of defining the filler or non-keyword model. An acoustic model of OOV/filler/non-keyword is created with the concatenation of syllabic models[4-5].

Since keyword spotting is a specific application of SR, the techniques which are used for SR is adopted with a variant in keyword spotting system. Discriminative based keyword spotting is carried out using Support Vector Machines(SVM), Associative Neural Networks(ANN) [6-8]. The challenges of discriminative based KWS is that it requires a lot of annotated data for training which is again a time-consuming process.

With recent advances in computing power, DTW based KWS is revisited [3,8] recently. Further Fast DTW [10-12] method implements DTW in O(N) computations instead of $O(N^2)$ computations. To speed up DTW, Sakoe-Chiba band and Itakura parallelogram constraints are analyzed along with a variant of DTW known as Segmental DTW(SDTW) [11]. Typically, DTW based keyword spotting system use Mel-Frequency Cepstral Coefficient (MFCC) as feature extraction method. Instead of the conventional MFCC, Gaussian posteriorgram vectors are used in the SDTW [12] algorithm for checking the similarity. Short-Time Fourier Transform (STFT) is used in MFCC which deteriorates the system performance if the environment is noisy. Wavelet transform captures time-frequency information [13] for analyzing transient signals such as speech signal. Also, Wavelet transform is used for speech enhancement [14]. The implementation of wavelet transform is done using successive digital filters [13] which can be imitated to implement a Mel scale like filter bank for speech recognition [15]. The conventional filter banks of speech recognition such as Mel scale and Bark scale are implemented using wavelet transform instead of Fourier transform [16-18]. In this work, Wavelet transform is chosen to design a feature extraction method for noisy environment because it captures time- frequency information of transient signal and gives multi resolution of time-frequency information. Traditional DTW deals with univariate time series(UTS) [3] which may not capture the similarities of all the dimension of the feature vector and the similarity check of two univariate time series sequence would not reflect correctly [19]. Moreover feature vectors are multi-dimensional, so there is a necessity for Multivariate Time Series (MTS) based DTW. Henceforth this research aims at developing MTS based DTW with less computational complexity. In addition, most of the real-time applications are handled in a noisy environment; wavelet transform based feature extraction technique is experimented with modified Euclidean DTW technique. The detailed discussion on wavelet cepstral coefficient and the basic concepts of dynamic time warping methods are presented in Section 2. Table 1 illustrates some notations used in this work. It is followed by the proposed changes in the modified DTW algorithm in Section 3. An insight into the experimental setup and offline simulated results are presented in Section 4. Elaborate results and discussion are presented in Section 5 and conclusion are given in Section 6.

2. Methodology

Pattern recognition system has a data acquisition and a data representation as a front end system. For this research work, a wavelet based feature extraction method considered as a feature extraction method.

2.1 Wavelet Cepstral Coefficient Feature Extraction Technique

The speech in the TIDIGITS database is sampled at 8 kHz, which leads to the permissible bandwidth of the speech signal limited to 4 kHz. The speech signal is framed at every 25 ms with an overlap of 15 ms in which initially the framed speech signal is hamming windowed. Auditory scale like filter bank is designed using wavelet packet decomposition. [13, 20-22]

A three level wavelet packet decomposition is applied on a windowed signal which results in eight sub-bands of 500 Hz bandwidth. Further, the lowest four sub-bands are applied with one level WP decomposition to get eight sub-bands of 250 Hz

Table 1 Some notations used in this work

Symbol	Definition
E_i	Energy of i^{th} sub band
s_{ik}	k^{th} sample of i^{th} sub band
n	Number of frames in keyword template
m	Number of frames in unknown utterance, m>n
U_i	Feature vector sequence of keyword template of i^{th} frame = $\{u_{i1}, u_{i2}, u_{i3}, \ldots u_{ik}\}$, where u_{ij} represents j^{th} feature vector of i^{th} frame
V_i	Feature vector sequence of unknown utterance of i^{th} frame = $\{v_{i1}, v_{i2}, v_{i3}, \ldots v_{ik}\}$, where v_{ij} represents j^{th} feature vector of i^{th} frame
U	k x n matrix
V	k x m matrix
D	local distance matrix of the order n x m
d_{ij}	i^{th} row and j^{th} column element of D matrix
A	Global distance matrix of the order n x m
a_{ij}	i^{th} row and j^{th} column element of A matrix

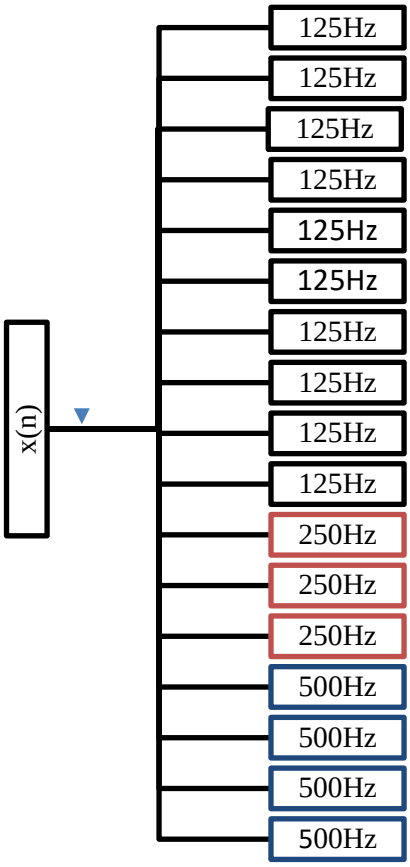


Figure. 1 Wavelet filter bank tree structure

bandwidth. Then the lowest five sub-bands of 250 Hz are further decomposed, resulting into ten sub-bands of 125 Hz. The optimal filter bank which imitates the auditory scale of the human ear is shown in Fig. 1. Henceforth there are totally seventeen filter banks starts from the lowest frequency of 0 Hz, ten sub-bands of 125 Hz, three sub-bands of 250 Hz and finally four sub-bands of 500 Hz up to 4 kHz.

The Hamming windowed signal is passed through these filter banks. The energy of these seventeen filter banks is calculated.

Assume i[th] sub-band with s_i samples, and then the energy E_i is calculated as given in Eq. (1). Then the log energy is computed as given by Eq. (1).

$$E_i = \frac{\sum_k s_{ik}^2}{\sum_i \sum_k s_{ik}^2} \quad , \; i=1 \text{ to } 17 \tag{1}$$

Discrete Cosine Transform (DCT) is applied to de-correlate the coefficients. Next, Cepstral Mean Normalisation (CMN)[23] is applied to the cepstral coefficient to compensate the channel noise. A sequence of cepstral coefficients $\{c_1, c_2, ..c_{17}\}$ obtained from one speech frame are considered, then CMN is calculated as given in Eq. (2), where μ_c is the mean feature vector from each vector c_t and σ_c^2 is the variance used to obtain the normalized vector $\hat{c}_t$

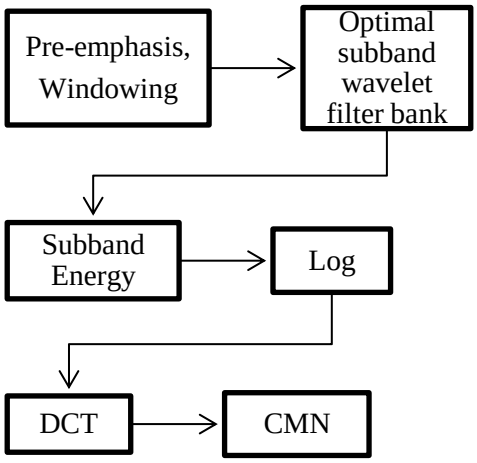


Figure. 2 Wavelet Feature Extraction method

$$\hat{c}_t = \frac{c_t - \mu_c}{\sigma_c} \tag{2}$$

where $\mu_c = \frac{1}{T}\sum_t c_t$ and $\sigma_c^2 = \frac{1}{T}\sum_{t=1}^{T}(c_t^2 - \mu_c^2)$.

The block diagram representation of Wavelet Cepstral Coefficient (WCC) method is as shown in Fig. 2.

2.2 DTW algorithm

Let the keyword template utterance be represented by a feature vector sequence $\{U_1, U_2, U_3, ...U_n\}$, where n is the number of frames in the keyword template. Each A_i is represented by $\{u_{i1}, u_{i2}, u_{i3}, ...u_{ik}\}$ where k is the number of feature vectors per frame in which k varies with the type of feature extraction technique being preferred. Similarly, assume the unknown spoken utterance be represented by a feature vector sequence $\{V_1, V_2, V_3, ...V_m\}$ in which m is the number of frames in the unknown spoken utterance where each Bi is represented by $\{v_{i1}, v_{i2}, v_{i3}, ...v_{ik}\}$ and finally the total number of features in keyword with unknown utterance are calculated as n x k, m x k correspondingly. To compare time series sequence of different lengths, the sequences must be warped in a dynamic manner [12, 13]. The DTW algorithm will find out the warping path between keyword and unknown utterance. Here, the objective is to find out whether there is a keyword present in the long spoken utterance so that the number of frames in the utterance 'm' is going to be always bigger than 'n'. The computational complexity of DTW is $O(n^2)$ where n is the maximum length of the two-time series.

Conventional DTW algorithm:

The time series sequence represented by U, V as below.

$$U=\{u_1, u_2, u_3 \ldots u_n\}$$
$$V=\{v_1, v_2, v_3 \ldots v_m\}$$

Here the absolute distance between the two elements u_i, v_j is d_{ij}. This results in a local distance matrix [D] of length $n \times m$ as given by the Eq. (3),

$$d_{ij}= abs(u_i- v_j) \quad (3)$$

The global distance matrix [A] calculated from the local distance matrix through the following steps.

1. Start with the calculation of $a(1,1) = d(1,1)$

2. Calculate the first row as given in Eq. (4),

$$a(i,1) = a(i,1)+d(i,1). \quad (4)$$

Calculate the first column as given in Eq. (5),

$$a(1,j) = a(1,j)+d(1,j). \quad (5)$$

3. The second row to the last row is calculated by Eq. (6),

$$a(i,2)=min[a(i,1),a(i-1,1),a(i-2,2)]+d(i,2) \quad (6)$$

4. Rest of the rows from left to right and from bottom to top with the rest of the grid are calculated as in Eq. (7),

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

5. Trace back the best path through the grid starting from $a(n, m)$ and moving towards $a(1,1)$ by following the minimum score path. Then the GWC is given by Eq. (8),

$$GWC = \frac{1}{N}\sum_{i=1}^{p} W_i \quad (8)$$

Where W_i is the cost along the warping path and $N= m+n$.

3. Proposed algorithm

3.1 Modified DTW algorithm

In the proposed modified DTW algorithm, let the keyword template utterance represented by a feature vector sequence is given by Eq. (9),

$$U = \{u_{11,}\, u_{12},..,u_{1k},u_{21},..\, u_{2k},..u_{n1,}..,u_{nk}\} \quad (9)$$

Similarly the unknown spoken utterance is represented by a feature vector sequence as given by Eq. (10),

$$V= \{v_{11,}v_{12},,..v_{1k},v_{21,}v_{22},..v_{2k},...v_{m1},v_{m3},..v_{mk}\} \quad (10)$$

Here in the Eqs. (9) and (10), k is the number of feature vectors per frame, n is the number of frames in keyword template and m is the number of frames in the unknown utterance. U, V could be treated a Multivariate Time Series (MTS) instead of a univariate time series (UTS) as matrices of the order $[k \times n]$, $[k \times m]$ which is given by Eq. (11),

$$U = \begin{bmatrix} u_{11} & \cdots & u_{1n} \\ \vdots & \ddots & \vdots \\ u_{k1} & \cdots & u_{kn} \end{bmatrix}, V = \begin{bmatrix} v_{11} & \cdots & v_{1m} \\ \vdots & \ddots & \vdots \\ v_{k1} & \cdots & v_{km} \end{bmatrix} \quad (11)$$

If the traditional UTS dynamic time warping algorithm is applied, it may not capture the properties of all the dimension of the feature vector and also similarity test would not reflect correctly. Subsequently, for this keyword spotting, feature vectors are treated as MTS.

In the traditional DTW, Eq. (3) is used for computing local distance matrix, which is replaced by MTS based Euclidean distance. Hence Eq. (3) is modified as given in Eq. (12),

$$d_{ij}= \sum_{p=1}^{K}(u_{pi} - v_{pj})^2 \quad (12)$$

On implementing these changes, local distance matrix reduced from $[nk \times mk]$ to an $[n \times m]$ matrix which in turn reduces the order of complexity by k^2 where k is the dimension of the feature vector per frame thereby it increases the speed up time in the search operation in the long utterance. Because of this, MTS based Euclidean DTW (EDTW) algorithm will speed up the search operation in the long utterance. Next, Global distance matrix $[A]$ equation or otherwise the Dynamic Programming (DP) Eq. (7) is modified with respect to the search path. The traditional DTW global distance, this will search for immediate minimum neighbour among $(a(i,j-1),a(i-1,j-1),a(i-1,j))$ instead if search path could skip the immediate neighbour , look in the very next neighbour. This is called skip a distance by one unit. The traditional search is shown in Fig. 3a. The modified skip distance is shown in Fig. 3b and 3c. This is also represented by the Eqs. (13) and (14).

$$a(i,j)=min\,[a(i-1,j-2), a(i-1,j), a(i-1,j-1)] + d(i,j) \quad (13)$$

$$a(i,j)=min\ [a(i-1,j-2),\ a(i-2,j-1),\ a(i-1,j-1)] + d(i,j). \quad (14)$$

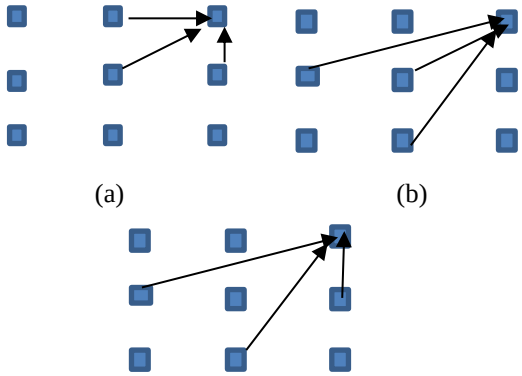


Figure. 3 DP equation: (a) Original, (b) modified skip distance with two, and (c) modified skip distance with three

The objective for this modification is to have a faster search of warping path and also the computation of global warping cost. Global warping cost equation is same as the traditional DTW. Threshold value defined by Eq. (15).

$$Threshold = mean - \alpha \times std \quad (15)$$

4. Experimental setup

The proposed keyword spotting system is shown in Fig. 4. Speech signals from TIDIGITS corpus are processed with a frame size of 25ms with an overlap of 15 msec. Baseline MFCC[24] features are extracted with a dimension of 39 vectors per frame which includes static and dynamic features. As discussed in section 2, WCC features of 34 vectors are extracted with static features of 17 and dynamic features of 17. The extracted features are aligned to a matrix form of the order of $k \times n$, $k \times m$, where k is the dimension of feature vectors per frame, n is the number of frames per keyword, m is the number of frames in the unknown spoken utterance.

Researchers have experimented DTW algorithm for KWS with a template length of half keyword to two keyword length [3] with a sequential shifting of the template along the unknown utterance of one frame to one whole length of keyword [8]. In this study, the experiment is conducted with a sliding window of one keyword length or half keyword length with a shift of one keyword length shift along the unknown utterance.

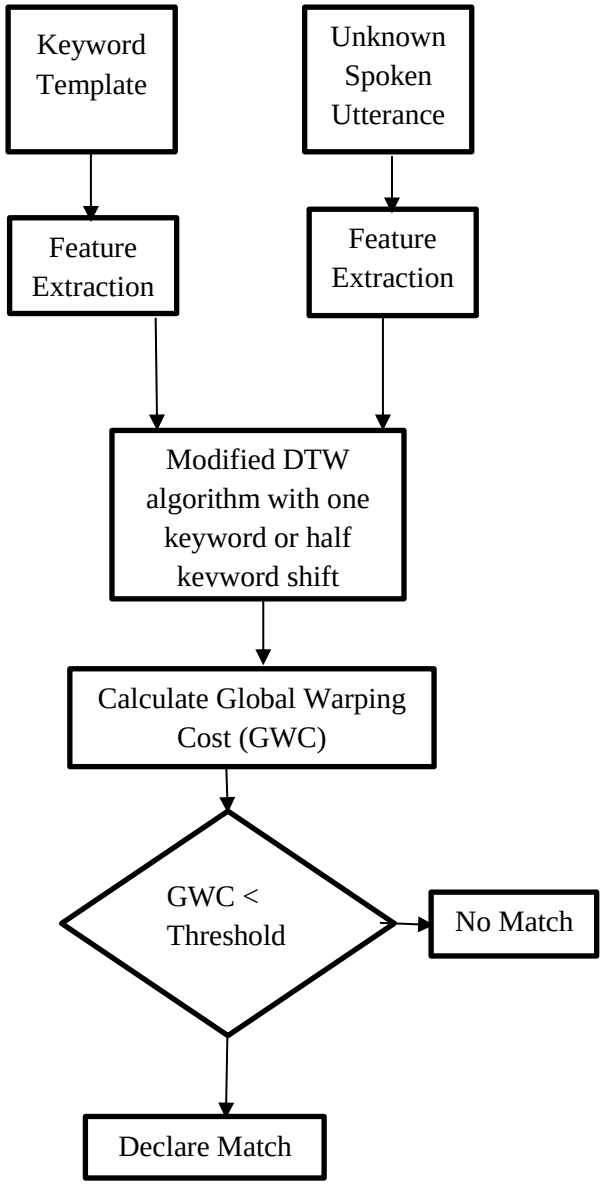


Figure. 4 Proposed spoken keyword spotting system

The baseline experiment considered for this paper is traditional DTW which has city block distance used for computing the local distance matrix and original DP equation for global measurement with MFCC and WCC features along with a hashing algorithm [8]. Furthermore, two experiments have been performed, one with whole KW length shift, other with half KW length shift. In these two experiments, the proposed modified EDTW has experimented.

4.1 Corpus

The experiment is done on LDC93S10-TIDIDGITS corpus[25] in which numbers from zero to nine are uttered continuously and also isolated digits uttered by the same speaker. The corpus has

two subsets one for training namely "train" and one for evaluation namely "test". Train set has a collection of utterances from 55 men speakers, 57 women speakers, 25 boys and 26 girls. Test set has a collection of utterances from 56 men speakers, 57 women speakers, 25 boys and 25 girls. Each speaker uttered 22 isolated digits, 11 two digit sequence, 11 three digit sequence, 11 four digit sequence, 11 five digit sequence, 11 seven digit sequence. For this research work, isolated digits are treated as keyword template; they are searched through multi-digit sequences from another speaker to test speaker independent spoken keyword spotting system.

4.2 Experiments

As explained in section 2, WCC features are extracted for keyword templates one to nine digits. WCC features with a dimension of 34 are extracted per frame. For MFCC features, 39 feature vectors are computed per frame.

The first experiment is conducted with one keyword template length and one keyword shift along unknown digit sequence. This experiment is performed on MFCC feature with a proposed multivariate time series based EDTW for local distance matrix calculation along with two skip distance for global measurement. The same experiment is repeated on WCC feature set. Another subsection of the experiment is proposed multivariate time series based EDTW for local distance matrix but with skip distance of three for global distance measurement.

The second experiment is conducted with template length of one keyword and half keyword length shift along unknown digit sequence. This experiment is performed same as the first experiment on WCC and MFCC feature with multi-variate time series based Euclidean local distance matrix and skip distance of two or three for global distance measurement. Digits from one to nine are searched through 9 different samples of connected digit sequence which of either a 5 digit sequence or 6 digit sequence or 7 digit sequence. If the keyword digit exists in the sequence and if it is detected by the proposed algorithm, then it is counted as correct and if it is not detected then it is treated as a miss. If the keyword does not exist in the sequence but it is detected as a keyword, then it is treated as a false match.

5. Results and discussion

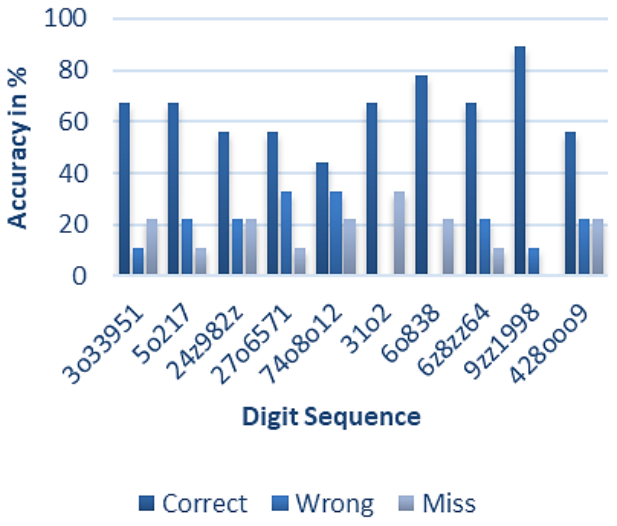
As mentioned in section 4, series of experiments are performed and the naming conventions of the experiment are given in Table 2.

Table 2. Experiment list

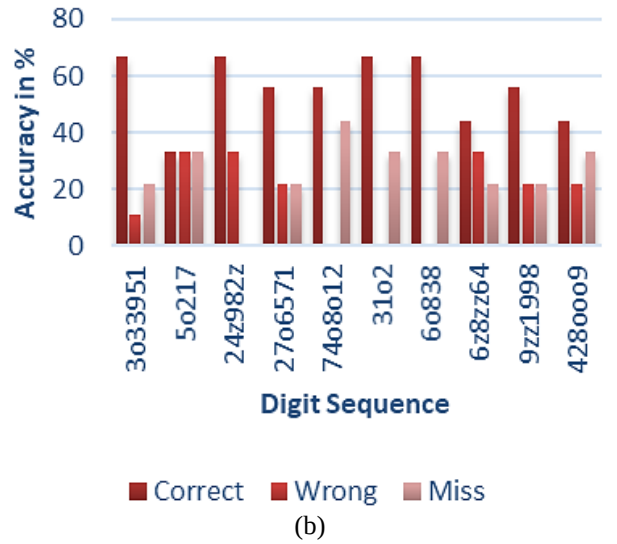
Abbreviation	Method
BASE_MFCC	Univariate traditional DTW with hashing algorithm on MFCC feature
BASE_WCC	Univariate traditional DTW with hashing algorithm on WCC feature
MEDTW_1KW_M	Multivariate Euclidean DTW with one keyword sequential shift on MFCC feature
MEDTW_1 KW _W	Multivariate Euclidean DTW with one keyword sequential shift on WCC feature
MEDTW_1KW _2S_M	Multivariate Euclidean DTW with two skip distance for global measurement with one keyword sequential shift on MFCC feature
MEDTW_1KW _2S _W	Multivariate Euclidean DTW with two skip distance for global measurement with one keyword sequential shift on WCC feature
MEDTW_1KW _3S _M	Multivariate Euclidean DTW with three skip distance for global measurement with one keyword sequential shift on MFCC feature
MEDTW_1KW_3S _W	Multivariate Euclidean DTW with three skip distance for global measurement with one keyword sequential shift on WCC feature
MEDTW_0.5KW_M	Multivariate Euclidean DTW with half keyword sequential shift on MFCC feature
MEDTW_0.5KW_ W	Multivariate Euclidean DTW with half keyword sequential shift on WCC feature
MEDTW_0.5KW_2S_M	Multivariate Euclidean DTW with two skip distance for global measurement with half keyword sequential shift on MFCC feature
MEDTW_0.5KW_2S_W	Multivariate Euclidean DTW with two skip distance for global measurement with half keyword sequential shift on WCC feature
MEDTW_0.5KW_3S_M	Multivariate Euclidean DTW with three skip distance for global measurement with half

	keyword sequential shift on MFCC feature
MEDTW_0.5KW_3S_W	Multivariate Euclidean DTW with three skip distance for global measurement with half keyword sequential shift on WCC feature

The keywords from one to nine are searched through ten sequences of various lengths. The baseline experiment is conducted with a traditional DTW algorithm along with hashing algorithm as in paper [8]. Since the time series of length *kn* and is reduced to ***n*** and *km* is reduced to *m* by the hashing algorithm, the local distance matrix is of dimension *nm* as that of the proposed MTS based EDTW. The results are shown in Fig. 5a, 5b.



(a)



(b)

Figure. 5 Keyword spotting accuracy of baseline system: (a) using MFCC feature and (b) using WCC feature

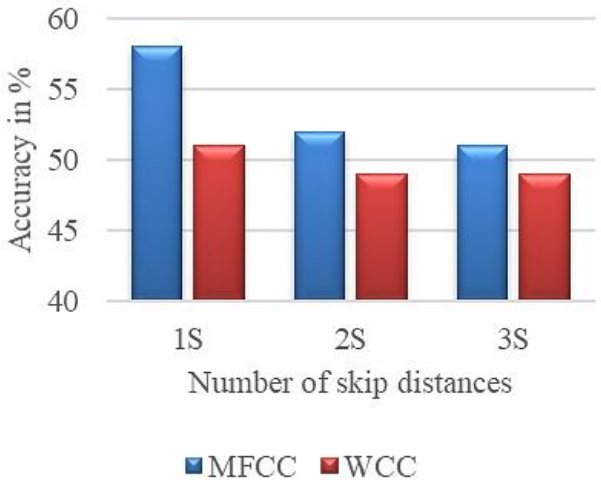


Figure. 6 Detection accuracies of the proposed modified DTW algorithm for one keyword shift

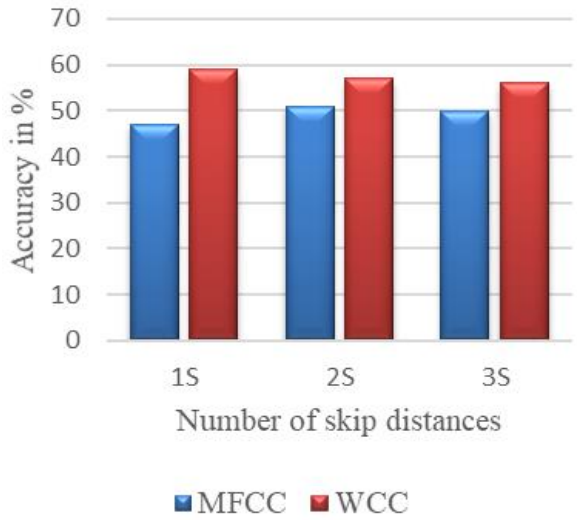


Figure. 7 Detection accuracies of the proposed modified DTW algorithm for half keyword length sequential shift

The proposed modified EDTW with a sliding shift of one keyword along the unknown utterance is evaluated with MFCC, WCC features are presented in Fig. 6. It can be observed that MFCC feature shows better performance than wavelet features for clean data. It is interesting to note that the overall correct detection is 48% to 57% with the different skip distance methods. The highest correct detection of 57% is observed for traditional global distance with MFCC features. The wavelet features applied on the proposed approach with 2–skip distance, 3-skip distance yielded the lowest score of 48%.

The results obtained from the proposed multivariate Euclidean DTW [MEDTW] with a half keyword sequential shift along the unknown utterance are shown in Fig. 7 for MFCC and wavelet features. It is interesting to note that there is a consistently better performance with wavelet features than traditional spectral feature such as MFCC in a

clean environment. It is also observed that static 17-dimensional wavelet feature per frame produced better detection rate than the static feature combined with a dynamic feature which is of 34 dimensions. So, we used the static wavelet features alone for experimentation instead of combined static and dynamic features which also added the advantage of fewer features which in turn results in less computational complexity. The highest score of 59% is noted for static wavelet features applied on the proposed MEDTW with traditional global distance calculation. The lowest score of 47% is indicated by MFCC feature of 39 dimensions original global distance calculation.

The traditional spectral features yielded high performance in a clean environment than wavelet feature for one keyword sliding shift along the unknown utterance. However, with half keyword sliding shift along unknown utterance, the static wavelet features yielded gain of 12% detection score compared to spectral features. When compared one keyword shift along the unknown utterance with half keyword shift, half keyword shift yielded small gain of 2% over one keyword shift.

From the previous results, the evaluation for noisy environment is performed on the proposed MEDTW with half keyword shift along the unknown utterance. The TIDIGITS corpus is added with Additive White Gaussian noise (AWGN) of 0dB SNR, 10dB SNR to evaluate the proposed system for noisy environment simulation. Fig. 8 present the performance of the proposed system for a noisy environment. For the 10dB SNR, both the wavelet feature and spectral feature showed the same performance. For 0dB SNR, wavelet feature indicated a gain of 3% over MFCC features. This small gain could be due to the localizing multi-resolution time-frequency information of wavelet transform.

Figure. 8 Detection accuracies of the proposed modified DTW algorithm for half keyword length sequential shift in noisy environment

6. Conclusion

Spoken keyword detection is a state-of-the-art research in the field of speech recognition because of massive digitization of speech data. Spoken keyword detection is used in interactive voice response systems, call center, call monitoring for customer care service and call surveillance of national security. Even though the dynamic time warping is one of the earliest template technique adopted for speech recognition, this method has been revisited because of the computing power evolution. Traditional dynamic time warping use single dimensional time series sequence. In this research work, multi-dimensional feature set such as MFCC, wavelet features along with Multivariate Euclidean Dynamic Time Warping (MEDTW) is proposed for keyword detection system and evaluated in the clean and noisy environment. MFCC features on MEDTW showed a gain over wavelet features for one keyword length sliding along the unknown utterance. However, wavelet feature yielded a gain of 12% over MFCC feature for half keyword sliding shift along the unknown utterance in a clean environment. In addition to that, it is observed that wavelet features yielded 3% improvement in the noisy environment over MFCC features [8]. It is remarkable to observe that the dimension of wavelet feature is less than half of the MFCC feature dimension. The proposed EDTW reduced the complexity of the algorithm from $O((nm)^2)$ to $O((m)^2)$ where n is the dimension of the feature vector per frame and m is the number of frames in the keyword template. In future, MEDTW can be applied to other set of speech features such as spectrograms to see the efficacy of the proposed method.

References

[1] J. Bridle, “An efficient elastic template method for detecting given keywords in the running speech”, In: *Proc. of Br. Acoust. Soc. Meet.*, Vol. 1, No. 1, pp. 1–4, 1973.

[2] W. Li and A. Billard, “Keyword Detection for Spontaneous Speech”, *Signal Processing*, Vol. 1, No. 1, pp. 1–5, 1920.

[3] M. S. Barakat, C. H. Ritz, and D. A. Stirling, “Keyword spotting based on the analysis of template matching distances”, In: *Proc. of 5th Int. Conf. Signal Process. Commun. Syst. ICSPCS’2011*, pp. 1–6, 2011.

[4] A. Jansen and P. Niyogi, “An Experimental Evaluation of Keyword-Filler Hidden Markov Models”, pp. 1–10, 2009. https://newtraell.cs.uchicago.edu/files/tr_authentic/TR-2009-02.pdf , Accessed on 05/16/2017

[5] M. Park and H. Kim, "Different Filler Models for Keyword Recognizer", pp. 1-4. https://srtlab.kaist.ac.kr/common/download.php?tbl=public_tbl&pname=pu...1%0A, Accessed on 05/06/2017

[6] T. M. English and L. C. Boggess, "Back-Propagation Training of a Neural Network for Word Spotting", In: *Proc. of ICASSP-92, Int. Conf. Acoust. Speech, Signal Process*, Vol. 3, pp. 357–360, 1992.

[7] S. Jothilakshmi, "Spoken keyword detection using autoassociative neural networks", *Int. J. Speech Technol.*, Vol. 17, No. 1, pp. 83–89, 2014.

[8] J. S. R. Alex and N. Venkatesan, "Spoken utterance detection using dynamic time warping method along with hashing technique", *Int. J. Eng. Technol.*, Vol. 6, No. 2, pp. 1100–1108, 2014.

[9] S. Salvador and P. Chan, "FastDTW : Toward Accurate Dynamic Time Warping in Linear Time and Space", *J. Intell. Data Anal.*, Vol. 11, No. 5, pp. 561–580, 2007.

[10] R. Saabni and A. Bronstein, "Fast Key-Word Searching Using 'BoostMap' based Embedding", In: *Proc. of Front. Handwrit. Recognit. (ICFHR), 2012 Int. Conf. ., Bari Italy*, pp. 734–739, 2012.

[11] A. K. Vuppala, "Analysis of Constraints on Segmental DTW for the Task of Query-by-Example Spoken Term Detection", In: *Proc. of India Conf. (INDICON), 2015 Annu. IEEE, New Delhi, India*, pp. 1–6, 2015.

[12] Y. Zhang and J. R. Glass, "Unsupervised Spoken Keyword Spotting via Segmental DTW on Gaussian Posteriorgrams", In: *Proc. of the 2009 IEEE Workshop on Automatic Speech Recognition and Understanding*, ASRU 2009, Merano, Italy, pp. 398–403, 2009.

[13] K.P. Soman, *Insight into Wavelets from Theory to Practice*, 2nd edition. Prentice-Hall of India, New Delhi. 2005.

[14] Z. Wang, J. Yang, and X. Zhang, "Combined discrete wavelet transform and wavelet packet decomposition for speech enhancement", In: *Proc. of Signal Process. 2006 8th Int. Conf. on, Beijing, China*, pp. 6–9, 2006.

[15] Z. Tufekci, J. N. Gowdy, S. Gurbuz, and E. Patterson, "Applied mel-frequency discrete wavelet coefficients and parallel model compensation for noise-robust speech recognition", *Speech Communication*, Vol. 48, pp. 1294–1307, 2006.

[16] X.Y. Zhang, J. Bai, and W.Z. Liang, "The speech recognition system based on bark wavelet MFCC", In: *Proc. of Int. Conf. Signal Process. Proceedings, ICSP*, Vol. 1, pp. 1–3, 2007.

[17] O. Farooq and S. Datta, "Mel Filter-Like Admissible Wavelet Packet Structure for Speech Recognition", *IEEE Signal Process. Lett.*, Vol. 8, No. 7, pp. 196–198, 2001.

[18] H. R. Tohidypour, S. A. Seyyedsalehi, and H. Behbood, "Comparison between wavelet packet transform, bark wavelet & mfcc for robust speech recognition tasks", In: *Proc. of Intl. Conf. on Industrial Mechatronics and Automation* Vol. 11, No. 2, pp. 329–332, 2010.

[19] J. Mei, S. Member, M. Liu, and Y. Wang, "Learning a Mahalanobis Distance based Dynamic Time Warping Measure for Multivariate Time Series Classification", *IEEE Transactions on Cybernetics* , Vol. 46, No. 6, pp. 1363 - 1374, 2015.

[20] S. Mallat, *A Wavelet Tour of Signal Processing* ,Academic, New York, 1998.

[21] I. Daubechies, *Ten lectures on Wavelets*, SIAM,Philadelphia,USA, 1992.

[22] T.N.G. Starang, *Wavelets and Filter Banks*, Wellesley,Wellesley-Cambridge press, MA, USA, 1997.

[23] N. S. Nehe and R. S. Holambe, "DWT and LPC based feature extraction methods for isolated word recognition", *EURASIP J. Audio, Speech, Music Process.*, Vol.1,No. 1, pp. 1-7, 2012.

[24] J.S.R. Alex and N. Venkatesan, "Modified MFCC methods based on KL- transform and power law for robust speech recognition", *J. Theor. Appl. Inf. Technol.*, Vol. 67, No. 2, pp. 527–532, 2014.

[25] R. G. Leonard, and G. Doddington, *TIDIGITS LDC93S10. Web Download. Philadelphia: Linguistic Data Consortium*, 1993.