Emotion Recognition from Speech using Teager based DSCC features

Santosh V. Chapaneri, Deepak J. Jayaswal, Ph. D
Department of Electronics and Telecommunication Engineering,
St. Francis Institute of Technology,
University of Mumbai

ABSTRACT
Emotion recognition from speech has emerged as an important research area in the recent past. The purpose of speech emotion recognition system is to automatically classify speaker's utterances into seven emotional states including anger, boredom, disgust, fear, happiness, sadness and neutral. The speech samples are from Berlin emotional database and the features extracted from these utterances are Teager-based delta-spectral cepstral coefficients (T-DSCC) which are shown to perform better than MFCC. Dynamic Time Warping (DTW) and its variant Improved Features for DTW (IFDTW) is used as a classifier to classify different emotional states. Unlike in conventional DTW, we do not use the minimum distance for classification. Rather, the median distance criterion is employed for improved emotion classification. The proposed emotion recognition system gives an overall classification accuracy of 97.52%.

General Terms
Speech Processing, Emotion Recognition

Keywords
Emotion recognition, MFCC, DSCC, Teager Energy Operator, Dynamic time warping

1. INTRODUCTION
Human emotion recognition is an important component for efficient human-computer interaction and has become a major research topic due to its many potential applications. It plays a critical role in communication, allowing people to express oneself beyond the verbal domain. It is being applied to growing number of areas such as humanoid robots, call centers, car industry, mobile communication, computer tutorials via virtual avatars, etc. [1]. The term emotion describes the subjective feelings in short periods of time which are related to events, objects, or persons. Since the emotional state of humans is a highly subjective experience, it is hard to find objective and universal definitions. This is the reason why there are different approaches to model emotions in literature.

Emotion recognition is a statistical pattern classification problem. It consists of two major steps, feature extraction and classification. While the theory of pattern classification is well-developed [2], the extraction of features for emotion recognition is a highly empirical issue and depends on the specific application and database. Various speech features containing emotion information are found in the literature such as energy, pitch frequency [3], formant frequency [4], Linear Prediction Coefficients, Linear Prediction Cepstrum Coefficients [5], Mel-Frequency Cepstrum Coefficients [6]. In [7], harmony features based on psychoacoustic harmony perception known from music theory are employed as features. In [8], modulation spectral features are used by using an auditory filter-bank and a modulation filter-bank for speech analysis. In [9], both linguistic and acoustic features are used for anger classification. To reduce the size of feature set and selecting the most relevant subset of features in emotion recognition system, following techniques have been employed in the recent years: Fisher's linear discriminant analysis [10], forward and backward feature selection [5], fast correlation-based filter [11], and sequential floating forward selection [12]. Also, the following methods have been used for emotion classification in the recent years: dynamic time warping (DTW) [13], Bayesian networks [14], Hidden Markov Models (HMM) [15], support vector machines (SVM) [15], artificial neural networks (ANN) [16], Gaussian Mixture Models (GMM) [17], k-nearest neighbor (KNN) [15], decision trees [15], and hybrid approaches [18].

This paper is organized as follows. Section 2 explains the Berlin emotion database used to train and test the proposed system followed by Section 3 which reviews the conventional MFCC feature extraction, DSCC features and the proposed Teager-based DSCC (T-DSCC) features. Section 4 explains the feature recognition technique using conventional DTW, Improved Features for DTW (IFDTW) algorithm and its modification using a median distance. Section 5 demonstrates the experimental results followed by conclusions in Section 6.

2. SPEECH DATABASE
In this paper, the Berlin Emotion Database (EMO-DB) [19] is used as a database for the experiments, which contains 535 utterances, as shown in Table I, of 10 professional native German-speaking actors (5 male, 5 female) simulating utterances which could be used in everyday communication and are interpretable in all applied emotions. The actors were advised to read pre-defined sentences in the targeted seven emotions of anger (*Wut*), boredom (*Langeweile*), disgust (*Ekel*), fear (*Angst*), happiness (*Freude*), sadness (*Trauer*), and a neutral emotional state. The length of the utterances varies from 2 to 8 seconds. The recordings were taken in an anechoic chamber with high-quality recording equipment at a sampling rate of 16 kHz with a 16-bit resolution and a mono channel. 70% of the utterances were used for training and the remaining for testing.

Table I. Number of Utterances in EMO-DB

Anger	Boredom	Disgust	Fear	Happiness	Sadness	Neutral
127	81	46	69	71	62	79

3. FEATURE EXTRACTION

3.1 Conventional MFCC

Mel Frequency Cepstral Coefficients (MFCC) features are widely used for speech recognition as well as emotion recognition obtaining a good recognition rate. MFCC is based on the characteristics of human ear's hearing, which uses a non-linear frequency unit to simulate the human auditory system. The detailed procedure of computing MFCC features for each speech utterance is explained in [20]. Pre-processing is performed on the speech signal for end-point detection, followed by framing and windowing since speech is a quasi-stationary random process. Pre-emphasis is performed on each frame to flatten the spectrum of the speech signal [21]. Spectral coefficients are then computed using FFT which are then filtered by a Mel-scale non-linear filter. The cepstral coefficients are then computed as the inverse Fourier transform of the log of the resulting coefficients. Since the log Mel filter bank coefficients are real and symmetric, the inverse Fourier transform operation is replaced by DCT to generate the cepstral coefficients [22]. Typically, only the first 13 cepstral coefficients are used since MFCC in the low frequency region has a good frequency resolution, and the high frequency coefficients do not obtain satisfactory accuracy. Mean Variance Normalization (MVN) is performed to eliminate the acoustic difference from the features other than emotion. Further, delta and double-delta features are also computed for each frame to recover the trend information in the frame-by-frame analysis. Thus, 39 MFCC features are obtained for each frame of each speech utterance, as illustrated in Fig. 1a.

3.2 DSCC

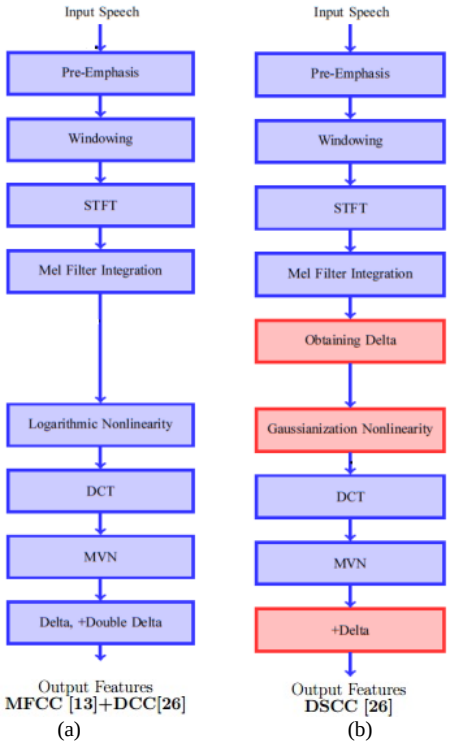

Fig. 1: (a) 13-dimensional MFCC features + 26-dimensional delta-cepstral coefficients (DCC), (b) 26-dimensional delta-spectral cepstral coefficients (DSCC) features [23]

The speech data can typically vary widely in recording quality. These data may have been recorded through different types of channels, such as a cell phone, and a room microphone. Additionally, the data may contain different levels and types of additive noise, such as white noise, babble noise, and music. Finally, speech may also be recorded in different acoustic environments with different impulse responses. Therefore, an ideal recognition system would need to be robust to channel effects, noise, and reverberation. In [23], a novel set of features were proposed for more robust recognition. This set of features, called delta-spectral cepstral coefficients (DSCC), was sought to improve recognition accuracy via performing the first delta operation in the spectral domain rather than the cepstral domain. It was shown that DSCC features were more robust to noise and reverberation when used in conjunction with MFCC [23].

The major changes compared to MFCC are that the initial time-differencing operation is now moved earlier in the processing and a new Gaussianization stage is added. Specifically, performing the delta operation in the spectral domain enhances the fast changing speech components, and suppresses the slowly-changing noisy components. The raw delta-spectral cepstral coefficients are highly non-Gaussian as observed in Fig. 2a. To adapt DSCC for emotion recognition, histogram normalization is applied to the delta-spectral features to give them a Gaussian distribution as illustrated in Fig. 2b. The DCT operation compresses the 40-dimensional delta-spectral features to a 13-dimensional vector of delta-spectral cepstral coefficients (DSCC). Double-delta features are then derived from the delta-spectral features in the cepstral domain.

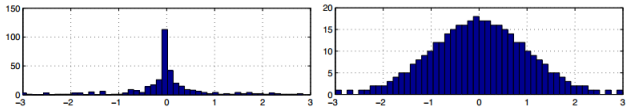

Fig. 2: Histogram of short-time power after the delta operation for a clean-speech sample (a) before and (b) after Gaussianization [23]

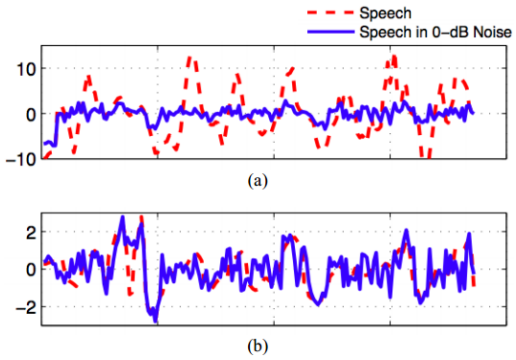

Fig. 3: Short time power plots of a single Mel-channel for (a) temporal difference over the logarithmic power of a speech signal (reflecting DCC), (b) Gaussianization operation over temporal difference of a speech signal (reflecting DSCC) [23]

The short-time power plots are illustrated in Fig. 3. The lines representing the clean and noisy speech signals are more similar in the plot representing DSCCs (Fig. 3b); this similarity suggests that DSCCs should be more robust features in noise than DCCs, possibly due to the Gaussianization nonlinearity which replaces

the logarithmic nonlinearity. Also, the DSCC features completely ignore the static-spectral contents, deriving their features instead entirely from the dynamic-spectral contents. The dynamic features in the DSCC features are not only good for emotion recognition but they are also very robust to additive noise.

3.3 Proposed Teager-based DSCC (T-DSCC)

The majority of studies in the field of speech emotion recognition have concentrated on the features derived from a linear speech production model which assume that airflow propagates in the vocal tract as a plane wave. This pulsatile flow is considered the source of sound production. According to studies by Teager [24], however, this assumption may not hold since the flow is actually separate and concomitant vortices are distributed throughout the vocal tract. Teager suggested that the true source of sound production is actually the vortex-flow interactions, which are non-linear. This observation was supported by the theory in fluid mechanics as well as by numerical simulation of Navier–Stokes equation [25]. Therefore, non-linear speech features are necessary for classification of different emotional status. In an effort to reflect the instantaneous energy of nonlinear vortex-flow interactions, Teager developed an energy operator, with the supporting observation that hearing is the process of detecting the energy. The simple and elegant form of the operator was introduced by Kaiser [26] as:

$$\psi\left[x(t)\right]=\left[\frac{d}{dt}x(t)\right]-x(t)\left[\frac{d^2}{d^2t}x(t)\right] \quad (1)$$

where ψ is Teager energy operator (TEO) and $x(t)$ is single component of the continuous speech signal. Since speech is represented in discrete form in most speech processing systems, Kaiser derived the operator for discrete-time sampled speech signal $x(n)$ from its continuous form as:

$$\psi\left[x(n)\right]=x^2(n)-x(n+1)x(n-1) \quad (2)$$

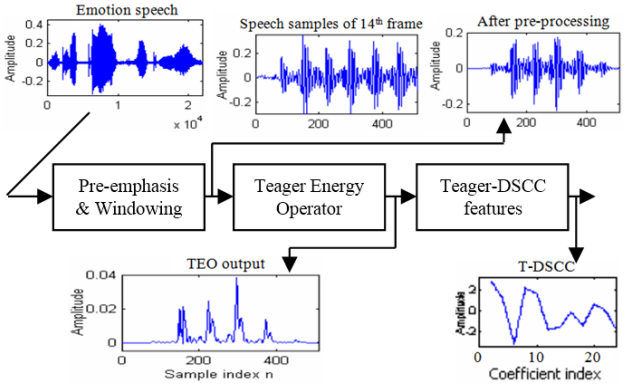

Fig. 4: Proposed Teager-based DSCC feature extraction

As a powerful nonlinear operator, TEO gives a remarkable performance in the field of background noise suppression and signal feature extraction. Thus, we propose to use TEO in the process of feature extraction to eliminate the effect of noise. The proposed system is illustrated in Fig. 4 where TEO is applied after pre-processing. After TEO, the usual steps of DSCC are performed as explained in Sec. 3.2. We obtain the resulting 26-dimensional T-DSCC features for each frame of each speech utterance of the emotion database.

4. FEATURE CLASSIFICATION

4.1 Conventional Dynamic Time Warping

Dynamic Time Warping (DTW) is an algorithm adopted by the speech recognition community to handle the matching of non-linearly expanded or contracted signals [27].Unlike Linear Time Warping (LTW) which compares two time series based on linear mapping of the two temporal dimensions, DTW allows a non-linear warping alignment of one signal to another by minimizing the distance between the two as shown in Fig. 5. DTW is a common technique for comparing time series by searching for optimal alignments using dynamic programming, described in terms of optimal *warp paths*. This warping between two signals can be used to determine the similarity between them and thus it is very useful for feature classification.

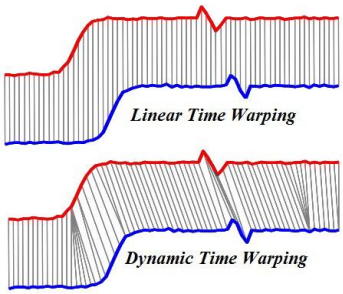

Fig. 5: DTW non-linear alignment of two time series

The classical DTW finds the optimal alignment between two one-dimensional sequences $X = \{x_1, x_2,\ldots x_N\}$ and $Y = \{y_1, y_2,\ldots y_M\}$ of length N and M respectively, in which one sequence is non-linearly spanned or shrunk in its time axis with respect to another sequence. The algorithm works by building an N x M cost matrix in which each element $d(i, j)$ corresponds to the pairwise local distance computed using the Euclidean distance. The local distance measure can either be the Euclidean distance or the Mahalanobis distance of which the latter incorporates a covariance matrix in the computation. This covariance matrix has to be estimated from a general statistical model of the features in the application area. Each element of the cost matrix defines an alignment between X and Y sequence and is computed using a recursive formula:

$$D(i, j) = d(i, j) + \min[D(i-1, j), D(i-1, j-1), D(i, j-1)] \quad (3)$$

$D(1, 1)$ is initialized to $d(1, 1)$. The alignment that results in the minimum distance between the two sequences has the value $D(M, N)$. The warping path must satisfy the conditions of monotonicity, continuity, boundary and slope constraints [28]. There is also 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 band of size R, directly above and to the right of the diagonal. Fig. 6 illustrates the window bands widely used in DTW.

In application to emotion recognition, the two time series corresponds to the two *numCoefficients* by *numFrames* T-DSCC feature vectors of different emotion speech utterances. A two-dimensional cost matrix is computed that stores the minimum

distance between two feature vectors *X* and *Y*. The test emotion signal's feature vector is compared to the reference feature vectors' using Median IFDTW (discussed in Section 4.3) and the one closest to the reference is chosen as the classification output.

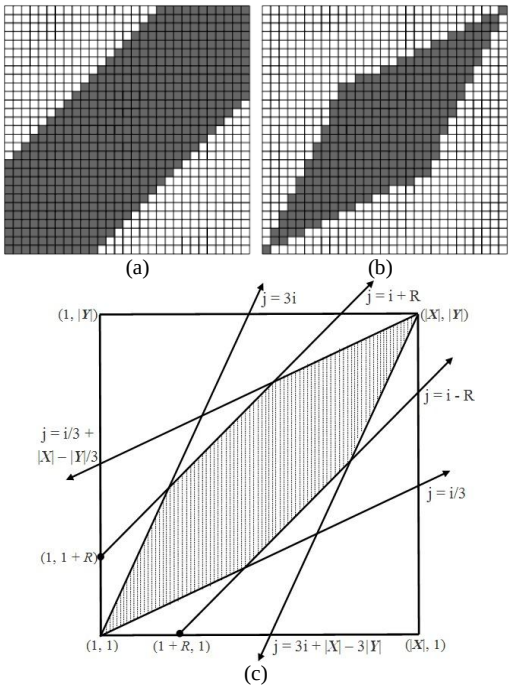

Fig. 6: Adjustment window constraints: (a) Sakoe-Chiba band [28], (b) Itakura Parallelogram [29], (c) IFDTW [20]

4.2 IFDTW

Several modifications of conventional DTW are found in the literature [30] since its fundamental flaw 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. Improved Features for DTW (IFDTW) technique was proposed in [20] where instead of using absolute feature value or derivative estimates, modified features are used since an absolute value or local feature is not sufficient to identify and match common trends and patterns in the feature vectors. Both local and global features of each data point are used to track more accurately their contribution towards pattern matching. Further, to reduce the computational complexity of IFDTW from O(N^2) to O(N), FIFDTW technique was proposed in [30] which uses a fast DTW algorithm [31] using the concepts of constraints and data abstraction approaches. The optimal warping path is determined through coarsening, projection and refinement stages as illustrated in Fig. 7, and this approach speeds-up the DTW computation significantly.

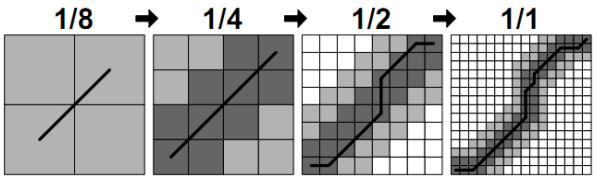

Fig. 7: Refinement of optimal warping path [31]

4.3 Proposed Median IFDTW

In this work, we propose the Median IFDTW technique to determine the emotion reference feature vector closest to the test emotion feature vector. The conventional DTW as well as other modifications and IFDTW algorithm recognize the test emotion/speech using the minimum Euclidean distance approach. However, consider the analysis of incorrect classification of "Happiness" emotion with "Anger" emotion using conventional DTW. Fig. 8 shows the distributions of distances between test emotion "Happiness" and reference emotions "Anger" and "Happiness". From Fig. 8a, we observe that the reference emotion of "Anger" yields the minimum distance of 91 which is thus misclassified with the test emotion. But using the median distance of 140 from Fig. 8b, the test emotion is correctly classified as "Happiness". Thus, we propose to use the median distance for feature classification using IFDTW.

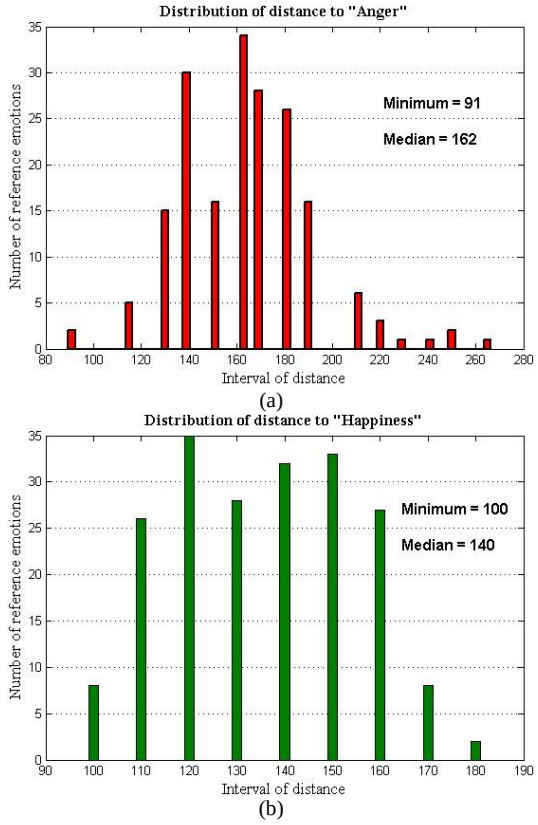

Fig. 8: Distribution of distances between feature vectors of test emotion "Happiness" and reference emotions "Anger" and "Happiness"

We assume that there are *m* reference emotions. Each reference emotion has *n* utterances from different people (as in EMO-DB). The IFDTW distance that the test emotion feature vector computes with the feature vector of i^{th} reference emotion is denoted as D_i ($1 \le i \le m$). The IFDTW distance that the test emotion feature vector computes with the j^{th} utterance of the i^{th} reference emotion is denoted as d_{ij} ($1 \le j \le n$). Thus, $D_i = [d_{i1}, d_{i2}, d_{i3},\ldots, d_{ij},\ldots, d_{in}]$. All IFDTW distances for the test emotion feature vector are denoted as $D = [D_1\ D_2 \ldots D_i \ldots D_m]^T$. We sort all distances of every reference emotion from small to large to get the new distance series D_i' as:

$$D_i' = \underset{(1\le i\le m)}{Sort}(D_i) = \left[d_{i1}' \; d_{i2}' \cdots d_{in}' \right] \quad (4)$$

Thus, d_{i1}' is the minimum distance and d_{in}' is the maximum distance among D_i'. We obtain the ordered distance matrix D' = [D_1' D_2' … D_i'… D_m']^T. The median distance a_i is then extracted from D_i', i.e. a_i = *Median*(D_i'), and the corresponding median distance vector is formed as $A = [a_1\ a_2 \ldots a_i \ldots a_m]$. The resulting distance for the test emotion feature vector is thus obtained as the minimum value of vector A, i.e. *Min*(A). Note that for the conventional DTW, the test emotion feature vector would be classified by *Min*([d_{11}' d_{21}' … d_{m1}']).

5. EXPERIMENTAL RESULTS

Conventional emotion recognition systems consist of feature extraction based on MFCC followed by feature recognition using DTW algorithm. We test the effectiveness of our proposed Teager-based DSCC (T-DSCC) feature extraction algorithm and Median IFDTW feature classification algorithm for the emotion utterances of EMO-DB [19] by adding noise with SNR of 10 dB in the original speech samples. From the emotion database, 70% of the utterances are used for training and 30% are used for testing, i.e. 161 utterances for testing with 23 utterances per emotion. In T-DSCC algorithm, the emotion 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. Table II shows the confusion matrix for each test emotion and Table III compares the overall recognition accuracy obtained by various algorithms.

Table II. Confusion Matrix (23 utterances/emotion)
(A: Anger, B: Boredom, D: Disgust, F: Fear, H: Happiness, S: Sadness, N: Neutral)

	A	**B**	**D**	**F**	**H**	**S**	**N**
A	**23**	0	0	0	0	0	0
B	0	**23**	0	0	0	0	0
D	0	0	**22**	0	0	0	1
F	0	0	0	**23**	0	0	0
H	0	0	0	0	**23**	0	0
S	0	1	0	1	0	**21**	0
N	0	1	0	0	0	0	**22**

Table III. Overall Recognition Accuracy (%)

	#Features	**DTW**	**IFDTW**	**Median IFDTW**
MFCC + Δ + ΔΔ	39	84.52	87.39	91.29
DSCC	26	93.82	95.14	96.73
T-DSCC	**26**	94.18	95.69	**97.52**

The above results demonstrate that the non-linear features based on TEO using DSCC are effective and have optimal emotion recognition capacity in the presence of noise. In comparison with MFCC, the recognition rates were increased significantly by using DSCC and Teager-based DSCC (T-DSCC) features. This is due to the fact that MFCC is known to be developed to mimic human perception process and since the problem of emotion recognition deals with identification of perceptually similar emotions, MFCC gets confused in discriminating the emotion-specific features. On the other hand, T-DSCC represents the combined effect of airflow properties in the vocal tract (which are known to be language and speaker dependent [32]) and human perception process. So, T-DSCC is able to capture the emotion-specific information better than MFCC and has better *class discrimination* power than MFCC.

There is also a significant improvement in performance of T-DSCC due to the median distance based IFDTW classification algorithm. It can be also inferred from Table II that anger, boredom, fear and happiness are the best emotions to be recognized but disgust, sadness and neutral emotions give confusing results in few cases.

6. CONCLUSION

The effort has been done through this work to explore the Teager-based DSCC features for recognizing the emotions using speech utterances. We have evaluated the proposed emotion speech recognition system on acted Berlin emotion database by simulating additive noise. The experimental results demonstrate that the recognition system with T-DSCC can achieve higher recognition rate than the systems using MFCC and DSCC. Also, by using the median distance as the criteria in IFDTW algorithm, we achieve higher classification accuracy of 97.52% compared to using the conventional DTW and its variants.

The results reveal that recognition rate of some emotions, including disgust, sadness and neutral still needs to be further improved. Further research should focus on the following aspects: first, the combination of special emotion information should be paid attention to, such as the fundamental frequency rise in the end of surprising sentence, the shaking sound of fear, etc. Second, fuzzy theory can be used to find the probability of some kind of emotions. Third, the trend of emotion recognition is not clearly known in the case of many other languages. It would be helpful to evaluate the proposed and established features on different Indian languages for emotion recognition. This will help to decide whether the methods and features used in literature are language independent or not.

7. REFERENCES

[1] M. Ayadi, M. Kamel, and F. Karray, "Survey on Speech Emotion Recognition: Features, Classification Schemes, and Databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572-587, Mar. 2011

[2] R. Duda, P. Hart, and D. Stork, *Pattern Classification*, 2nd ed., Wiley, New York, 2001

[3] D. Ververidis, C. Kotropoulos, and I. Pitas, "Automatic emotional speech classification", *IEEE Intl. Conf. Acoustics, Speech and Signal Processing*, vol.1, pp. 593-596, Montreal, May 2004

[4] Z. Xiao, E. Dellandrea, W. Dou, and L. Chen, "Features extraction and selection for emotional speech classification", *IEEE Conference on Advanced Video and Signal Based Surveillance*, pp. 411-416, Sep 2005

[5] T. Pao, Y. Chen, J. Yeh, and Y. Chang, "Emotion recognition and evaluation of Mandarin speech using weighted D-KNN classification", *17th Conf. on Computational Linguistics and Speech Processing*, pp.203-212, Sep 2005

[6] T. Pao, Y. Chen, J. Yeh, and P. Li, "Mandarin emotional speech recognition based on SVM and NN", *Proc. 18th Intl. Conf. on Pattern Recognition* (ICPR'06), vol.1, pp. 1096-1100, Sep 2006

[7] B. Yang, and M. Lugger, "Emotion recognition from speech signals using new harmony features", *Journal of Signal Processing* vol. 90, no. 5, pp. 1415-1423, 2010

[8] S. Wu, T. Falk, and W. Chan, "Automatic speech emotion recognition using modulation spectral features", *Speech Communication*, vol. 53, no. 5, pp. 768-785, 2011

[9] T. Polzehl, A. Schmitt, F. Metze, and M. Wagner, "Anger recognition in speech using acoustic and linguistic cues", *Speech Communication*, vol. 53, pp. 1198-1209, 2011

[10] S. Haq, P. Jackson, and J. Edge, "Audio-visual feature selection and reduction for emotion classification", *Proc. Intl. Conf. on Auditory Visual Speech Processing*, pp. 185-190, 2008

[11] D. Gharavian, M. Sheikhan, A. Nazerieh, and S. Garoucy, "Speech emotion recognition using FCBF feature selection method and GA-optimized fuzzy ARTMAP neural network", *Neural Computing and Applications, Springer*, vol. 12, no. 8, pp. 2115-2126, Nov 2012

[12] A. Batliner, S. Steidl, B. Schuller, D. Seppi, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, V. Aharonson, L. Kessous, and N. Amir, "Whodunnit - Searching for the most important feature types signalling emotion-related user states in speech", *ACM Journal of Computer Speech and Language*, vol .25, no. 1, pp.4-28, Jan 2011

[13] M. Krishna, P. Lakshmi, Y. Srinivas, and S. Devi, "Emotion recognition using dynamic time warping technique for isolated words", *Intl. Journal Computer Science Issues*, vol. 8, no. 5, pp. 306-309, Sep 2011

[14] E. Fersini, E. Messina, and F. Archetti, "Emotional states in judicial courtrooms: an experimental investigation", *Speech Communication*, vol. 54, no. 1, pp. 11-22, Jan 2012

[15] E. Ayadi, M. Kamel, and F. Karray, "Survey on speech emotion recognition: features, classification schemes, and databases", *Pattern Recognition*, vol. 44, no. 3, pp. 572-587, Mar 2011

[16] J. Wang, Z. Han, and S. Lung, "Speech emotion recognition system based on genetic algorithm and neural network," *IEEE Intl. Conf. Image Analysis and Signal Processing*, pp.578-582, Oct 2011

[17] M. Kockmann, L. Burget, and J. Cˇernocky, "Application of speaker and language identification state-of-the-art techniques for emotion recognition", *Speech Communication*, vol. 53, no. 10, pp. 1172-1185, Nov 2011

[18] R. Lo´pez, J. Silovsky, and M. Kroul, "Enhancement of emotion detection in spoken dialogue systems by combining several information sources", *Speech Communication*, vol. 53, no. 10, pp. 1210-1228, Nov 2011

[19] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of German emotional speech", *Proc. Interspeech-2005*, Lisbon, Portugal, pp. 1-4, Jan 2005

[20] S. Chapaneri, "Spoken digits recognition using weighted MFCC and improved features for dynamic time warping", *Intl. Journal Computer Applications*, vol. 40, no. 3, pp. 6-12, Feb 2012

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

[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] K. Kumar, C. Kim, and R. Stern, "Delta-Spectral Cepstral Coefficients for robust speech recognition", *IEEE Intl. Conf. Acoustics, Speech, and Signal Processing*, pp. 4784-4787, May 2011

[24] H. Teager, and S. Teager, "A phenomenological model for vowel production in the vocal tract," *Speech Science: Recent Advances*, pp.73-109, 1985

[25] T. Thomas, "A finite element model of fluid flow in the vocal tract", *Journal of Computer Speech and Language*, vol. 1, no. 2, pp. 131-151, Dec 1986

[26] J. Kaiser, "On a simple algorithm to calculate the 'energy' of a signal", *IEEE Intl. Conf. Acoustics, Speech and Signal Processing*, vol. 1, pp. 381-384, Apr 1990

[27] S. Theodoridis, and K. Koutroumbas, *Pattern Recognition*, 4th Ed., Academic Press, 2008

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

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

[30] S. Chapaneri, and D. Jayaswal, "Efficient speech recognition system for isolated digits", *Intl. Journal Computer Science and Engineering Technologies*, vol. 4, no. 3, pp. 228-236, Mar 2013

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

[32] H. Patil, and T. Basu, "Detection of bilingual twins by Teager energy based features," *Proc. Intl. Conf. Signal Processing and Communication*, pp. 32-36, Dec 2004