Spoken Digits Recognition using Weighted MFCC and Improved Features for Dynamic Time Warping

Santosh V. Chapaneri
Department of Electronics and Telecommunication Engineering,
St. Francis Institute of Technology,
University of Mumbai

ABSTRACT
In this paper, we propose novel techniques for feature parameter extraction based on MFCC and feature recognition using dynamic time warping algorithm for application in speaker-independent isolated digits recognition. Using the proposed Weighted MFCC (WMFCC), we achieve low computational overhead for the feature recognition stage since we use only 13 weighted MFCC coefficients instead of the conventional 39 MFCC coefficients including the delta and double delta features. In order to capture the trends or patterns that a feature sequence presents during the alignment process, we compute the local and global features using Improved Features for DTW algorithm (IFDTW), rather than using the pure feature values or their estimated derivatives. The experiments based on TI-Digits corpus demonstrate the effectiveness of proposed techniques leading to higher recognition accuracy of 98.13%.

General Terms
Speech Processing, Speech Recognition

Keywords
Speech recognition, MFCC, Dynamic time warping

1. INTRODUCTION
The objective of automatic speech recognition (ASR) systems is to recognize the human speeches, such as words and sentences, using algorithms evaluated by a computer without the interference of humans. ASR is essentially a pattern recognition task, the goal is to take one pattern, i.e. the speech signal, and classify it as a sequence of previously learned patterns, e.g. words or sub-word units such as phonemes [1]. Speech recognition systems can be characterized by many parameters, such as speaking model, speaking style, vocabulary, etc. An isolated-word speech recognition system requires that the speaker pause briefly between words, whereas a continuous speech recognition system does not. Spontaneous, or extemporaneously generated, speech contains disfluencies, and is much more difficult to recognize than speech read from script. Some ASR systems require speaker enrollment, a user must provide samples of his or her speech before using them, whereas other systems are speaker-independent, in that no enrollment is necessary. Speech recognition is the process of converting an acoustic signal captured by a microphone to a set of words. The recognized words can be the final result, as for applications such as commands and control, data entry and document preparation. They can also serve as the input to further linguistic processing in order to achieve speech understanding [2]. Speech recognition techniques are often seen as an alternative to typing on a keyboard or touching smart phones or tablet. They help people with a variety of disabilities to communicate with a computer.

There are two main phases in a speech recognition system: training and recognition. 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 training vectors (patterns) are stored in a database for subsequent use in the recognition phase. During the recognition phase, the user speaks any word for which the system was trained. A test pattern is generated for that word and the corresponding text string is displayed as the output using a pattern comparison technique. The system block diagram is shown in Fig. 1.

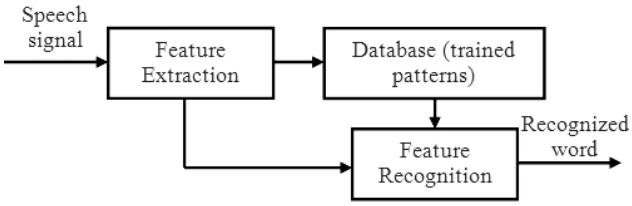

Fig. 1: Speech recognition system block diagram

Several different techniques for feature extraction exists, the most common being linear predictive coding (LPC) and Mel frequency cepstral coefficients (MFCC). LPC is a time-domain technique and suffers from variations in the amplitude of the speech signal due to noise [3, 4]. The preferred technique for feature extraction is MFCC [5, 6, 7] 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 [8].

For feature recognition stage, several techniques are available including analysis methods based on Bayesian discrimination [9], Hidden Markov Models (HMM) [10], Dynamic Time Warping (DTW) based on dynamic programming [11, 12, 13], Support Vector Machines [14], Vector Quantization [15], and Neural Networks [16]. DTW is an algorithm developed by the speech recognition community to handle the matching of non-linearly expanded or contracted speech signals. In this work, we use DTW because of its simplicity in hardware implementation and it is also widely used in small-scale embedded systems (e.g. cell phones, mobile applications, etc.).

The paper is organized as follows. Section 2 explains the pre-processing done on spoken speech digits followed by Section 3 which details the procedure of conventional MFCC feature extraction and the proposed Weighted MFCC (WMFCC) technique. Section 4 explains the feature recognition technique using conventional DTW and the proposed Improved Features for DTW (IFDTW) algorithm. Section 5 demonstrates the experimental results followed by conclusions in Section 6.

2. PRE-PROCESSING OF SPEECH

Accurate detection of speech in the presence of background noise is important to constrain the amount of processing that is needed for recognition. An endpoint detection algorithm [17] is applied to the speech signal to find the beginning and end of each spoken digit, and to remove the silence and noise region. The algorithm uses signal features based on energy level and zero-crossing rate. Fig. 2 shows the result of endpoint detection algorithm for spoken digit "2" along with its energy and zero-crossing rate plots. The solid vertical lines in the top plot indicate the portion of resultant speech void of silence regions.

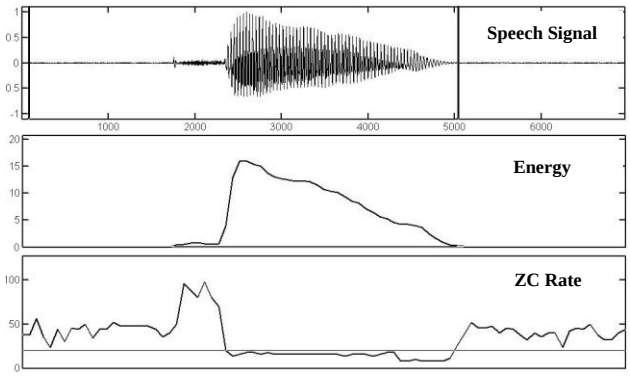

Fig. 2: Result of Endpoint Detection for spoken digit "2"

3. WEIGHTED MFCC

The first step of speech recognition process is to extract the features from the speech signal. The purpose of feature extraction is two-fold: first is to compress the speech signal into features, and second is to use features that are insensitive to speech variations, changes of environmental conditions and independent of speaker. The procedural steps of MFCC feature extraction is described as follows:

3.1 Pre-emphasis

The speech signal spectrum is pre-emphasized by approximately 20 dB per decade to flatten the spectrum of the speech signal [5]. The pre-emphasis filter is used to offset the negative spectral slope of the speech signal to improve the efficiency of the spectral analysis [18].

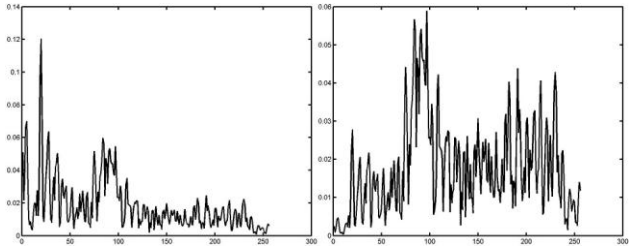

Fig. 3: Spectrum before (left) and after Pre-emphasis (right)

The filter transfer function is given by $H(z) = 1 - az^{-1}$, where '*a*' is between 0.9 and 1. Fig. 3 shows the effect of pre-emphasis on the spectrum of a speech signal with the value of 0.97 for *a*.

3.2 Framing

Since the human speech signal is slowly time varying, it can be treated as a stationary process when considered under a short time duration [5]. Therefore, the speech signal is usually separated into small duration blocks, called frames, and the spectral and cepstral analysis is performed on these frames. Typically, the frame length is kept as 25 milliseconds and the neighboring frames are overlapped by 10 milliseconds. The frame shift is the frame length minus the frame overlap.

3.3 Windowing

After being partitioned into frames, each frame is multiplied by a window function prior to the spectral analysis to reduce the discontinuity introduced by the framing process by attenuating the values of the speech samples at the beginning and end of each frame. Typically, Hamming window is used [19].

3.4 Spectral Estimation

The spectral coefficients of the speech frames are estimated using the Fast Fourier Transform (FFT) algorithm. These coefficients are complex numbers containing both magnitude and phase information. However, for speech recognition, the phase information is usually discarded and only the magnitude of the spectral coefficients is retained [19].

3.5 Mel Filtering

The spectrum of speech signal is then filtered by a group of triangular bandpass filters as shown in Fig. 4 that simulate the characteristics of human's ear. The purpose of Mel filtering is to model the human auditory system that perceives sound in a nonlinear frequency binning [5].

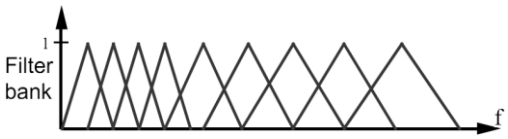

Fig. 4: Triangular bandpass Mel filter bank

The ears analyze the spectrum of the sound in groups according to a series of overlapped critical bands. The critical bands are distributed in a way that the frequency resolution is high in low frequency region and low in high frequency region. The bandwidth of the window is narrow in low frequency and gradually increases for high frequency. The edge of the window is arranged so that it coincides with the center of the neighboring window [20]. To decide the location of the Mel frequency of the center of the windows, the Mel frequencies for minimum and maximum linear frequency are first calculated using:

$$f_{\text{Mel}} = 2595 \times \log_{10}(1 + f/700) \qquad (1)$$

where f_{Mel} is the Mel frequency corresponding to the linear frequency *f*. The windows are linearly distributed in the Mel frequency scale, but when converted back to linear frequency, the center frequencies of the windows are logarithmically distributed.

3.6 Logarithmic Compression

While the Mel filtering approximates the non-linear characteristics of the human auditory system in frequency, the

natural logarithm deals with the loudness non-linearity. It approximates the relationship between the human's perception of loudness and the sound intensity [21]. Besides this, it converts the multiplication relationship between parameters into addition relationship [19]. The convolutional distortions, such as the filtering effect of microphone and channel, and the multiplication in frequency domain, become simple addition after the logarithm. The log Mel filter bank coefficients are computed from the filter outputs as:

$$S(m) = 20\log_{10}\left(\sum_{k=0}^{N-1}|X(k)||H(k)\right), \quad 0 < m < M \qquad (2)$$

where M is the number of Mel filters (20 to 40), *X(k)* is the N-point FFT of the specific window frame of the input speech signal, and *H(k)* is the Mel filter transfer function [19].

3.7 Discrete Cosine Transform (DCT)

The cepstrum is defined as the inverse Fourier transform of the log magnitude of Fourier transform of the signal. Since the log Mel filter bank coefficients are real and symmetric, the inverse Fourier transform operation can be replaced by DCT to generate the cepstral coefficients [22]. This step is crucial in speech recognition as it can separate the vocal tract shape function from the excitation signal of the speech production model. The lower order cepstral coefficients represent the smooth spectral shape or vocal tract shape, while the higher order coefficients represent the excitation information [18]. The cepstral coefficients are the DCT of the M filter outputs obtained as:

$$c(n) = \sum_{m=0}^{M-1} S(m)\cos\left(\frac{\pi n(m-1/2)}{M}\right) \qquad (3)$$

Typically, the first 13 cepstral coefficients are used. Another benefit of DCT is that the generated MFCC coefficients *c(n)* are less correlated than the log Mel filter bank coefficients.

3.8 Log Energy

In addition to the above MFCC features, the energy of the speech frame is also used as a feature [19]. The log energy, denoted as logE, is calculated directly from the time-domain signal of a frame as:

$$\log \mathrm{E} = \log \sum_{n=1}^{N} x(n)^2 \qquad (4)$$

where *x(n)* is the speech sample and *N* is the length of the frame. In this work, the cepstral coefficient *c(0)* is replaced by logE to give a more accurate energy feature.

3.9 Liftering

The higher order cepstral coefficients tend to be numerically small so that there is a large variation of cepstral coefficients between the low-order and high-order coefficients [23]. To re-scale the coefficients, a raised sine window bandpass liftering (filtering in cepstral domain) is used as follows:

$$c(n) \leftarrow \left(1 + \frac{L}{2}\sin\left(\frac{\pi n}{L}\right)\right)c(n) \qquad (5)$$

Typically, *L* is 12 for a 12-order cepstral vector.

3.10 Cepstral Mean Normalization (CMN)

This step ensures that all the features contribute equally. Without normalization, the feature with large dynamic range, such as the log energy feature, may dominate the distance metric in the feature recognition stage. CMN reduces the influence of additive white noise [5]. By subtracting the estimated mean of every channel noise in the cepstral domain, the average value of noisy speech can be reduced to almost zero. Assuming that the speech signal is divided into *K* frames, the normalized cepstral coefficients for frame *k* are calculated as

$$c_k(n) = c_k(n) - \frac{1}{K}\sum_{k=1}^{K} c_k(n) \qquad (6)$$

3.11 Delta and Double Delta Features

The trend of the speech signals in time is lost in the frame-by-frame analysis. To recover the trend information, the time derivatives (delta) and accelerations (double delta) are used [5]. For speaker-independent speech recognition system, these features are especially important. Although the location of the formant of the speech varies from person to person, the time trend of the formant is quite constant among different speakers. The trend information, represented by delta and double delta features, is important for improving the robustness of the recognition. The delta $\Delta c(n)$ features are calculated as follows:

$$\Delta c(n) = \frac{1}{\sum_{i=1}^{D} i^2}\sum_{i=1}^{D} i \times (c(n+i) - c(n-i)) \qquad (7)$$

where *c(n)* are the MFCC coefficients for each frame, and *D* is typically set to 2. The double delta features $\Delta\Delta c(n)$ are calculated similarly from the delta features. These derived features are concatenated to the original cepstral features, thus giving us a 39-dimensional MFCC feature vector for each frame, which are 12 MFCC, 1 energy, 12 delta MFCC, 1 delta energy, 12 double delta MFCC, and 1 double delta energy features.

3.12 Proposed Weighted MFCC (WMFCC)

The delta and double delta features improve the overall accuracy of the speech recognition system; however, this approach increases the dimension of the feature vector leading to higher computational complexity overhead in the recognition stage. Several modifications to MFCC feature extraction have been reported in the literature [24, 25]. However, to reduce the dimensions of feature vector while still retaining the advantages of delta and double delta features, we propose a simple technique of weighted MFCC (WMFCC) feature vector as follows:

$$wc(n) = c(n) + p \bullet \Delta c(n) + q \bullet \Delta\Delta c(n) \qquad (8)$$

where the delta and double features are weighted according to *p* and *q*, respectively. Since these derivative features contribute slightly less than *c(n)*, the weights are constrained to be $q<p<1$. The final feature vector *wc(n)* is 13-dimensional thus reducing the complexity overhead of the recognition stage. Also as shown in Fig. 5, we note that WMFCC and conventional MFCC have similar amplitude curves and thus can be effective in speech recognition as demonstrated in Section 5.

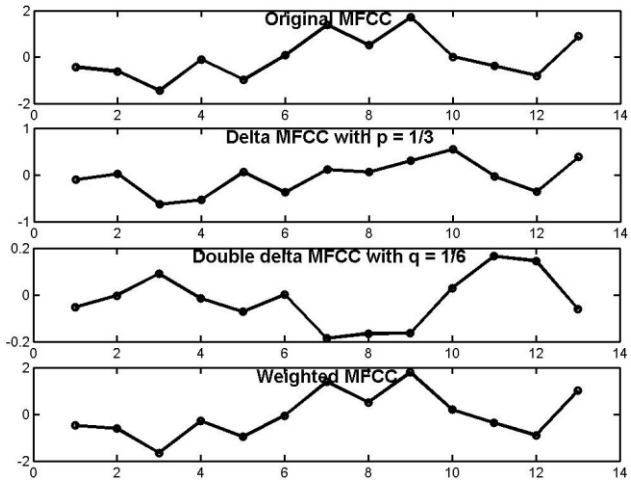

Fig. 5: Weighted MFCC (13-dimensional) of a speech frame

4. IMPROVED FEATURES FOR DTW

4.1 Conventional Dynamic Time Warping

Unlike Linear Time Warping (LTW) which compares two time series based on linear mapping of the two temporal dimensions, Dynamic Time Warping (DTW) allows a non-linear warping alignment of one signal to another by minimizing the distance between the two as shown in Fig. 6.

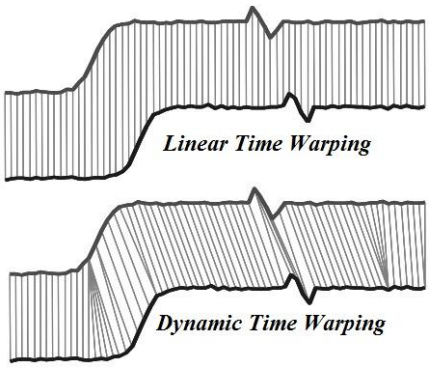

Fig. 6: DTW non-linear alignment of two time series

This warping between two signals can be used to determine the similarity between them and thus it is very useful for feature recognition. In a speech waveform, the duration of each spoken digit can vary but the overall speech waveforms are similar for the same digit. DTW is a pattern matching algorithm with a non-linear time optimization effect based on Bellman's principle of optimality [19], which states that given an optimal path from A to B and a point C lying somewhere along this path, the path segments AC and CB are optimal paths from A to C and C to B respectively. The DTW problem can be formulated as follows: Given two time series X and Y of lengths $|X|$ and $|Y|$,

$$X = x_1, x_2, \ldots, x_i, \ldots x_{|X|}$$
$$Y = y_1, y_2, \ldots, y_j, \ldots y_{|Y|} \qquad (9)$$

construct a warp path W:

$$W = w_1, w_2, \ldots, w_K \quad \max(|X|,|Y|) \le K \le |X| + |Y| \qquad (10)$$

where K is the length of the warp path and the k^{th} element of warp path is $w_k = (i, j)$, where i is an index from time series X, and j is an index from time series Y. The warp path must start at the beginning of each time series at $w_1 = (1, 1)$ and finish at the end of both time series at $w_K = (|X|, |Y|)$. This ensures that every index of both time series is used in the optimal warping path. There is also a constraint on the warp path that forces i and j to be monotonically increasing and every index of each time series must be used.

$$w_k = (i, j), w_{k+1} = (i', j') \quad i \le i' \le i+1,\ j \le j' \le j+1 \qquad (11)$$

The optimal warp path is the warp path with the minimum distance, where the normalized distance of a warp path W is:

$$Dist(W) = \sum_{k=1}^{K} Dist(w_{ki}, w_{kj}) \Big/ (|X| + |Y|) \qquad (12)$$

where $Dist(w_{ki}, w_{kj})$ is the distance metric, either Euclidean or City-block [26], between the two data point indices (one from X and one from Y) in the k^{th} element of the warp path. Instead of attempting to find the minimal distance all at once, a dynamic programming approach is used by finding solutions to sub-problems and using this repeatedly to find solutions to a slightly larger problem until the final minimum distance is obtained [5, 19]. A two-dimensional cost matrix D is constructed where the value at $D(i, j)$ is the distance of the warp path. Fig. 7 shows an example of a cost matrix and a minimum-distance optimal warp path traced through it.

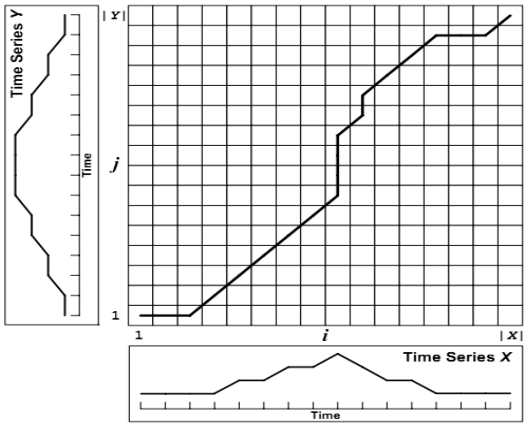

Fig. 7: Minimum-distance warping path for two time series

The rationale behind dynamic programming approach is that since the value at $D(i, j)$ is the minimum warp distance, then if the minimum distances are already known for smaller portions that are a single data point away from i and j, then $D(i, j)$ is the minimum distance of all possible warp paths for time series that are one data point smaller than i and j, plus the distance between the two points x_i and y_j. Thus, the distance can be evaluated recursively as:

$$D(i, j) = Dist(i, j) + \min[D(i-1, j), D(i-1, j-1), D(i, j-1)] \qquad (13)$$

The slope constraint condition on DTW states that the warp path should not be too steep or too shallow [5]. This prevents very short sub-sequences to match very long ones. This slope is expressed as a/b, where b is the number of steps in x direction and a is the number of steps in y direction. After b steps in x, the path must take a step in y, and vice-versa. Further, there is a constraint on adjustment window to speed up the calculations since an intuitive alignment path is unlikely to drift very far from the diagonal. The distance that the warp path is allowed to wander is limited to a window or band of size R, directly above and to the right of the diagonal. Fig. 8 illustrates the two window bands widely used in DTW.

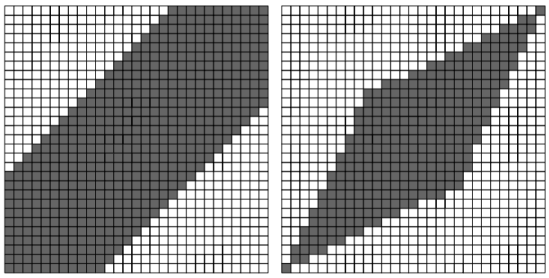

Fig. 8: Adjustment window constraints: Sakoe-Chiba band (left) [11] and Itakura Parallelogram (right) [27]

In application to speech recognition, the two time series corresponds to the two *numCoefficients* x *numFrames* MFCC feature vectors of different speech signals. A two-dimensional cost matrix is computed that stores the minimum distance between two feature vectors $\boldsymbol{x}_i$ and $\boldsymbol{y}_j$. The spoken digit's feature vector is compared to the template feature vectors using DTW and the one with the minimum distance is chosen as recognition output.

4.2 Derivative Dynamic Time Warping

The fundamental flaw of conventional DTW is that the numerical value of a data point in a time series does not represent the complete picture of the data point in relation to the entire sequence. In [28], derivative DTW was proposed in which each data point is replaced by its first derivative. This estimated derivative serves as the local feature of a point that expresses its relationship with two adjacent neighboring data points. The derivative estimates for MFCC feature vector $\boldsymbol{X}$ are computed as:

$$D(x_i) = \frac{(x_i - x_{i-1}) + (x_{i+1} - x_{i-1})/2}{2}, \quad 1 < i < K \qquad (14)$$

where K is the number of frames. This estimate is not defined for the first and last vectors of the feature sequence. Similarly, derivative estimates are computed for the feature vector $\boldsymbol{Y}$ and conventional DTW algorithm is applied to these derivative features. Fig. 9 illustrates the alignment by conventional DTW and derivative DTW where we note that the conventional DTW produces multiple singularities.

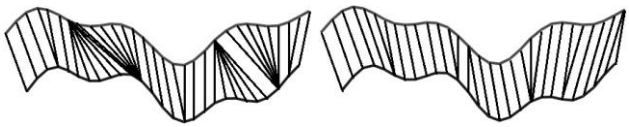

Fig. 9: Alignment produced by conventional DTW (left) and derivative DTW (right) [adapted from 28]

4.3 Proposed Improved Features for Dynamic Time Warping (IFDTW)

Several refinements exist in the literature for improving the performance of DTW algorithm [29, 30, 31]. In this work, we propose improved features for dynamic time warping instead of using absolute feature value or derivative estimates since an absolute value or local feature is not sufficient to identify and match common trends and patterns in the feature vectors. We use both local and global features of each data point to track more accurately their contribution towards pattern matching. For MFCC feature vector $\boldsymbol{X}$, we compute the local and global features as:

$$f_{\text{local}}(x_i) = x_i - \left(\frac{x_{i+1} - x_{i-1}}{2} \right)$$

$$f_{\text{global}}(x_i) = x_i - \left(\frac{\sum_{k=i+1}^{K} \frac{x_k}{K-i} - \sum_{k=1}^{i-1} \frac{x_k}{i-1}}{2} \right) \qquad (15)$$

The local feature is simply the feature value minus the slope of the line through its left and right neighbors. The global feature should reflect the position of a feature value in the global shape of the feature sequence. The derivative features in (14) contain no global information and so we propose a global feature which is computed as the feature value minus the difference between the average values of the last *K-i* points and the average value of the first *i-1* points in the feature vector $\boldsymbol{X}$. Similarly, local and global features are computed for the feature vector $\boldsymbol{Y}$. These local and global features are not defined for the first and last vectors of the sequences.

Based on the local and global features, we compute the distance between feature vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$ as follows:

$$dist(x_i, y_j) = dist_{\text{local}}(x_i, y_j) + dist_{\text{global}}(x_i, y_j)$$

$$dist_{\text{local}}(x_i, y_j) = \left| f_{\text{local}}(x_i) - f_{\text{local}}(y_j) \right| \qquad (16)$$

$$dist_{\text{global}}(x_i, y_j) = \left| f_{\text{global}}(x_i) - f_{\text{global}}(y_j) \right|$$

Note that the time complexity of the proposed IFDTW is same as that of conventional DTW and derivative DTW, i.e. $O(N^2)$. Fig. 10 illustrates the alignment between two feature vectors for the same digit spoken by two different speakers where we can observe the advantage of the proposed IFDTW technique since it has fewer singularities compared to conventional DTW and derivative DTW.

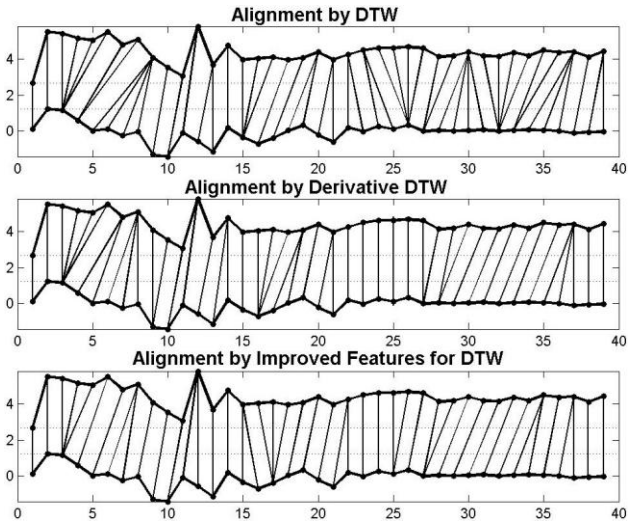

Fig. 10: Alignment by various DTW techniques with minimum distance by DTW: 116, DDTW: 108, IFDTW: 94

In this work, we also modify the adjustment window constraint by considering the different slopes of Itakura parallelogram. Itakura proposed the window constraint using slope 2 and 1/2 [27], however, our experiments indicate that using constrained slope of 3 and 1/3 gives the minimum matching distance using IFDTW. Table I illustrates the minimum matching distance for same digits spoken by

different speakers corresponding to different slopes of the parallelogram.

TABLE I. Slopes for Adjustment Window Constraint

	Matching distance for digit		
Slope	**"2"**	**"4"**	**"9"**
5 ~ 1/5	138	132	123
4 ~ 1/4	126	125	106
3 ~ 1/3	**95**	**102**	**82**
2 ~ 1/2	112	118	97

The matching distance is minimum at the constrained path slope of 3 and 1/3 and accordingly, our adjustment window constraint is the shaded region as shown in Fig. 11. *R* is the width of the Sakoe-Chiba band which is typically 10% of max(|***X***|, |***Y***|) [11].

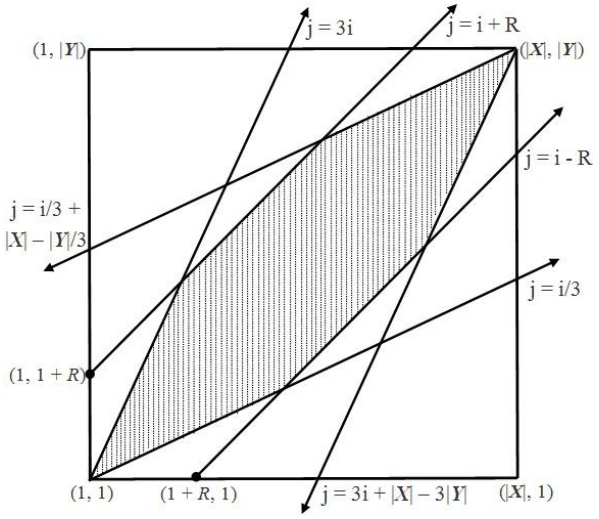

Fig. 11: Adjustment window constraints for IFDTW

5. EXPERIMENTAL RESULTS

Conventional speech recognition systems consist of feature extraction based on MFCC followed by feature recognition using DTW algorithm. We test the effectiveness of our proposed Weighted MFCC (WMFCC) features and Improved Features for DTW (IFDTW) algorithm for speaker-independent isolated spoken digits 0 to 9. The entire recognition system is implemented using MATLAB. The training and test speech data are taken from TI-Digits database [32] from which we have used samples from 10 male and 10 female speakers. Each digit is spoken twice by each speaker and total 400 utterances are collected in our experiments. We have used 240 utterances (60%) for training and 160 utterances (40%) for testing. In all our experiments, the speech signals are sampled at 16 kHz and represented by 16 bits. The speech signal is divided into frames of duration 25 ms with 10 ms overlap between adjacent frames. The number of Mel filters used for feature extraction is 40 and 512-point FFT is used for WMFCC feature extraction stage.

Fig. 12 shows the recognition accuracy by the proposed techniques for each digit, Table II shows the confusion matrix for 16 test utterances of each digit from 0 to 9 and Table III compares the recognition accuracy obtained by various techniques.

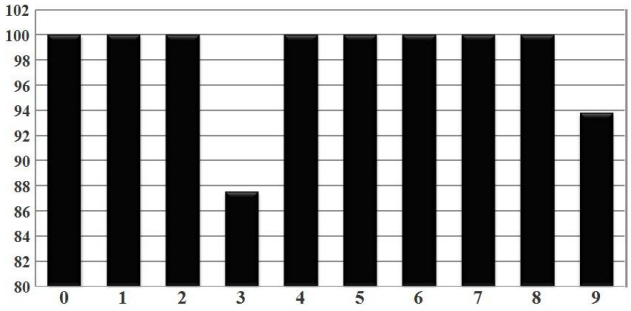

Fig. 12: Recognition accuracy of different spoken digits using WMFCC and IFDTW

TABLE II. Confusion Matrix

	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
0	**16**	0	0	0	0	0	0	0	0	0
1	0	**16**	0	0	0	0	0	0	0	0
2	0	0	**16**	0	0	0	0	0	0	0
3	0	0	0	**14**	0	0	0	0	**2**	0
4	0	0	0	0	**16**	0	0	0	0	0
5	0	0	0	0	0	**16**	0	0	0	0
6	0	0	0	0	0	0	**16**	0	0	0
7	0	0	0	0	0	0	0	**16**	0	0
8	0	0	0	0	0	0	0	0	**16**	0
9	0	**1**	0	0	0	0	0	0	0	**15**

TABLE III. Overall Recognition Accuracy (%)

	#Features	**DTW**	**DDTW**	**IFDTW**
MFCC	13	86.68	88.95	90.72
MFCC + Delta	26	92.65	94.28	95.40
MFCC + Delta + Double Delta	39	94.25	96.75	97.42
Weighted MFCC	**13**	95.30	96.15	**98.13**

The above results indicate that the proposed WMFCC feature extraction and IFDTW feature recognition techniques are superior to the existing techniques for isolated spoken digits recognition. With a smaller number of cepstral coefficients by taking into account both delta and acceleration coefficients, WMFCC surpass the recognition accuracy relative to conventional MFCC and also have the benefit of low computational overhead to the recognition stage. In all cases, using the Improved Features for DTW gives us more accuracy compared to conventional DTW and derivative DTW.

6. CONCLUSION

In this paper, we proposed an improved technique for feature extraction using Weighted MFCC that considers both the voiceprint and the dynamic characteristics of the spoken digit, and an enhanced technique for feature recognition using Improved Features for DTW (IFDTW). The experimental results demonstrate that the recognition system with WMFCC can achieve higher recognition rate than the systems using MFCC and its delta and double delta coefficients, thus leading to a lower computational overhead on the DTW algorithm. Also, by improving the WMFCC features considering their local and global trend over the entire spoken speech signal in IFDTW, we achieve higher accuracy compared to using the conventional DTW (pure value based) and derivative DTW

(local feature based) algorithms. Our future focus will be to improve the recognition speed of IFDTW without compromising its recognition accuracy.

7. REFERENCES

[1] R. Cox, C. Kamm, L. Rabiner, J. Schroeter, and J. Wilpon, "Speech and language processing for next-millennium communications services", *Proc. of the IEEE*, vol. 88, no. 8, Aug 2000

[2] D. Jurafsky, and J. Martin, *Speech and Language Processing*, Prentice Hall, 2000

[3] J. Tierney, "A study of LPC analysis of speech in additive noise", *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 28, no. 4, pp. 389-397, 1980

[4] A. Paul, D. Das, and M. Kamal, "Bangla speech recognition system using LPC and ANN", *7th Intl. Conf. Advances in Pattern Recognition*, 2009

[5] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, 1993

[6] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 28, no. 4, pp. 357-366, Aug 1980

[7] A. Mishra, M. Chandra, A. Biswas, and S. Sharan, "Robust features for connected Hindi digits recognition", *Intl. Journal of Signal Processing, Image Processing and Pattern Recognition*, vol. 4, no. 2, pp. 79-90, June 2011

[8] Z.Jun, S. Kwong, W. Gang, and Q. Hong, "Using Mel-frequency cepstral coefficients in missing data technique", *EURASIP Journal on Applied Signal Processing*, vol. 2004, no. 3, pp. 340-346, 2004

[9] O. W. Kwon, K. Chan, J. Hao, and T. W. Lee, "Emotion recognition by speech signals", in *Proc.8th European Conf. Speech Communication and Technology*, pp. 125-128, Geneva, Switzerland, 2003

[10] L. Rabiner, B. Juang, S. Levinson, and M. Sondhi, "Recognition of isolated digits using hidden markov models with continuous mixture densities", *AT&T Tech. Journal*, 64(6), 1985

[11] H. Sakoe, and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition", *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. ASSP-26, 1978

[12] L. Rabiner, A. Rosenberg, and S. Levinson, "Considerations in dynamic time warping algorithms for discrete word recognition", *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. ASSP-26, 1978

[13] W. Fu, X. Yang, and Y. Wang, "Heart sound diagnosis based on DTW and MFCC", *3rd IEEE Intl. Congress on Image and Signal Processing*, pp. 2920-2923, Oct 2010

[14] F. Yu, E. Chang, Y. Xu, and H. Shum, "Emotion detection from speech to enrich multimedia content", in *Proc.2nd IEEE Pacific Rim Conf. Multimedia*, pp. 550-557, Beijing, China, 2001

[15] S. Singh, and E. Rajan, "Vector Quantization approach for speaker recognition using MFCC and inverted MFCC", *International Journal of Computer Applications*, vol. 17, no. 1, Mar 2011

[16] R. Tato, R. Santos, R. Kompe, and J. Pardo, "Emotional space improves emotion recognition", in *Proc. 7th Intl. Conf. Spoken Language Processing*, vol. 3, pp. 2029-2032, Denver, USA, 2002

[17] L. Rabiner, and M. Sambur, "An algorithm for determining the endpoints of isolated utterances", *Bell System Technical Journal*, vol. 54, no. 2, pp. 297-315, Feb 1975

[18] J. Picone, "Signal modeling techniques in speech recognition", *Proc. of the IEEE*, vol. 81, no. 9, Sep 1993

[19] J. Deller, J. Proakis, and J. Hansen, *Discrete Time Processing of Speech Signals*, Prentice Hall, NJ, USA, 1993

[20] S. Kopparapu, and M. Laxminarayana, "Choice of Mel filter bank in computing MFCC of a resampled speech", *Proc. IEEE Intl. Conf. Information Sciences Signal Processing and their Applications*, pp. 121-124, May 2010

[21] G. Bekesy, *Experiments in Hearing*, Mc-Graw Hill, New York, 1960

[22] H. Hassanein, and M. Rudko, "On the use of Discrete Cosine Transform in cepstral analysis", *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 32, no. 4, pp. 922-925, 1984

[23] B. Juang, L. Rabiner, and J. Wilpon, "On the use of bandpass liftering in speech recognition", *IEEE Intl. Conf. Acoustics, Speech, and Signal Processing*, pp. 765-768, Apr 1986

[24] W. Hong, P. Jingui, "Modified MFCCs for robust speaker recognition", *IEEE Intl. Conf. Intelligent Computing and Intelligent Systems*, pp. 276-279, Oct 2010

[25] W. Junqin, and Y. Junjun, "An improved arithmetic of MFCC in speech recognition system", *IEEE Intl. Conf. Electronics, Communications and Control*, pp. 719-722, China, Sep 2011

[26] S. Ong, and C. Yang, "A comparative study of text-independent speaker identification using statistical features", *Intl. Journal on Computer Engineering Management*, vol. 6, no. 1, 1998

[27] F. Itakura, "Minimum prediction residual principle applied to speech recognition", *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. ASSP-23, pp. 52-72, 1975

[28] E. Keogh, and M. Pazzani, "Derivative dynamic time warping", *Proc. of the 1st SIAM Intl. Conf. Data Mining*, Chicago, USA, 2001

[29] S. Salvador, and P. Chan, "FastDTW: toward accurate dynamic time warping in linear time and space", *Proc. of 3rd KDD Workshop on Mining Temporal and Sequential Data*, pp. 70-80, 2004

[30] L. Yan-Sheng, and J. Chang-Peng, "Research on improved algorithm of DTW in speech recognition", *IEEE Intl. Conf. Computer Application and System Modeling*, pp. 418-421, Oct 2010

[31] K. Chanwoo, and S. Kwang-deok, "Robust DTW-based recognition algorithm for hand-held consumer devices", *IEEE Intl. Conf. Consumer Electronics*, pp. 433-434, Jan 2005

[32] R. Leonard, "A database for speaker-independent digit recognition", *IEEE Intl. Conf. Acoustics, Speech, and Signal Processing*, pp. 328-331, Mar 1984