Context Dependent *Tri-Phone* Automatic Speech Recognition using Novel Spectrum Analysis Approach

Amr M. Gody #1, Tamer M. Barakat #2, Sayed A. Zaky #3
Electrical Engineering, Faculty of Engineering
Fayoum University, Egypt

Abstract: *In this research, speech recognition is implemented for English Tri-Phone unit recognition. Newly developed features called Best Tree Encoding [1] is used to define the recognition parameters. HMM is used as recognition engine. To verify the proposed model, all results are compared against the most popular features used in similar approaches in ASR the Mel Frequency Cepstral Coefficients (MFCC). The most popular HMM tool kit the HTK is used for designing and implementing HMM. The most popular corpus database the TIMIT database is used in all experiments through this research. The proposed model gives success rate for tri phone recognition94%with respect to the success rate of MFCC for tri phone recognition of the same samples. The D and A may be ignored in both MFCC and BTE.*

Keywords: *Automatic Speech recognition, English Phone Recognition, Wavelet packets, Mel scale, MFCC, HTK and Best Tree Encoding.*

I. INTRODUCTION

Speech is a kind of pressure waves called acoustic waves which have features that could be used to recognize a certain speaker or word. The mathematical analysis or processing of speech signal may be sub-divided into main classes which are speech encoding, speech synthesis and speech recognition. Speech recognition, the field of this thesis, is the process to convert speech signal to a sequence of words and when this process is done by a computer it is called Automatic Speech Recognition (ASR). The object of automatic speech recognition systems is recognizing a certain language unit, like a word or sub-word unit. Syllable, tri phone and mono phone (single phone) are examples of sub-word language unit, in this paper all experiments were made to recognize tri phone only. ASR systems have many applications which could make our life easier such as speech to text dictation, automatic route and automatic translation. However, it is a complex field which related to many fields of science, mathematics, physics, engineering and linguistics.

Automatic speech recognition system consists of three main stages. The first stage is the pre-processing which prepares the speech signal for the next stage (feature extraction stage) so its functions are dependent on the used feature extraction approach. The feature extraction stage processes the required speech signal to extract the acoustic features which describe a certain utterance. The acoustic features are represented by numerical values in the form of a vector called feature vector. The feature vectors must be able to differentiate between different classes while being unaffected by any environmental conditions where the performance of ASR system is based on how these vectors are discriminated. There are many feature extraction approaches used, such as Mel Frequency Cepstral Coefficients (MFCC), the most popular approach; the discrete wavelet transforms (DWT) and the linear predictive coding (LPC).The final stage is the classification stage which uses a generative approach called Hidden Markov Model (HMM) to find the joint probability distribution over a given observations and classes. Then it used to predict the output for a new input. The Hidden Markov Model (HMM) is a statistical approach that widely used for automatic speech recognition modeling. In HMM the system is assumed to be a Markov process with unknown parameters, and it required to find these unknown parameters, from the observations. Then the extracted parameters can be used to perform further analysis, for example for speech recognition. Hence, by using Hidden Markov model the probability of generating a certain utterance could be obtained.

As mentioned before, MFCC feature extraction technique is the most popular one used in speech recognition systems. There are numerous works that used MFCC for tri phone recognition, for example the work in [2] which apply MFCC and HMM for an English database (TIMIT) to find the baseline values of MFCC which reach about 70% without language model. However, MFCC technique has a main issue that the performance is decreased as the environment became noisy, so there is a need to other techniques to increase the speech recognition system robustness.

New feature using wavelet Packets Best Tree Encoding (BTE) [1] is intended to enhance the efficiency and robustness of Automatic Speech Recognition (ASR) by providing human like processing of speech stream and by moving the problem to another domain which can be utilized for solving the recognition problem. To improve and increase BTE performance various BTE models were designed. The model in [3] is a mono phone based recognition model using 4- points BTE features vector, where each component is a 7-bit component. The previous model resolution was increased by adding a new level [4] and encoding the speech information into 15 bits. By increasing the number of levels to 7 and encode speech information to 64 bits another phase was created [5]. A hybrid model based on Mel frequency and BTE is introduced [6] as a new entropy function to increase the amount of information in the tree leaves. The work in [10] improves the performance of previous BTE increased resolution versions by applying the Mel-based entropy using a hybrid model for mono phone recognition. The previous BTE works concern only context-independent phone recognition, this paper try to enhance the performance of the speech recognition system by using BTE for English tri-phone recognition. To evaluate the performance of the proposed model a sample of an English database TIMIT is used. TIMIT speech corpus consists of 6300 utterances from 630 different speakers of American English (70% of them are female & 30% male).

All results are verified against MFCC where it is the most common feature extraction technique used. First of all MFCC is used to obtain the reference results for the available sample of the TIMIT speech corpus. Then experiments are run to get the results for the proposed features and models. The results are provided as comparison to the reference MFCC baseline values [2] on TIMIT. The acceleration (A) and Delta (D) coefficients also increased the performance measure. The proposed approach (BMDE) for tri phone recognition achieves 64.22 % by adding "D" plus "A" qualifiers. When applied the same conditions on the MFCC and using TIMIT database, the recognition accuracy of MFCC achieves 68.56%, which a percentage success rate about 94% with reference to MFCC.

Section 2 gives an overview for best tree encoding (BTE) and band mapped. Section 3 is a detailed description of the proposed model. Section 4 explains HMM design for tri-phone recognition. Section 5presents the experiments procedures and steps; section 6discuss results and finally section 7 gives the overall conclusion.

II. BEST TREE ENCODING OVERVIEW

There are many applications based on the Wavelet Transform, such as speech recognition applications. The wavelet transform representation basically consists of the decomposition of the represented signals using small waves called wavelets. The wavelet transform could be used for processing many types of signal efficiently, where it maps the signal energy into a group of coefficients and the processed signal could be reconstructed perfectly from these coefficients without losing most of the signal features. To produce a speech signal many transformations occurred at different stages: semantics, linguistics, articulators and acoustics. So, any changes in these transformations will change the speech signal acoustic features. The challenge in speech recognition systems is to find a well adapted representation for extracting speech signals content. To find a better signal representation that has the ability to classify the different signals into different classes the signal is transformed to another domain, and where the speech signal frequency is vary with time, see figure 1, so transformation like Fourier transform and Short Fourier transform cannot be used in this case where too much signal information will be lost.

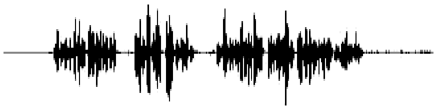


Fig.1: Speech signal

So there was a need for another transformation that could be used with speech signal. The wavelet transform (WT) is that transformation which represents the signal in a time-frequency domain, and so it has the ability to represent different parts of the signal at different scales. The continuous one dimensional WT decomposes the speech signal *f* (t) into a group of basis function (wavelets) $\Psi_{a,b}(t)$. These wavelets are generated from a single mother wavelet Ψ (t) by dilation and translation.

Amplitude
Time
Wavelet Transform
Scale
Time

Fig.2: Wavelet transform

$$W (a, b) = \int f(t)\, \Psi^{*}_{a,b} (t)\, dt \qquad (1)$$

$$\Psi_{a,b} (t) = 1/\sqrt{a}\ \Psi ((t-b)/a) \qquad (2)$$

Where: f (t) is the signal to be analyzed, (a) is the scale factor, and (b) is the translation factor. W (t) is the mother wavelet. To decompose the speech signal filters of different cutoff frequencies are used. There are two types of wavelet transform the continuous wavelet transform and the discrete wavelet transform. The continuous WT operates at every scale with shifting the analyzing wavelet over the full domain of the analyzed function. In the discrete wavelet transform, scales and positions of powers of two are chosen. For example, a speech signal (S) of length (N), therefore the discrete wavelet transform has a number of stages equal (log2 N). First, the signal (S) is decomposed to approximation coefficient A1 and the detail coefficient D1, see figure 3, by convolving the signal (S) with the function of a low pass filter to obtain the approximation coefficient, and with a high pass filter for detail coefficient, then a dyadic decimation is applied. The next step is decomposing the approximation coefficient A1 into two coefficients following the same way.

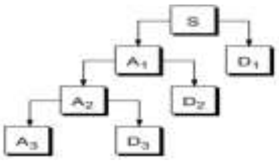


Fig.3: Wavelet decomposition

Wavelet transform is effective in case of short-time phenomena or discontinuities detection, however a lot of information is lost due to not all the signal is decomposed. Therefore, for BTE to be lossless it based on another decomposition scheme called wavelet packet decomposition, where the details as well as the approximations are decomposed, see figure 4.

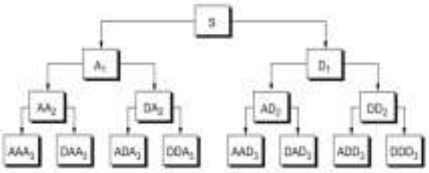


Fig.4: Wavelet packet decomposition

For speech signal not all the generated wavelet tree leaves have useful information so another criteria called entropy function is used to obtain a best tree which has a reduced number of leaves, hence the analysis time and complexity are reduced, see figure 5.

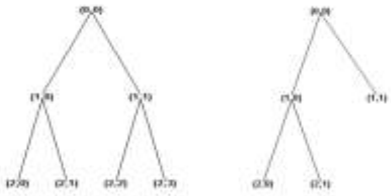


Fig.5: Two-level wavelet packet and best tree

A. Best Tree Encoding Process Description

BTE was first introduced in [1], the BTE first generation, as novel features for automatic speech recognition. Figure 6 shows a block diagram of BTE feature extraction process. The input speech signal is framed, using Hamming window, into frames of length 20 ms, then the frame is decomposed into wave packet tree which consists of a number of levels. Each level contains many nodes. Each node gives the contribution of the signal power into certain frequency band. The proper entropy function is applied to obtain the optimal nodes that contribute the most signal information. The obtained nodes collection is called the best tree. Shannon entropy is used in this approach to obtain the best tree. The best tree is then encoded to construct the features vector which is called Best Tree Encoding or simply BTE. This encoder generates fixed vector size of 4 components that can describe the best tree nodes positions.

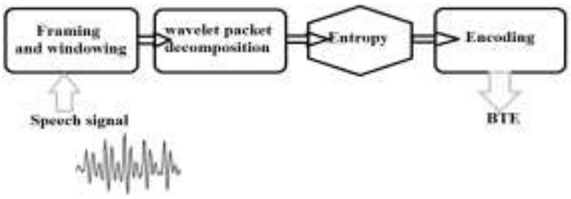


Fig.6: BTE block diagram

Figure 7 illustrates the frequency relation of BTE components as well as the encoding strategy.

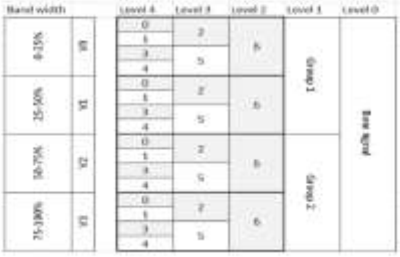
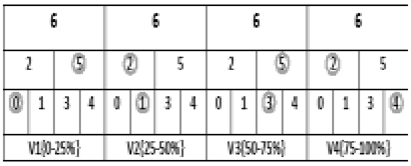


Fig.7: Wave packet tree for 4 points encoding [1]

The leave nodes encoded based on its location to a 7 bit value figure 8 shows an example for encoding a frame of speech signal.



6				6				6				6			
2		⑤		②		5		2		⑤		②		5	
⓪	1	3	4	0	①	3	4	0	1	③	4	0	1	3	④
V1{0-25%}				V2{25-50%}				V3{50-75%}				V4{75-100%]			

Fig.8: Best tree 4 point encoding example

Circled numbers in figure 8 represents leave nodes in the best tree decomposition, which will be encoded into features vector of 4 elements as shown in table 1. The numbers in Figure 8 is the bit order. The band width is divided into 4 groups. Each group is 7 bits. Each bit represents node position in the complete wavelet packets tree. Bits are ordered in such that to keep the distance between two components minimum according to the frequency adjacency.

Table 1
Best tree 4 point encoding evaluation

Element	Binary value	Decimal value	Frequency band
V1	0100001	33	0-25%
V2	0000110	6	25-50%
V3	0101000	40	50-75%
V4	0010100	20	75-100%

Features vector for this example, in figure 8, will be:

$$F= [33; 6; 40; 20] \qquad (3)$$

The second generation of BTE was developed in [7], where the decomposing level increased to 5 levels to increase the discrimination of the features vector, the third generation BTE7 was developed in [8], by increasing the decomposing levels to 7 levels and uses the Log energy entropy instead of Shannon's. The key parameters in BTE development are manly Resolution and Entropy. The resolution is intending to increase the encoder output resolution. This will affect the size of the number that representing each BTE vector component. This in turn will increase the discrimination ability of BTE. The other factor is the Entropy function. This factor is affecting the structure of the Best tree itself. Obtaining better structure that best fit the variations between the recognized speech units is the target of this factor. Another factor affect the performance of the system is the sampling rate used where the speech inside the ear is affected by the frequencies of the pine, conch, and the outer ear canal, which produces a broad peak of 15-20 dB at 2500 Hz and spreading relatively uniformly from 2000-7000 Hz. The most important issue here is the non-linear behavior of the ear, where the middle ear transfer function is not uniform but has a peak at 1000 Hz and gradually drops off to about 20 dB below peak level at 100 Hz and 10,000 Hz. So, the frequency range containing speech information is from 300 Hz to 7000 Hz where the variation is less than 10 dB.

B. PRE-EVENT PROCESSING

Before starting BTE process, the down sampling should be considered in order to scale the speech signal into the proper analysis band. The signal is properly rescaled into the frequency band in order that to be moved within the 10 (KHZ) band. This band is covering the human speech. So in the pre event processing two factors are included, frequency mapping to 10 KHz and speech normalization.

III. PROPOSED MODEL

The proposed model Band Mapped Distributed Energy BTE (BMDE-BTE) decrease the non useful information in the sampled speech signal by using a sampling rate 10 KHz.

To increase the feature vector discrimination the BMDE-BTE model add more information about the best tree leaves to the vector, many kind of information could be used to represent the leaves , this research use the leaves energy as extra information appended to the feature vector. As done for BTE4 the bandwidth divided into four sub- bands “V1, V2, V3 and V4”. Energy elements of each best tree node will be appended to the feature vector so the feature vector becomes:

F=[P1;E01;E11;E21;E31;E41;E51;E61;P2;E02;E12;……Pi; Eji] (4)

Where:

Pi component represents the node position code; Eji component represents the leaf energy (distributed energy) code; i the bandwidth sub-band number (i=1,…4), and j the best tree leaf number (j=0,1,…..6). The output is a 32 components feature vector and where each node in each level has different energy then the vector become more discriminated.

BTE previous work was focused on mono phone recognition only, here the proposed model adapted for tri phone recognition.

IV. HMM DESIGN FOR TRI-PHONE RECOGNITION

Hidden Markov Model is a statistical model that consists of observations sequence. When phoneme based HMMs are being used, they must be concatenated to construct word or phrase HMMs. For example, an HMM for `cat' can be constructed from the phoneme HMMs for /k/ /a/ and /t/, then `cat' HMM model will has nine states. The main problem in the phoneme based models that they do not take into account any contextual variation in phoneme production, to solve this problem a larger sub-word units should be used or using models that depend on the context. Tri-phone models are the most common approach to solve this issue, in tri-phone models there is one different phoneme model for every different left and right phoneme context, so there are distinct models for the /ai/ in /k-ai-t/ and in /h-ai-t/. Now, a word model is made up from the appropriate context dependant tri-phone models: 'cat' would be made up from the three models [/sil-k-a/ /k-a-t/ /a-t-sil/]. See figure 9.

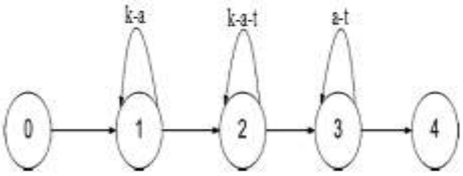


Fig.9: Tri-phone model for the English word Cat

Tri-phone recognition solves the problem of context sensitivity but another problem is presented, there is no enough data to train all tri-phone models (for 45 English phonemes there are 91125 possible tri-phone models). One approach is using state tying where only word internal tri-phones are used instead of the cross word tri-phones. The difference between the cross word tri-phones and the internal tri-phones is it captures co-articulation effects across word boundaries so it used for continuous speech recognition, but the word internal tri-phone model uses tri-phones only for word internal triples and di-phones for word final phonemes; `cat' would become: [sil /k-a/ /k-a-t/ /a-t/ sil] and that will be less accurate for continuous speech recognition.

The database used is sample of TIMIT, TIMIT is an English database consists of 6300 utterances from 630 different speakers of American English (70% of them are female & 30% male), recorded with a high-fidelity microphone in a noise-free environment. The sample database used consists of 160 audio files divided into two sets for training and testing, the total number of unique triphones in the training database used is 3277 triphones, 1033 triphones appear less 5 times, see figure 10, and where the low frequency triphones represents a problem due to there are no enough training data and so it may be assigned to a wrong state. Therefore, the triphones which appeared more than 16 times (more than 70%) in the training database were modeled.

V. EXPERIMENTS PROCEDURE AND STEPS

A. *Experiment Procedures*

In the experiments reported in this section, continuous word recognition experiments were performed using sample of English database (TIMIT). The database is classified into two sets one for training the Hidden Markov Model (HMM) models using HTK toolkit and the other for testing.

Experiments are performed to obtain comparison results of the proposed BTE in comparison with the previous versions of BTE and with respect to the reference features MFCC. Finally, the above experiments repeated with adding Dleta and Acceleration coefficients to the features vector. In is also included the effect of the Gaussian Mixture Model (GMM)

B. *Experiment Steps*

Figure 10 represents a block diagram for the experiments steps which done with C# interface Speech Tool Kit (STK) [8] and HTK tools. STK uses the HTK tools as an engine to execute the recognition tasks, and especially for controlling the configuration of the Hidden Markov Model Toolkit tools. Also, STK includes the Matlab functions and codes that are used for generating the BTE feature. The experiment steps are divided into two main classes, the first one to create HMM models for mono phones where the HTK tools use mono phone as basis for tri phone recognition. The second class is creating tri phone from mono phone to do recognition task [9].

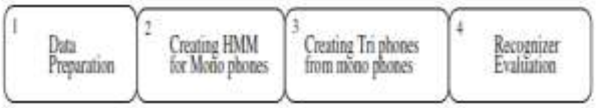

1
Data Preparation
2
Creating HMM for Mono phones
3
Creating Tri phones from mono phones
4
Recognizer Evaluation

Fig.10: Experiment steps block diagram

VI. RESULTS AND DISCUTION

This section report the results of all experiments and discus these results to evaluate the new BTE algorithm for English tri phone recognition. Here, all results verified against the most popular feature extraction technique MFCC to estimate the performance of the proposed model using the percentage success rate (SR %) with reference to MFCC, which defined as:

$$SR = BTE\ SR / MFCC\ SR \qquad (4)$$

A. *MFCC baseline values*

This section presents the appropriate values of MFCC feature extraction technique. Figure 11 shows that the success rate for mono phone recognition using MFCC with multi GMM count and using TIMIT database is about 57% and the same for Tri phone is about 70%.

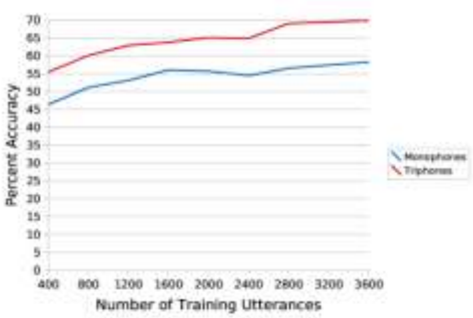


Fig.11: MFCC success rate with 10 GMM [2]

For the sample TIMIT used in experiment MFCC gives tri-phone recognition success rate about 68% which considered as reference for the BTE success rate.

B. *Comparative parameters*

This discussion and evaluation is based on some parameters used to compare the performance of each feature extraction approach:

I. Gaussian Mixture Model (GMM)

GMM is the number of Gaussian mixtures used in the emitting states of HMM. The value of GMM ranges between the values (2, 4).

II. Frame size

As a first stage of ASR system the input speech signal is sampled to frames with a constant size. For speech recognition application the frame size has values 10-30 ms to represent the signal as a stationary signal and reduce the effect of speech variability. Here, the frame size assigned values 20 and 25 ms.

III. Entropy Function

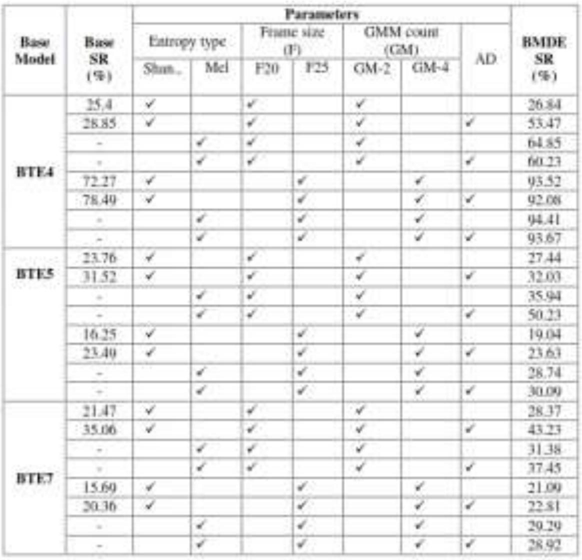
Two Entropy functions were used for all experiments Shannon and Mel based entropy, for more details see [6]. The Shannon entropy provides a way to estimate the average minimum number of bits needed to encode a string of symbols, based on the frequency of the symbols.

IV. Qualifiers

As the feature vector become distinctive the feature extraction technique produce better results, so a number of qualifiers added to increase the difference between vectors. Here, two qualifiers were used, Delta and Acceleration [8].

C. *English tri-phone recognition results*

Figure 11 summarize all experiments results for different comparative parameters with reference to the MFCC baseline values and will be discussed in the following:



Base Model	Base SR (%)	Parameters							BMDE SR (%)
		Entropy type		Frame size (F)		GMM count (GM)		AD	
		Shan..	Mel	F20	F25	GM-2	GM-4		
BTE4	25.4	✓		✓		✓			26.84
	28.85	✓		✓		✓		✓	53.47
	-		✓	✓		✓			64.85
	-		✓	✓		✓		✓	60.23
	72.27	✓			✓		✓		93.52
	78.49	✓			✓		✓	✓	92.08
	-		✓		✓		✓		94.41
	-		✓		✓		✓	✓	93.67
BTE5	23.76	✓		✓		✓			27.44
	31.52	✓		✓		✓		✓	32.03
	-		✓	✓		✓			35.94
	-		✓	✓		✓		✓	50.23
	16.25	✓			✓		✓		19.04
	23.49	✓			✓		✓	✓	23.63
	-		✓		✓		✓		28.74
	-		✓		✓		✓	✓	30.09
BTE7	21.47	✓		✓		✓			28.37
	35.06	✓		✓		✓		✓	43.23
	-		✓	✓		✓			31.38
	-		✓	✓		✓		✓	37.45
	15.69	✓			✓		✓		21.09
	20.36	✓			✓		✓	✓	22.81
	-		✓		✓		✓		29.29
	-		✓		✓		✓	✓	28.92

Fig.11: Experiments result

I.GMM=2 and Frame size=20 ms

From Figure 13, encoding best tree (BT) using distributed energy and Mel based entropy gives the best results (64.85%). By increasing the decomposition level (BTE5, BTE7) the process takes more time than BTE4 and the results not enhanced.

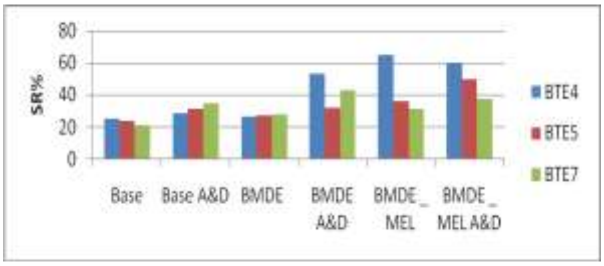


Fig.13: Results for tri phone recognition (GMM=2& F=20ms)

II. GMM=4 and Frame size=25 ms

Increasing the frame size and GMM values and using Mel-based entropy enhance the success rate too much where it become close to the MFCC values (94.41%).As in the previous section, the success rate not increased by increasing the decomposition level, figure 14.

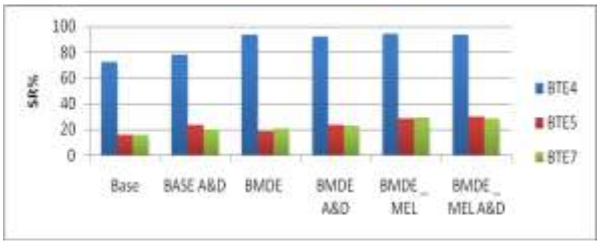


Fig.14: Results for tri phone recognition (GMM=4& F=25ms)

VI. CONCLUSION

Encoding the wavelet best tree based on leaves energy and position give a feature vector consists of 32 components which increase vector discrimination. The entropy function based on Mel increases the process speed and accuracy where a lot of non useful information outside the human perception was omitted. By increasing GMM to be 4 and frame size to 25 ms force BTE4 to a new era of tri phone recognition where its result with reference to MFCC is about94%.

REFERENCES

[1] Amr M. Gody, "*Wavelet Packets Best Tree 4 Points Encoded (BTE) Features*", The Eighth Conference on Language Engineering, Ain-Shams University, Cairo, Egypt,PP 189-198, December 2008.

[2] Barnard, E, Gouws, E, Wolvaardt, K and Kleynhans, N. "*Appropriate baseline values for HMM-based speech recognition*". 15th Annual Symposium of the Pattern Recognition Association of South Africa, Grabouw, South Africa, November 2004.

[3] Amr M. Gody, Rania Ahmed AbulSeoud,Mohamed Hassan "*Automatic Speech Annotation Using HMM based on Best Tree Encoding (BTE) Feature"*, The Eleventh Conference on Language Engineering, Ain-Shams University, Cairo, Egypt PP. 153-159 ,December 2011.

[4] Amr M. Gody, Rania Ahmed AbulSeoud,Maha M. Adham, Eslam E. Elmaghraby "*Automatic Speech Using Wavelet Packets Increased Resolution Best Tree Encoding"*, The Twelfth Conference on Language Engineering, Ain-Shams University, Cairo, Egypt PP. 126-134, December 2012.

[5] Amr M. Gody, Rania Ahmed AbulSeoud,Eslam E. Elmaghraby "*Automatic Speech Recognition Of Arabic Phones Using Optimal- Depth – Split –EnergyBesttree Encoding*", The Twelfth Conference on Language Engineering, Ain-Shams University, PP. 144-156, December 2012, Cairo, Egypt.

[6] Amr M. Gody, Rania Ahmed AbulSeoud, Mai Ezz El-Din,"Using Mel-Mapped Best Tree Encoding for Baseline-Context-Independent-Mono-Phone Automatic Speech Recognition" Ain-Shams journal,2015.

[7] Maha M. Adham, "Phone Level Speech Segmentation Using Wavelet Packets*"*, Fayoum University, 2013.

[8] Eslam E. Elmaghraby "Enhancement Speed Of Large Vocabulary Speech Recognition System", Fayoum University, 2013.

[9] HTK Book documentation, "http://htk.eng.cam.ac.uk/docs/docs.shtml".

[10] Amr M. Gody, Rania Ahmed Abul Seoud ,Marian M.Ibraheem , " Hybrid Model design for Baseline-Context-Independent-Mono-Phone Automatic Speech Recognition", International Journal of Engineering Trends and Technology (IJETT) – Volume 27 Issue :2231-5381- September 2015