Hybrid Techniques based Speech Recognition

Ahlam Hanoon Shini
Computer Eng. Depart.
University of Baghdad, Iraq

Zainab Ibrahim Abood
Electrical Eng. Depart
University of Baghdad, Iraq

Tariq Ziad Ismaeel
Electronic Eng. Depart.
University of Baghdad, Iraq

ABSTRACT

Information processing has an important application which is speech recognition. In this paper, a two hybrid techniques have been presented. The first one is a 3-level hybrid of Stationary Wavelet Transform (S) and Discrete Wavelet Transform (W) and the second one is a 3-level hybrid of Discrete Wavelet Transform (W) and Multi-wavelet Transforms (M). To choose the best 3-level hybrid in each technique, a comparison according to five factors has been implemented and the best results are WWS, WWW, and MWM. Speech recognition is performed on WWS, WWW, and MWM using Euclidean distance (Ecl) and Dynamic Time Warping (DTW). The match performance is (98%) using DTW in MWM, while in the WWS and WWW are (74%) and (78%) respectively, but when using (Ecl) distance match performance is (62%) in MWM. So, in speech recognition to get the high alignment and high performance one must use DTW distance measurement.

Keywords

Hybrid techniques, speech recognition, multi-wavelet transform, wavelet transform, stationary wavelet transform, feature extraction, dynamic time warping.

1. INTRODUCTION

Speech recognition is the process of automatically choosing and determining language information conveyed by the speech signal using electronic circuits or computers [1].

Sylvio introduced dynamic time warping for speech recognition, which is based on alignment of the template models with the input signal. Dynamic time warping has a drawback of a high computational cost that appears as the length of the signal increases. So, DTW based on discrete wavelet transform was introduced to overcome this problem [2]. A multi-resolution time-frequency wavelet transform is presented by Nitin. By using different Wavelets, decomposition the speech signal into different frequency channels has been implemented, and then the wavelet coefficients are considered as feature vectors. Feed forward network with three layers is used for classification the words and the result is that for 5-level DWT and Daubechies 8 wavelet the accuracy is (90.42%) [1].

Zainab introduced image recognition using 2^i techniques of 3-level stationary wavelet transform (SWT) and discrete wavelet transform (DWT), a comparison between them has been implemented. In image recognition, SWW technique has a match performance of (100%) which is higher performance than the WWW technique [3].

Feature extraction is a process of removing redundant and unwanted information and retaining the useful information. In practice some important information may be lost when using this process. The feature extraction goal is to find out a set of properties which is called as utterances' parameter by processing the utterances' signal waveform. These parameters are called the features. After the preprocessing of the speech signal feature extraction is achieved, it produces the meaningful representation of a speech signal. Feature extraction includes a process of converting the speech signals into a digital form and measuring important characteristics of the signal i.e. frequency or energy and augment these measurements with the meaningful derived measurements [4].

For solving a global distance matrix, John introduced an adapted DTW by which template digit utterances are compared with TIDIGITs data. The performance of his proposed technique (DTW + DWT level5) is tested with the recognition accuracy is 79% while the conventional approach has an accuracy of 66% [5].

2. WAVELET TRANSFORM

2.1 Discrete Wavelet Transform

Wavelet transform is the technique that processes the data at different scale and resolution. In the wavelet transform the output has two sets of coefficients, the approximation coefficients and the detail coefficients.

In discrete wavelet transform, for computing coefficients of the wavelet transform, the analysis must be transformed to a pyramidal and fast algorithm [6]. The scaling function is given by [7]**:**

$$\phi[t] = \sqrt{2}\sum_{k=-\infty}^{\infty} h_k . \phi[2t-k] \qquad (1)$$

and Wavelet function is given by

$$\psi[t] = \sqrt{2}\sum_{k=-\infty}^{\infty} g_k . \phi[2t-k] \qquad (2)$$

where h_k is the scaling filter coefficient, g_k is the wavelet filter coefficient [7**]** and ϕ[2t-k] is the scaling function with dilations and translations [6].

2.2 Stationary Wavelet Transform

Stationary wavelet transform is a wavelet transform algorithm that was designed to overcome non ability of translation-invariance of the discrete wavelet transform. Translation-invariance is implemented by removing the down-sampling and up-sampling in the wavelet transform, and then up-sampling the coefficients of the filter by a factor of $2^{(m-1)}$ in the level m^{th} of the algorithm. The important advantage of SWT is that it preserves the original signal sequence's time information at each level. In some applications, the SWT is used for modeling ECG beats and denoising process [8].

2.3 Multi-Wavelet transform

In multi-wavelet, to represent a signal, two or greater than two scaling and wavelet functions must be used. For multi-wavelet, the dilation and wavelet equations can be represented by [7, 9]

$$\phi[t] = \sqrt{2}\sum_{k=-\infty}^{\infty} H_k . \phi[2t-k] \qquad (3)$$

$$\psi[t] = \sqrt{2}\sum_{n} G_k . \phi[2t-k] \qquad (4)$$

where H_k and G_k are the matrix filters. They are representing the r×r matrices of each integer k. In these filters the matrix elements provide degrees of freedom higher than that in the traditional scalar wavelet [7].

3. FEATURE EXTRACTION

In speech recognition, the most important part is the feature extraction which can help in distinguishing between one speech signal from the other. The aim of the feature extraction is to compute feature vectors to provide a compact representation of the input signal [4].

4. COMPARISON FACTORS

In this paper, five factors, Compression Ratio (cr), Root Mean Square Error (RMSE), Peak Signal to Noise Ratio (PSNR), Energy (en) and Mean, are used for the comparison between 3-level hybrid in each technique.

4.1 Compression ratio (cr)

Compression ratio is the measuring of the ratio between the size of the speech signal before and after compression [10]:

$$(\mathrm{cr}) = \frac{\text{size befor compression}}{\text{size after compression}} \qquad (5)$$

4.2 Root Mean Square Error (RMSE)

The root mean square error between the original and compressed speech signals is calculated by:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\mathrm{X(i,j)} - \mathrm{C(i,j)})^2} \qquad (6)$$

where $\mathrm{X(i,j)}$ and $\mathrm{C(i,j)}$ are the original and compressed speech signals [11].

4.3 Peak Signal to Noise Ratio (PSNR)

Peak signal to noise ratio is the measure of maximum error [3]:

$$\mathrm{PSNR} = 20\log_{10}\frac{1}{\sqrt{\mathrm{MSE}}} \qquad (7)$$

4.4 Energy (en)

The speech signal consists of small frames, where each frame has ω samples. Frame by frame, the energy of the speech can be calculated as in the following eq. [12]:

$$\mathrm{en} = \sum_{i=1}^{w} x_i^2 \qquad (8)$$

4.5 Mean

Average the columns of speech signal as vectors, and then returning the row vector of the mean values [13].

$$\mathrm{Mean} = \sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\mathrm{X(i,j)}}{n} \qquad (9)$$

4.6 Speech Recognition

Speech recognition is the task of features extraction from the speech signal and classifying these features. The aim is to distinguish between speech signals with high accuracy. The speech recognition process consists of two stages. The first one is the training stage and the second one is the recognition stage. In training stage the features of the speech are firstly extracted and then saved as a reference template. The recognition stage may be divided into two stages. The first stage is a feature extraction stage in which short time spectral or temporal features are extracted. The second stage is a classification stage wherein a comparison between the derived parameters and the stored reference parameters is implemented and then decisions are made according to some type of the minimum distance rule [6].

4.7 Euclidean Distance

Euclidean distance is the distance measurement between the feature vectors of the reference and tested speech signal.

$$\mathrm{Ecl} = \left(\sum(\mathrm{f}-\mathrm{t})^2\right)^{\frac{1}{2}} \qquad (10)$$

where f and t are the two feature vectors of the reference and test words [13]

4.8 Dynamic Time Warping

Dynamic time warping is a technique used to find the best alignment between two time-dependent sequences. In order the sequences match each other, they must be warped nonlinearly [14]:

$$\mathrm{d(i, j) = d\ (f\ (n), t\ (n'))} \qquad (11)$$

Where f(n) is the feature vector extracted from ith frame while t(n') is the feature vector extracted from jth frame [6].

5. THE PROPOSED TECHNIQUES ARCHITECTURE

A two hybrid techniques based Speech recognition has been presented in this paper, the first technique is a 3-level hybrid using (W) and (S) transforms, SSS, SSW, SWS, SWW, WSS, WSW, WWS and WWW (i.e., in WSW, the first level is wavelet transform, the stationary wavelet is applied to the low-low sub-band of wavelet as a second level and the third level is wavelet transform applied to the low-low sub-band of the stationary wavelet and so on..), while the second technique a 3-level hybrid using (M) and (W) transforms, WWW, WWM, WMW, WMM, MWW, MWM, MMW and MMM as shown in "figure 1".

There are 21 sets of words used as reference sets, each set contains 5 words recorded in different time and environment. These words are (Accountable (Acc), Beautiful (Beaut), Blackboard (Black), Confused (Con), Congratulations (Cong), Darkness (Dar), Decisions (Dec), Despaired (Des), Everything (Every), Forever (For), Infatuation (Inf). Ourselves (Our), Prostrate (Pro), Refreshment (Ref), Mistakes (Mis), Revolutions (Rel), Sleepers (Sleep), Spreading (Spr), Translation (Tran), Understanding (Und) and Yourself (Your)).

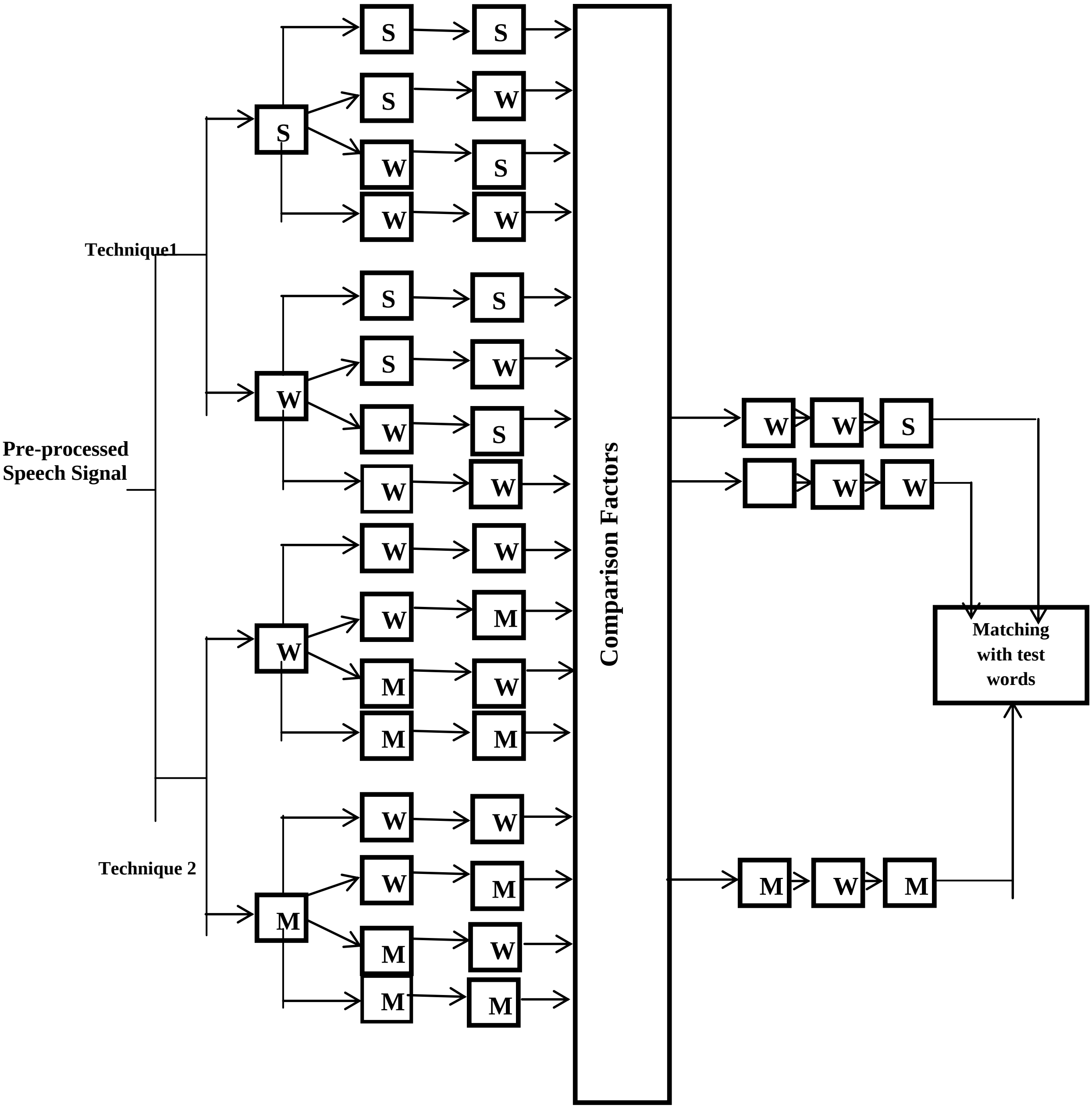

Fig 1: The Proposed Techniques Architecture

6. PROPOSED TECHNIQUE ALGORITHM

1. Input the recorded speech to be preprocessed. The preprocessing is applied to the test and reference sets.
2. Segment the speech into frames. There is 25% overlap.
3. With hamming window, the frames are windowed.
4. Resize the windows to be of size (256*256).
5. In the first technique, the 3-level transform such as SSS , is produced by applying a stationary wavelet transform to each word in the test and reference sets, then applying a second level stationary wavelet to the low-low sub-band of the first level, finally applying a third level stationary wavelet to the low-low sub-band of the second level.
6. The 3-level transform such as SSW, is produced by applying the stationary wavelet transform to each word in the test and reference sets, then applying a second level, finally applying a third level wavelet transform to the lowlow sub band of the second level.
7. The same procedure is followed with SWS, SWW, WSS, WSW, WWS and WWW.
8. In the second technique, the 3-level transform such as WWW, is produced by applying wavelet transform to each word in the test and reference sets (after steps from 1 to 4), then applying a second level wavelet transform to the low-low sub-band of the first level, finally applying a third level wavelet transform to the low-low sub-band of the second level.

9. The 3-level transform such as WWM, is produced by applying Wavelet transform to each word in the test and reference sets, then applying a second level wavelet transform to the low-low sub-band of the first level, finally applying a third level multi-wavelet transform to the low- low sub-band of the second level.
10. The same procedure is applied with WMW, WMM, MWW, MWM, MMW and MMM.
11. A comparison between these 3-level hybrid for each technique according to the values of the factors determined using eq's.5, 6, 7, 8 and 9 has been implemented. The best 3-level in the first technique are WWS and WWW, while in the second technique are WWW and MWM, so the best 3-level are WWS, WWW and MWM.
12. The coefficients of the low-low sub-band of the third level of each word of the WWS, WWW and MWM in the test and reference sets are considered as feature vectors to find the minimum distance that used for speech recognition by calculating Euclidean and DTW distances between the reference and test feature vectors using eq's.10 and 11.

8. IMPLEMENTATION AND RESULTS

The algorithm of the proposed techniques was simulated on Matlab. The sampling frequency used is 8 kHz.

"Table 1" shows the comparison between the different three levels of the first technique according to the five factors, as it's obvious from the results that the WWW is the best one in the PSNR, RMSE, en and cr, WWS is also good in PSNR and RMSE, so the best two 3-level in this technique are WWW and WWS. This also illustrated in the "figure 2".

Table1. Comparison between the differentthree levels of first technique

	sss	ssw	sws	sww	wss	wsw	wws	www
PS-NR	**24.0**	**26.9**	**27.2**	**29.08**	**27.9**	**30.7**	**33.1**	**36.7**
RM-SE	**0.07**	**0.05**	**0.05**	**0.03**	**0.05**	**0.03**	**0.02**	**0.01**
en	**1.8 e-06**	**3.6 e-06**	**2.5 e-06**	**5.0 e-06**	**2.40 e-06**	**4.4 e-06**	**2.8 e-06**	**5.80 e-06**
cr	**1**	**4**	**4**	**16**	**4**	**16**	**16**	**64**
Me- an	**0.34**	**1.20**	**0.16**	**0.13**	**0.18**	**0.13**	**0.10**	**0.07**

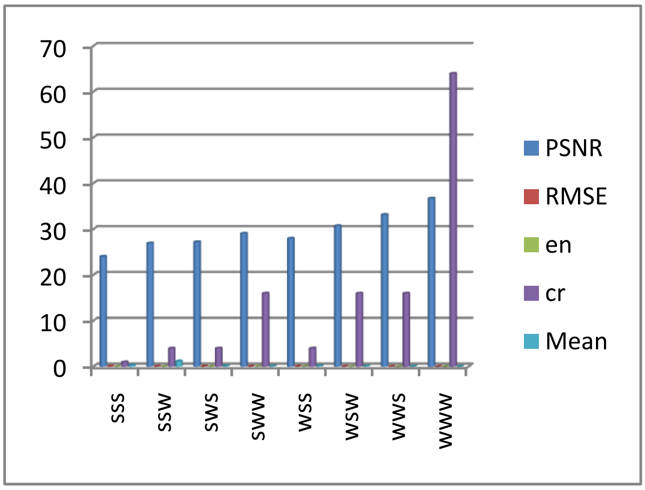

Fig 2: Comparison between the differentthree levels of first technique

"Table 2" shows the comparison between the different 3-level of the second technique, as it's obvious from the results that the MWM is the best one in PSNR, en and cr, while WWW is the best one in PSNR and RMSE. MMM has a higher values of cr and en, so it is good in speech compression applications, therefore; the best two 3-level in this technique are MWM and WWW. This also illustrated in the "figure 3".

Table2. Comparison between the differentthree levels of second technique

	WWW	WWM	WMW	WMM	MWW	MWM	MMW	MMM
PSNR$_{av}$	**36.7**	**30**	**29.2**	**23.7**	**30.6**	**31.0**	**26.3**	**26.2**
RMSE$_{av}$	**0.018**	**0.05**	**0.04**	**0.082**	**.046**	**0.06**	**0.08**	**0.07**
en$_{av}$	**0.585e-05**	**4.49 e-05**	**5.03 e-05**	**34.27 e-05**	**3.56 e-05**	**37.8e -05**	**26.4e -05**	**291 e-05**
cr$_{av}$	**64**	**256**	**256**	**1024**	**256**	**1024**	**1024**	**4096**
Mean$_{av}$	**0.072**	**0.10**	**.108**	**0.139**	**0.08**	**0.14**	**0.12**	**0.19**

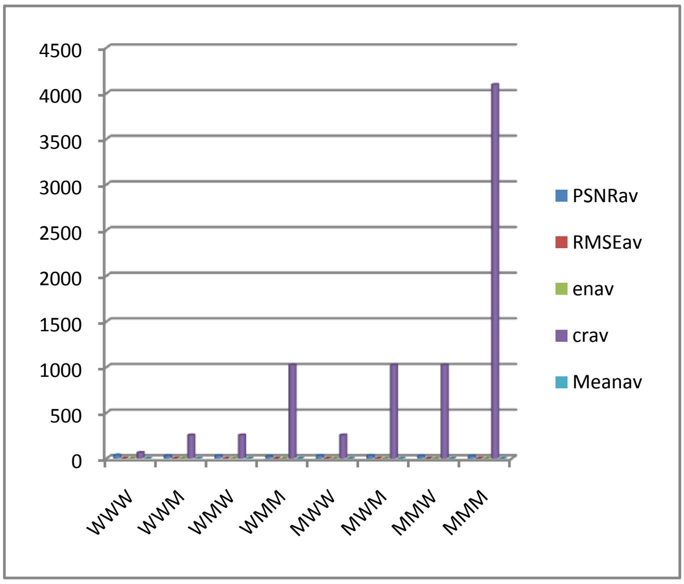

Fig 3: Comparison between the differentthree levels of second technique

As shown in tables (1 and 2), the cr in WWS, WWW and MWM are 16, 64 and 1024 respectively, because the size of the 3rd level of each one is (64*64), (32*32) and (8*8) respectively; Therefore, the time required for computation in MWM is very short as compared with WWS and WWW.

“Table 3” shows a sample of minimum distance measurements between the reference and test feature vectors for MWM using (Ecl) and DTW. Number “1” refers to “Recognize” and number “0” refers to “Not recognize”. In (Ecl) distance there is a minimum distance between the test word (Con) and the reference words (Con2) and (Con5), while in DTW there is a minimum distance between the test word (Con) and the (Con) set of the reference words (Con2), (Con3), (Con4) and (Con5) that is means that the DTW is better than (Ecl) in the recognition of the word (Con), and so on for the other words, so DTW is the best technique used for speech recognition.

Table3. Sample of minimum distance measurements

MWM		Ecl						DTW		
Reference\Test	CON	DAR	DES	DEC	MIS	Data\Test	CON	DAR	DES	DEC	MIS
CON1	0	0	0	0	0	CON1	0	0	0	0	0
CON2	1	0	0	0	0	CON2	1	0	0	0	0
CON3	0	0	0	0	0	CON3	1	0	0	0	0
CON4	0	0	0	0	0	CON4	1	0	0	0	0
CON5	1	0	0	0	0	CON5	1	0	0	0	0
DAR1	0	0	0	0	0	DAR1	0	1	0	0	0
DAR2	0	0	0	0	0	DAR2	0	1	0	0	0
DAR3	0	0	0	0	0	DAR3	0	0	0	0	0
DAR4	0	1	0	0	0	DAR4	0	1	0	0	0
DAR5	0	1	0	0	0	DAR5	0	1	0	0	0
DES1	0	0	1	0	0	DES1	0	0	1	0	0
DES2	0	0	1	0	0	DES2	0	0	1	0	0
DES3	0	0	0	0	0	DES3	0	0	1	0	0
DES4	0	0	0	0	0	DES4	0	0	1	0	0
DES5	0	0	1	0	0	DES5	0	0	1	0	0
DEC1	0	0	0	1	0	DEC1	0	0	0	1	0
DEC2	0	0	0	1	0	DEC2	0	0	0	1	0
DEC3	0	0	0	0	0	DEC3	0	0	0	1	0
DEC4	0	0	0	1	0	DEC4	0	0	0	1	0
DEC5	0	0	0	1	0	DEC5	0	0	0	1	0
MIS1	0	0	0	0	0	MIS1	0	0	0	0	1
MIS2	0	0	0	0	1	MIS2	0	0	0	0	1
MIS3	0	0	0	0	1	MIS3	0	0	0	0	1
MIS4	0	0	0	0	0	MIS4	0	0	0	0	1
MIS5	0	0	0	0	1	MIS5	0	0	0	0	1

“Table4” shows sample of match performance for WWS, WWW and MWM. As shown from the results, the matching performance in WWS and WWW using Ecl distance is very poor, while in MWM is better. By using DTW, the performance of the three types of hybrid is good and the last one is the best due to the alignment between the test and reference words especially in the words (Des), (Dec) and (Mis). Therefore; the match performance is (98%) in MWM, while in the WWS and WWW is (74%) and (78%) respectively as shown in “table 5” and the chart of “figure 4”.

Table4. Sample of match performance of WWS, WWW and MWM

	WWS		WWW		MWM	
Test	Con		Con		Con	
Reference	Ecl	DTW	Ecl	DTW	Ecl	DTW
Con1	0	0	0	0	0’	0
Con2	0	1	0	1	1	1
Con3	0	1	0	1	0	1
Con4	0	0	0	0	0	1
Con5	1	1	1	1	1	1
	WWS		WWW		MWM	
Test	Dar		Dar		Dar	

Reference	Ecl	DTW	Ecl	DTW	Ecl	DTW
Dar 1	0	1	0	1	0	1
Dar 2	0	1	0	1	0	1
Dar 3	0	0	0	0	0	0
Dar 4	0	1	0	1	1	1
Dar 5	1	1	1	1	1	1
	WWS		WWW		MWM	
Test	Des		Des		Des	
Reference	Ecl	DTW	Ecl	DTW	Ecl	DTW
Des 1	0	0	0	0	1	1
Des 2	0	0	0	0	1	1
Des 3	0	0	0	0	0	1
Des 4	0	0	0	1	0	1
Des 5	1	1	1	1	1	1
	WWS		WWW		MWM	
Test	Dec		Dec		Dec	
Reference	Ecl	DTW	Ecl	DTW	Ecl	DTW
Dec 1	0	1	0	1	1	1
Dec 2	0	1	0	1	1	1
Dec 3	0	0	0	0	0	1
Dec 4	0	1	0	1	1	1
Dec 5	1	1	1	1	1	1
	WWS		WWW		MWM	
Test	Mis		Mis		Mis	
Reference	Ecl	DTW	Ecl.	DTW	Ecl	DTW
Mis 1	0	1	0	1	0	1
Mis 2	1	1	1	1	1	1
Mis 3	0	1	0	1	1	1
Mis 4	0	1	0	1	0	1
Mis 5	1	1	1	1	1	1

Table5. Percentage matching performance

Recognition	WWS		WWW		MWM	
	Ecl	DTW	Ecl	DTW	Ecl	DTW
Matching %	30	74	30	78	62	98

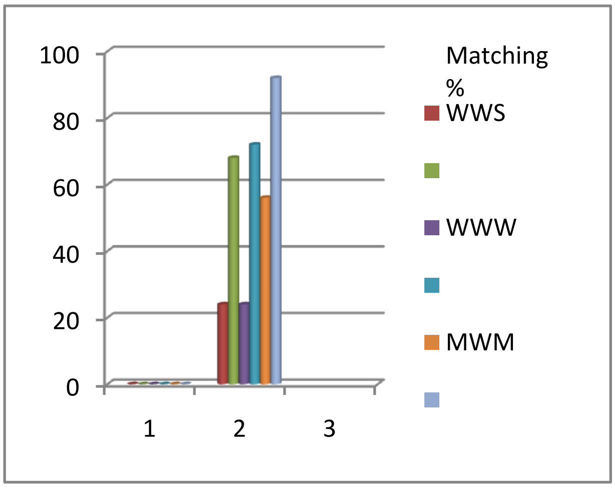

Fig 4: Percentage matching performance

9. CONCLUSIONS

In this paper, speech recognition was performed on 21 sets each of five words. Two hybrid techniques are used each of 8, 3-level hybrid transform, one of them is S and W and the other one is W and M. WWS ,WWW and MWM have a best results according to the five factors, cr, RMSE, PSNR, en and mean. The speech recognition is implemented on WWS, WWW and MWM using (Ecl) and DTW, the match performance is (98%) using DTW in MWM, while in the WWS and WWW are (74%) and (78%) respectively, but when using (Ecl) distance match performance is (62%) in MWM. Also, the time of computation in the MWM is very short as compared with WWS and WWW so it's preferred to be used in speech compression. In speech recognition, to get the high alignment and high performance DTW must be used. In future, these proposed techniques can be used for image recognition and compression, MMM is the best technique for speech compression, also; the speech recognition can performed on 3 - level hybrid techniques using neural network and DTW

10. REFERENCES

[1] N. Trivedi., V. Kumar., S. Kumar, S. Ahuja, R. Chadha, "Speech Recognition by Wavelet Analysis", International Journal of Computer Applications, Vol.15– No.8, Feb. 2011.

[2] S. B. Jr., R. C. Guido, L. S. Vieira, E. S. Fonseca, F. L. Sanchez, P.R. Scalassara, C. D. Maciel, J. C. Pereira and S. H. Chen, "Wavelet-based dynamic time warping", Journal of Computational and Applied Mathematics 2009.

[3] Z. I. Abood, A. H. Al-sudani, "3-Level Techniques Comparison based Image Recognition", International Journal of Computer Applications, Vol.97– No.11, July 2014.

[4] K R. Ghule, R. R. Deshmukh, "Feature Extraction Techniques for Speech Recognition: A Review", International Journal of Scientific & Engineering Research, Vol. 6, Issue 5, May-2015.

[5] J. Sahaya, R. Alex, T. S. Shivkumar and N. Venkatesan, "Adapted DTW Joint with Wavelet Transform for Isolated Digit Recognition", ARPN Journal of Engineering andApplied Sciences, Vol.10, No.1, Jan.2015.

[6] A. Chugh,P. Rana, S. Rana, “Speech Recognition System Using Wavelet Transform”, Research Article, International Journal of Computer Science and Mobile Computing, Vol. 3, Issue 8 Aug. 2014, PP 63-71.

[7] M. B. Martin and A. E. Bell, “New Image Compression Techniques using Multi-Wavelets and Multi-Wavelet Packets”, IEEE Transactions on Image Processing, Vol. 10, No. 4, Apr. 2001.

[8] S. Saminu, N. Özkurt, “Stationary Wavelet Transform and Entropy-Based Features for ECG Beat Classification”, International Journal of Research Studies in Science, Engineering and Technology Vol. 2, Issue 7, July 2015, PP 23-32.

[9] S. Bhatnagar and R. C. Jain, “A Comparative Analysis and Applications of MultiWavelet Transform in Image Design”, International Journal on Cybernetics & Informatics, Vol. 4, No. 2, Apr. 2015.

[10] S.R. Kodituwakku, U. S. Amarasinghe, “Comparison of Lossless Data Compression Algorithms for Text Data”, Indian Journal of Computer Science and Engineering, Vol 1 No. 4, PP 416-425.

[11] K. Kannan, S. A. Perumal, K. Arulmozhi, “Optimal Decomposition Level of Discrete, Stationary and Dual Tree Complex Wavelet Transform for Pixel based Fusion of Multi-focused Images”, Serbian Journal of Electrical Engineering, Vol. 7, No. 1, May 2010, PP 81-93.

[12] M. R Gamit, K. Dhameliya, “Isolated Words Recognition using MFCC, LPC and Neural Network”, International Journal of Research in Engineering and Technology, Vol.04 Issue: 06, June 2015.

[13] Z. I. Abood, I. J. Muhsin, N. J. Tawfiq, “Content-based Image Retrieval (CBIR) using Hybrid Technique”, International Journal of Computer Applications, Vol. 83 – No 12, Dec. 2013.

[14] M. Müller,m “Information Retrieval for Music and Motion”, Book, 2007.