Proceedings of the Institute of Acoustics

ACOUSTIC CLASSIFICATION USING TIME-FREQUENCY DISTRIBUTIONS

H. Marvi
I. Paraskevas
E. Chilton

Centre for Vision Speech and Signal Processing
University of Surrey
Guildford, Surrey, UK

1 INTRODUCTION

Due to the increasing demand for both speech and audio classification a new method is introduced for the effective recognition of both artificial and non-artificial (speech) utterances. The issue of accurate classification is divided into two parts: the feature extraction and the classification. Feature extraction is a process that computes certain feature vectors from the acoustic signal. The proposed method is based on cepstral analysis. The Two-dimensional cepstrum TDC, introduced recently by Ariki et al [1] is a special case of cepstral analysis. The acoustic signal's features are represented in matrix form. The coefficients located at the lower index portion of the TDC matrix seem to include the most significant information. The TDC provides a compressed approach to simultaneous time and frequency analysis.

Pai and Wang [9], have suggested the use of TDC based on the Bayesian classifier for the recognition of vowels and isolated Mandarin utterances in a speaker dependent manner. In [10], a method based on the TDC distance and vector quantisation is suggested for classifying TDC patterns. In that paper, the distance measure of cepstra has been used. This method has been tested by applying it to the recognition of Mandarin digits. The results show that this method is very promising for the speech recognition of syllabic languages such as Mandarin Chinese. Lin and Nein and Hwu [11] have used Mel TDC for noisy speech recognition by applying genetic algorithms so as to find the robust coefficients in the Mel TDC matrix. It was found that GA-based Mel TDC provides higher recognition results compared to the original TDC approach in noisy environments.
In [12] is shown a speaker recognition model using a TDC Mel-cepstrum combined with predictive neural network. In this paper, it has also been shown, that the TDC Mel-cepstrum is very effective for speaker recognition.

The application of the TDC method as a static and dynamic feature extraction technique for speech recognition has been presented in [3]. In [4] a novel method of feature extraction for speech recognition have been introduced based on root TDC and it has been modified by using LDA implementation [5]. In previously reported research, some of the extracted features commonly used were: energy function, average zero-crossing rate, fundamental frequency, spectral peak tracks brightness [4]. Also, another approach involves the statistical feature extraction from the Wigner-Ville distribution [5] as well as the combination of statistical features derived from both spectrogram and Wigner-Ville [6].

In this paper, the statistical TDC is presented as a new feature extraction technique, for acoustic classification. Also, three commonly used distance metric classifiers will be applied to classify the acoustic pattern. The TDC-based feature extraction process is applied to both speech and non-speech (general audio) cases in order to present the effectiveness of the proposed technique.

This paper is organized as follows, in section (2) the feature extraction method will be introduced, while in section (3) a brief discussion of the classifiers used are presented. Finally, in (4) and (5), the results and the observations / conclusions are presented.

2 FEATURE EXTRACTION

Feature extraction is a process that computes certain feature vectors from the speech /audio signal by short-term spectral analysis techniques. For each short frame of speech /audio (e.g 20ms) the extraction procedure outputs a feature vector that describes the acoustical characteristics of this frame. Feature extraction has a great influence on the efficiency of recognition and classification. The selection of the features varies from one application to another. However, it is generally desired that dissimilar acoustic vectors would be clearly separable from each other in the feature space and correspondingly, similar vectors would be close to each other. In other words, the inter-class variances of the vectors should be large and intra-class variances should be small for good recognition accuracy.

The feature extraction method which is adopted here is based on the analysis of the two-dimensional cepstrum (TDC) matrix. The TDC matrix implementation can be divided into four stages [2]:

I. Apply the discrete Fourier Transform to the samples of each frame of the utterance.

$$S(m,k) = \sum_{n=0}^{N-1} s(m,n) e^{-j2kn\pi/N} \quad (1) \qquad 0 \le k \le N-1, \quad 0 \le m \le M-1$$

where *s(m,n)* is the *n* th sample in frame *m*, *N* is the number of sample data in a frame, and *M* is the number of frame used for computing the TDC matrix.

II. Evaluation of the magnitude of *S(m,k)* stage (I)

III. Apply the logarithm function in stage (II)

i.e. $S_l(m,k) = log|S(m,k)|$ (2)

IV. Finally, apply the two-dimensional inverse Fourier Transform to the matrix implemented in stage (iii) i.e. $S_l(m,k)$. Analytically,

$$\hat{s}(u,v) = \frac{1}{N} \sum_{m=0}^{M-1} \sum_{k=0}^{N-1} S_l(m,k) e^{j2\pi kv/N} e^{j2\pi mu/M} \qquad 0 \le u \le N-1, \quad 0 \le v \le M-1 \quad (3)$$

So, the TDC matrix is defined as the absolute value of every element in matrix

$\hat{s}(u,v)$ i.e. $c(i,j) = |\hat{s}(u,v)|$ (4)

The axis *v* is called quefrency and represents the time dimension, where as the axis *u* is called time ferquency and represnts the frequency dimension.

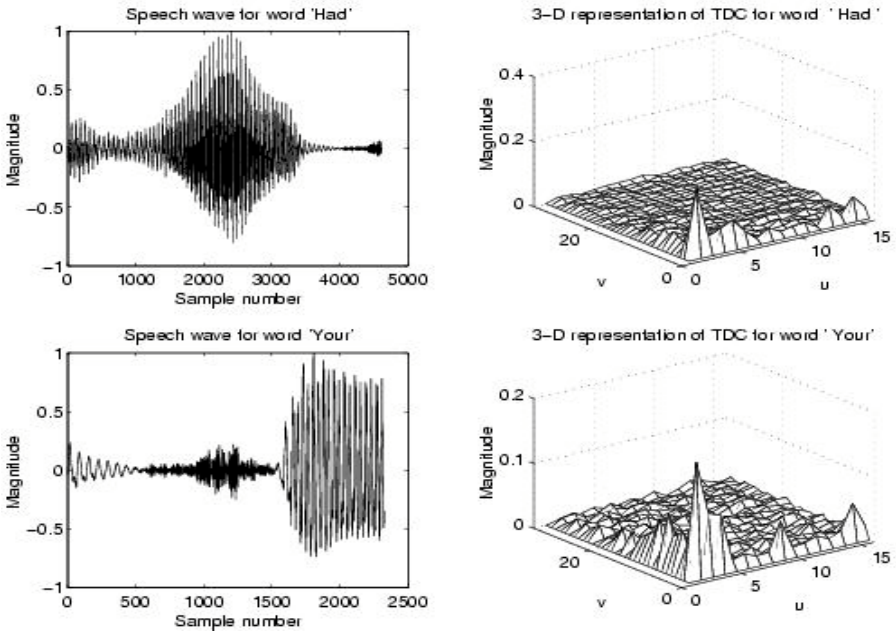

Fig. 1: Time-domain and TDC signal representation

Figure (1) shows the time domain and TDC reprentation of two words from the TIMIT database. It can be observed that the most inmportant information is concentrated in the low frequency and low quefrency region.

2.1 Feature Compression

After the TDC is implemented, it is necessary to reduce the size of the matrix in order to present it to the classifier. There are two ways to effectively compress the matrix. The first one is to use a limited number of coefficients that are enough in order to efficiently represent the utterance. The second way is to derive statistical features from the chosen coefficients so as to substitute them with an even more limited number of descriptive statistical values. The statistical analysis process includes the use of the following features:

i) Two-dimensional variance: is a measure of the spread of a distribution. It is defined as the mean of the squares of the differences between the data samples and the data mean value.
ii) Two-dimensional skewness: is a measure of asymmetry of the samples around the mean value. It is defined as the mean of the third power of the differences between the data samples and the data mean value.
iii) Two-dimensional kurtosis: is a measure of the outlier proness of a distribution. It is defined as the mean of the fourth power of the differences between the data samples and the data mean value.
iv) Entropy: describes information-related features and aims the accurate signal representation. In this case the 'log energy' entropy is employed.
v) Inter-quartile range: is a measure of dispersion. In most cases, the 50% of the data located in the middle is most important. So, the interquartile range is the difference between the 75^{th} and 25^{th} data percentile.
vi) Median: is a measure of location. It is defined as the middle value of the data set. It divides the data set in half, fifty percent of the measurements being over and fifty percent below it.
vii) Mean absolute deviation: is a measure of dispersion. It evaluates the average of the absolute differences between the data samples and the mean of the data samples.

3 CLASSIFICATION

Three different kinds of classifiers were employed in order to classify the acoustic pattern.
The first of them is the Euclidean distance metric which is a common classifier, defined as:

$$d(c_t, c_r) = \sqrt{\sum_{i=0}^{m} [c_r(i) - c_t(i)]^2} \quad (5)$$

where c_t and c_r are the reference and the test cepstral vectors respectively and *m* is the dimension of vectors.
The second is the Minkowski measure:

$$d(c_t, c_r) = \sum_{i=0}^{m} | c_r(i) - c_t(i) | \quad (6)$$

where c_t and c_r are the reference and the test cepstral vectors respectively and *m* is the dimension of vectors.
And the third is the Mahalanobis:

$$d(x_t, x_r) = (x_r - x_t) C_r^{-1} (x_r - x_t)^T \quad (7)$$

where x_t and x_r are the reference and the test cepstral vectors respectively and C_r is the covariance method of the reference data.

A codebook is derived from the mean values of each class, so each class is represented by each codeword. Then, the distance between each test pattern and the codeword is calculated using each one of the three different distance metrics. Therefore, the test pattern is classified based on the minimum distance between itself and the codeword.

4 EXPERIMENTAL RESULTS

4.1 Experimental set up

Tests were carried out using two databases. The first one, tested for speech, is a set of isolated words selected from the TIMIT database which is a fully labelled database of American English. It consists of utterances of 360 speakers that represent the major dialects of American English. It is divided into eight dialect regions with separate testing and training sections.

The second one includes acoustic utterances recorded from sounds of different kinds (classes) of gunshots. To name but a few of them: Firing a revolver with echo, Firing a .22caliber handgun, Firing an M-1 rifle, Firing a World War II German rifle, Firing a cannon, Firing a 30-30 rifle outdoors with echo, Firing of a .38 calibre semi-automatic pistol, Firing and cocking a lever action Winchester rifle outdoors with echo, Firing a 37mm anti-tank gun etc.

All utterances (speech/audio) are divided into 32 frames. Each frame is 256 samples long and there is 128 sample spacing between each frame. So each matrix has 32*256 size although only a limited number of coefficients is used in order to represent each utterance.

4.2 Experimental results in Audio

Tables (1) and (2) show the performance of the TDC and the statistical TDC applied to audio using various number of features and the three different types of distance measures. According to table 1,

TDC method, the number of features equals the number of coefficients extracted from the TDC whereas table 2, statistical TDC method, indicates that the number of features is always constant (seven statistical values) irrespective of the number of coefficients.

As can be seen from Tables 1 and 2, it is clear that the TDC method provides higher classification scores compared to the statistical TDC. However, the number of features used for the statistical TDC is always constant and independent to the number of coefficients. As it is shown, in the TDC case, the 100% classification is achieved using 16 features, whereas, for the statistical TDC (STDC), the same percentage is achieved using only the seven statistical features extracted from the same number of features (i.e. 16). This indicates that the STDC is an effective compression technique for classification purposes.

No. of Coef.	No. of Feature	Euclidean	Minkowski	Mahalanobis
8	8	82.6	69.6	97.1
10	10	89.8	87.0	97.1
16	16	94.2	89.8	100.0
24	24	98.6	94.2	100.0
32	32	98.6	94.2	100.0

Table 1: Classification score (%) using the TDC method – audio case

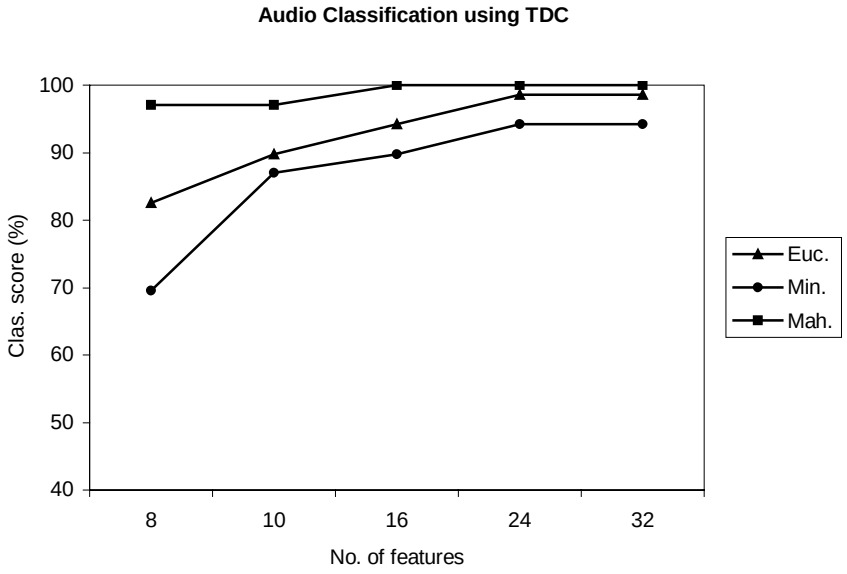

Figure 2: Audio classification using TDC

No. of Coef.	No. of features	Euclidean	Minkowski	Mahalanobis
8	8	71.8	59.1	88.4
10	8	84.9	83.6	94.5
16	8	92.7	88.8	100.0
24	8	94.2	90.5	100.0
32	8	96.3	91.4	100.0

Table 2: Classification rate (%) using the statistical TDC method – audio case

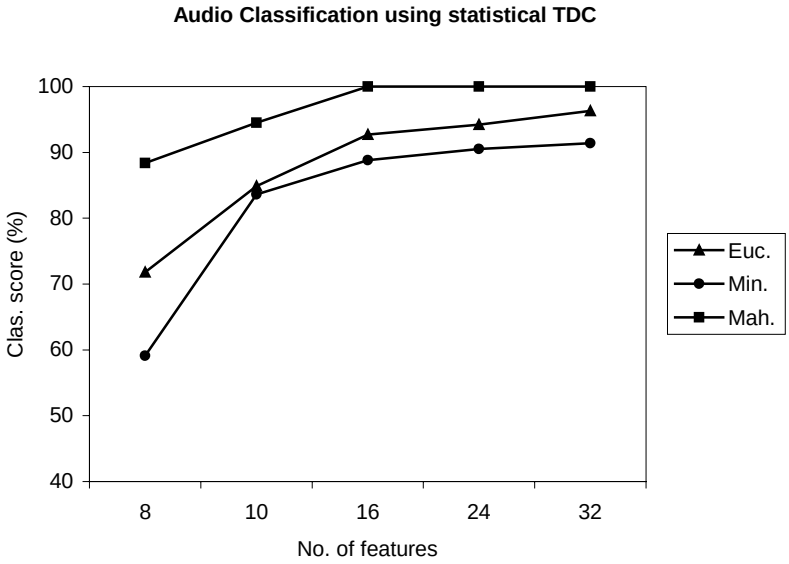

Figure 3: Audio classification using statistical TDC

In the cases of TDC and STDC all three different kinds of classifiers are used in order to compare the results and specify the optimum combination between classifier and number of coefficients. Based on figures 1 and 2, the Mahalanobis distance classifier exhibits the highest classification scores compared to the Euclidean and Minkowski.

4.3 Experimental results in Speech

Tables 3 and 4 present the results of the TDC and the STDC applied to speech for different distance measure. Similarly to the audio case, the TDC has better performance than the STDC, whereas its compression ability is lower compared to the STDC.

It can be seen from figures 3 and 4 that the higher the number of features the higher the classification rate. Also, from figure 1 Mahalanobis presents the highest classification rate for the whole range of number of features.

No. of Coef.	No. of features	Euclidean	Minkowski	Mahalanobis
8	8	76.7	75.6	80.0
16	16	82.2	80.0	86.7
24	24	90.0	88.1	93.3
32	32	96.7	93.9	98.9
40	40	97.8	96.4	98.9
48	48	98.9	97.1	100.0
54	54	98.9	98.2	100.0
64	64	100.0	100.0	100.0

Table 3: Classification score (%) using the TDC method – speech case

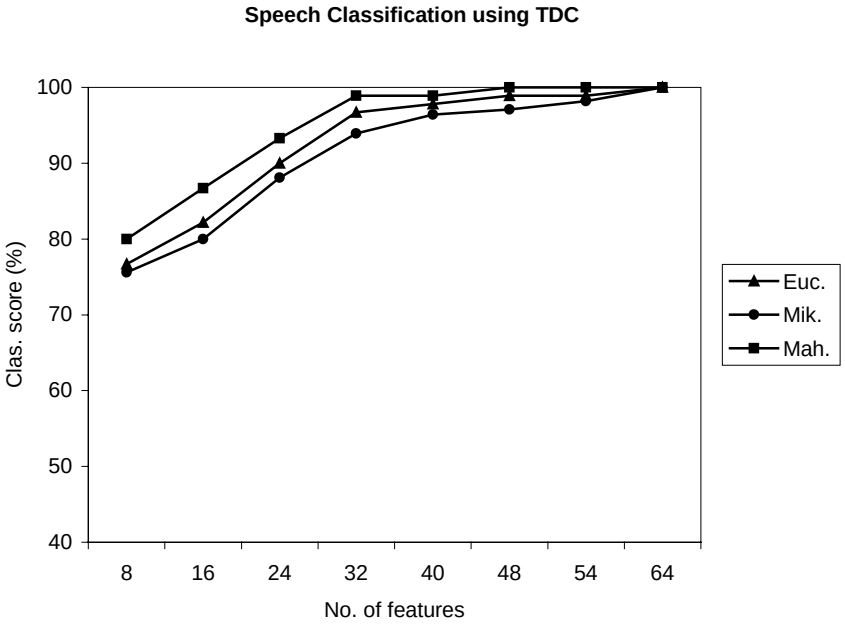

Figure 4: Speech classification using TDC

No. of Coef.	No. of features	Euclidean	Minkowski	Mahalanobis
8	8	51.2	49.8	64.6
16	8	68.7	65.2	77.8
24	8	78.1	74.6	83.8
32	8	84.9	80.0	85.9
40	8	92.4	91.1	93.9
48	8	95.6	92.4	98.1
54	8	97.1	94.9	100.0
64	8	100.0	98.3	100.0

Table 4: Classification rate (%) using the statistical TDC method – speech case

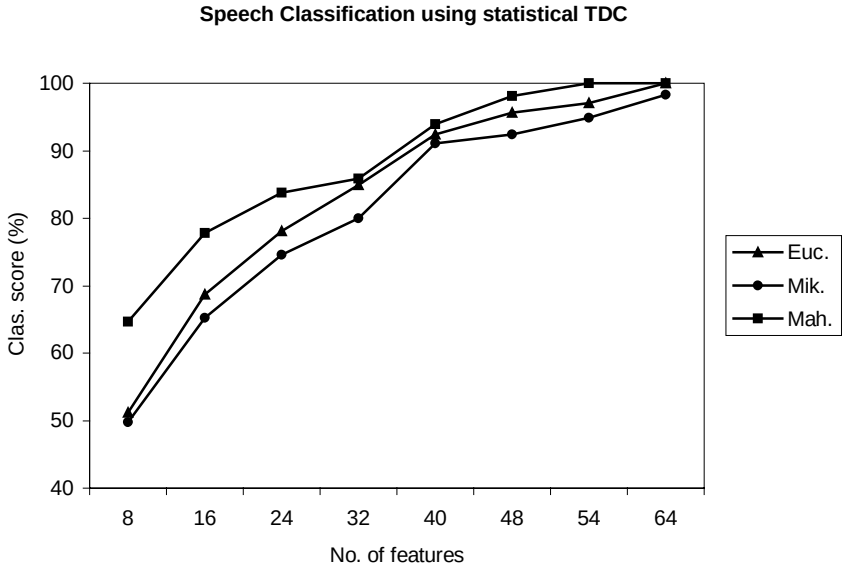

Figure 5: Speech classification using statistical TDC

Another observation is that by using the same number of features, audio presents higher classification rates compared to speech. Figure 3 and 4 are presenting graphically the classification scores for each combination of classifiers and number of features.

5 CONCLUSIONS

There are many significant observations that can be derived from the experimental results. The TDC method is an effective compression method. It can be observed that most of the important information is concentrated in the low frequency and low quefrency region. Applying the statistical analysis to the TDC coefficients, the compression ratio significantly increases whereas the classification score has similar values (STDC case), especially for the case of audio. Finally, from the tables it can be observed that when the number of coefficients exceeds a certain limit, the classification score remains high irrespective of the number of the coefficients chosen (saturation region).
Summarising, the STDC is suggested as a new feature extraction method with effective compression ability for both speech and audio cases, as illustrated from the experimental results.

6 REFERENCES

1. Y..Ariki, S.Mizuta, M.Nagata, and T.Sakai., "Spoken-word recognition using dynamic features analysed by two-dimensional cepstrum," *IEE Proceedings*, 136:133-140, 1989.

2. H.F. Pai and H.C. Wang., "A study on two-dimensional cepstrum approach for the speech recognition," In *Proceeding of International Computer Symposium*, pages 975-980, 1990.

3. H.Marvi and E.Chilton, "Dynamic and static feature extraction for speech recognition using two-dimensional cepstrum," In *the 9th Electrical and Electronic Engineering seminar of researchers in Europe*, Birmingham, U.K., June 2002.

4. E.Chilton and H.Marvi, "Two-dimensional root cepstrum as a feature extraction method for speech recognition," *Electronics Letters*, 39[10]:815-816, May 2003.

5. H.Marvi and E.Chilton, “Application of LDA to improve the accuracy of TDRC,” In *the proceeding of SCI 2003, Orlando, Florida, USA*, p.p. 328-331.

6. T.Zhang, and C.C.J. Kuo, “Audio content analysis for online audiovisual data segmentation and classification,” *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 4, May 2000

7. I.Paraskevas and E.Chilton, “Fine classification for acoustic images,” In *Proceeding of International Signal Processing Conference ISPC ‘03*.

8. I.Paraskevas and E.Chilton “Audio Classification using acoustic images for retrieval from multimedia databases,” In *Proceeding of 4th EURASIP Conference focused on Video/Image Processing and Multimedia Communications, 2-5 July 2003, Zagreb, Croatia,p.p. 187-192*.

9. H. F. Pai and H.C. Wang, “A study on two-dimensional cepstrum approach for the speech recognition” In *the proceeding of Internationla computer symposium, pages 975-980, 1990*

10. H. F. Pai and H.C. Wang, “ Two-dimensional cepstral distance measure for speech recognition”
In the proceeding of International conference on Acoustics speech and signal processing, 1993.

11. C. T. Lin, H. W. Nein and J. Y. Hwn , “Ga-based noisy speech recogntion using two-dimensional cepsturm .” IEEE Transaction on speech and Audio processing, 8(6):664-675, November 2000.

12. T. Kitamora and S. Takei, “ Speaker recognition model using two-dimensional mel-cepstrum and predictive neural network ” In the Fourth International conference on spoken Language processing 1996.