Hybrid Model Design for Baseline-Context-Independent-Mono-Phone Automatic Speech Recognition

Amr M. Gody [#1], Rania Ahmed AbulSeoud [*2], Marian M.Ibraheem [#3]
Electrical Engineering, Faculty of Engineering
Fayoum University, Egypt

Abstract: *In this research new hybrid model for Automatic Speech Recognition is introduced. The model is constructed as a hybrid of Mel-Scale and 15-Bit Best Tree Encoding (BTE). Best Tree Encoding is first introduced in [1] as new feature for solving Automatic Speech recognition (ASR). The model is compared with MFCC. HTK is used as Recognition Engine. The model is also compared with the old generations of BTE to evaluate the performance of Context-independent mono-phone recognition of English language. Sub class of TIMIT database is used in all experiments through this research .The proposed model gives success rate equals to 96%with respect to the success rate of the reference MFCC in solving the same problem but vector size is 33% of MFCC vector size.*

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

INTRODUCTION

People interact together using speech as a natural mean that doesn't require any specific tool or learning. Speech is skill we gain at the early stages of age, baby learn quickly to respond to his mother voice and to make a noise when necessary. Speech may be considered as a spontaneous action where speaking to somebody doesn't need any determined tasks and many tasks can be done during speaking.

Automatic Speech Recognition (ASR) is a system that makes a computerized framework able to recognize the speech of a person through any transducer like a microphone. ASR achieved to convert a speech signal into a sequence of words. Also, it aims to develop techniques that enable computers to accept speech input, which makes human-computers interface easily and in effective way. There are many approaches to model an ASR system, but the most common used approach is the statistical model that based on the Hidden Markov Model (HMM). Statistical approach using HMM is basically recognizing the speech by determining the probability of each phoneme at continuous frames of the speech signal. There are numerous speech recognition toolkits. HTK, Sphinx, CSLU and Julius are examples of these toolkits which are widely used for research purposes.

Many previous researches used the same TIMIT speech corpus ,in [15]the accuracy of MFCC for mono phone data is 50.7% using single mixture recognizers ,where reaches to 53.2% by using 10 mixtures recognizers. Many previous researches used the same TIMIT speech corpus ,in [15]the accuracy of MFCC for mono phone data is 50.7% using single mixture recognizers ,where reaches to 53.2% by using 10 mixtures recognizers. The work in [18] propose a large-vocabulary, speaker-independent, continuous speech recognition system. The system is based on hidden Markov modeling (HMM) using MFCC, The system has been evaluated on the TIMIT database. The system accuracy is 60.1% without grammar.

Another research [19] is a survey to provide baseline results for the TIMIT phone recognition using many techniques. One of these approaches based on MFCC give a result of 72.2% for phone recognition.

In this paper, all tests used the TIMIT speech corpus. TIMIT 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 recognition accuracy of MFCC for mono-phone data ranging from 43.70% without any qualifiers to 52.18% by adding the Delta "D" and Acceleration "A" qualifiers.

In this paper, the experiments are implemented using the open-source toolkit HTK and were restricted to English-language TIMIT. Results are compared with MFCC as it is the most popular feature model used in ASR. MFCC is used to generate the reference results for the available samples from TIMIT speech corpus. Then experiments are run to get the results for the proposed features and models. The results are presented as comparison to the reference MFCC results on TIMIT.

Best Tree Encoding (BTE) is new features developed for ASR. The key of these features is moving the ASR problem to new space where speech units can be

effectively discriminated. Many phases of enhancement for BTE are done to enhance the efficiency. In [3] is introduced a completely automated phone recognition system based on Best Tree Encoding (BTE) 4-point speech feature. The System identified spoken phone at 45.7% with respect to MFCC for solving the same problem using the same samples. Another BTE enhancement in [4] is introduced. In [4] Information related to speech phoneme is encoded into 15 bits instead of 7 bits in the original version of Best Tree Encoding (BTE4). This is achieved by introducing one more analysis level in the wavelet decomposition process that constructs the core of BTE. Five levels in wavelet decomposition stage instead of 4 levels is implemented to extract the 5 level BTE or simply BTE5. BTE5 is intended to test the effect of increased resolution on the recognition efficiency. Another phase of BTE development is introduced in [5]. It aims to enhance BTE encoder by adding two more factors to the encoder. The first factor is the increased resolution by adding more Analysis levels in wavelet packets to become 7 and the second factor is to add the Energy components to the features vector. The Energy is also split into 4 components instead of single component in the first version of BTE to enhance the discrimination of speech units into the new features space.

In this research, Section two provides review on the development of BTE because it is new features. Section three introduces the proposed hybrid model. Section four provides analysis of HMM models that are used in this research as well as HTK procedures and demonstrates the experiments parameters. Finally the discussion and conclusion on the obtained results.

I. Model Best Tree Encoding (BTE)

A. Best Tree Encoding Diagram

In this section, Best Tree Encoding feature is illustrated. Wavelet Packets Best Tree Encoded feature (BTE) was first introduced by Amr M. Gody in [3].The procedure of extracting BTE will be illustrated through the block diagram in Fig 1.

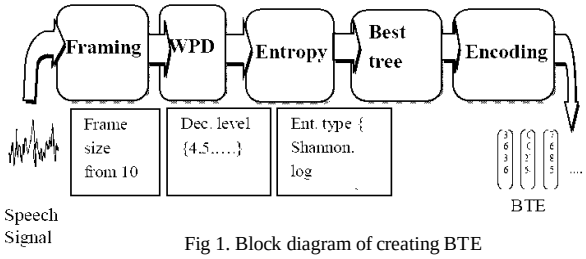

Fig 1. Block diagram of creating BTE

As shown in Fig 1 the process of creating BTE file starts with converting the speech stream into short time duration frames. The second step is the preprocessing phase. Wavelet packet decomposition (WPD) is used in the preprocessing phase as shown in Fig 1 by WPD Block. The next step is to select the proper entropy type. In this step only tree nodes with sufficient information will be kept but the others will be cut from the binary tree that constructs the wavelet packet frequency bands. The last step is to encode the tree structure into 4-Dimensional vector of integer values [1]. Here is below quick description of each block in Fig 1;

1) Framing

It is the process of segmenting the speech signal into small duration scaled frames in order to deal with it as stationary signal as shown in Fig 2. Frame length is most likely chosen from 10 to 30(ms). In this paper the experiments were made firstly using frame length of 20 (ms) then using 25(ms).

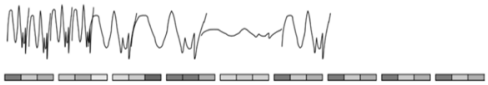

Fig 2. Speech signal is segmented into sequence of features frames [13]

2) Wavelet Packet Decomposition (WPD):

It is a one-dimensional wavelet packet analysis function, which returns wavelet packet tree corresponding to the wavelet packet decomposition of the vector X at level for example 4, with a particular wavelet. It is implemented using MatLab [18].

```
t= wpdec(X, 4,'db4', 'Shannon')   1
```

3) Entropy

Entropy provides a complexity measure of a time series, such as discretized speech signal. Entropy is the key step to enhance BTE. Entropy is used to measure information in each tree node in Fig 3. Accordingly the best tree is decided by removing all low informative tree nodes. In MatLab; there are various popular types of entropy as Shannon, log energy, threshold, sure and norm. In this paper Shannon entropy is chosen as given in equation 1.

4) Best Tree Encoding

The Best tree function [9] utilizes the entropy to evaluate the low information tree nodes. Best tree selection model is detailed in [9]. Simply starting from the higher level tree nodes, each 2 nodes have one parent node. If the entropy of the parent node is higher than the sum of entropies of both Childs, then Childs will be removed. The process is recursively continues till end with the best tree. Keep in mind

that each tree node is representing single frequency band. The component at each node is the signal projection on the underlying frequency band. This is the key in Best tree Encoding.

5) Encoding

The last step is the Encoding process. In this step, the obtained best tree in step 5 is encoded into 4 component features vector. Each component represents quarter of the band width of the signal. Each component can be used to recall the best tree nodes that fall into the corresponding quarter that is represented by the component. In section 2, quick description of BTE generations will be provided to clarify the encoding.

B. Best Tree Encoding Generations

In this section, the previous generations of BTE will be introduced. The following subsections provide quick abstract of each of the elder revision of BTE.

1) BTE4

BTE4 was first introduced in [1]. It has four wavelet decomposition levels. As mentioned in 2.1, BTE components each is representing quarter of the bandwidth as shown in table 1. V1 is the component that represents the first quarter of the bandwidth; V2 is the component that represents the second quarter of the bandwidth and so on. Each cell in table 1 represents Tree Node of wavelet packet decomposition. For example Cell 0 is representing the base signal. Cell 1 represents the signal at the frequency band starts from 0 and ends at half of the bandwidth of the signal. Cell 15 represents the projection of the signal in the frequency band starts from 0 and ends with $\frac{1}{16}$ of the bandwidth of the base signal. The encoding can be illustrated through table 2. Each cell in table 2 represents binary digit into 7 bit number. If tree node exists the corresponding bit is 1 else the corresponding bit is 0. Each group of 7-bits construct vector component as shown in table 2.The first group at the top is used to construct V1. The second group from the top is used to construct V2 and so on.

The second point in the encoding that should be focused on is the bit ordering-. Bits in each group are ordered in such that nodes with adjacent frequency bands are adjacent orders. This is to ensure using Euclidian distance will be frequency biased.

TABLE 1
CLUSTERING CHART TO EXPLAIN THE 4 POINTS ENCODING ALGORITHM BEFORE ARRANGEMENT

	Level 4	Level 3	Level 2	Level 1	Level 0
V_1	15	7	3	1	0
	16				
	17	8			
	18				
V_2	19	9	4		
	20				
	21	10			
	22				
V_3	23	11	5	2	
	24				
	25	12			
	26				
V_4	27	13	6		
	28				
	29	14			
	30				

TABLE 2
CLUSTERING CHART TO EXPLAIN THE 4 POINTS ENCODING ALGORITHM AFTER ARRANGEMENT

	Level 4	Level 3	Level 2	Level 1	Level 0
V_1	0	2	6	Low band	Base signal
	1				
	3	5			
	4				
V_2	0	2	6		
	1				
	3	5			
	4				
V_3	0	2	6	High band	
	1				
	3	5			
	4				
V_4	0	2	6		
	1				
	3	5			
	4				

2) Increased Resolution BTE

The key factor in this version of BTE is the increased resolution of the encoder. Adding more decomposition level to the wavelet packet is the methodology of this version of BTE. Two versions of BTE are developed BTE5 and BTE7. BTE5 was introduced in [9] and it is considered as being the second generation of BTE4 by adding a new level of decomposition to increase information resolution. The strategy of encoding the tree nodes in BTE5 is the same as of BTE4.

It is also 4-Dimensional vector; "Shannon Entropy" is used for extracting the best tree. Adding the new analysis level makes that, the information encoder is

15 bits instead of 7 bits in BTE4.Table 3 and table 4 explains the encoding process of BTE5. The key enhancement factor in this generation is increasing the resolution to enhance the discrimination between the different classes

TABLE 3
CLUSTERING CHART TO EXPLAIN THE 5 POINTS ENCODING ALGORITHM BEFORE ARRANGEMENT

	L5	L4	L3	L2	L1	L0
V1	31	15	7	3	1	0
	32					
	33	16				
	34					
	35	17	8			
	36					
	37	18				
	38					
V2	39	19	9	4		
	40					
	41	20				
	42					
	43	21	10			
	44					
	45	22				
	46					
V3	47	23	11	5	2	
	48					
	49	24				
	50					
	51	25	12			
	52					
	53	26				
	54					
V4	55	27	13	6		
	56					
	57	28				
	58					
	59	29	14			
	60					
	61	30				
	62					

TABLE 4
CLUSTERING CHART TO EXPLAIN THE 5 POINTS ENCODING ALGORITHM AFTER ARRANGEMENT

	L5	L4	L3	L2	L1	L0
V1	0	2	6	14	Low Signal	Base signal
	1					
	3	5				
	4					
	7	9	13			
	8					
	10	12				
	11					
V2	0	2	6	14		
	1					
	3	5				
	4					
	7	9	13			
	8					
	10	12				
	11					
V3	0	2	6	14	High signal	
	1					
	3	5				
	4					
	7	9	13			
	8					
	10	12				
	11					
V4	0	2	6	14		
	1					
	3	5				
	4					
	7	9	13			
	8					
	10	12				
	11					

3) BTE7

BTE7 is formed by adding one more decomposition level to the wavelet packet analysis. Adding the new analysis level makes that information encoder is 63 bits. In addition to increased resolution, BTE7 is supported with Energy components of the clustered regions. The bandwidth is divided into 4 equal parts. Each part is a cluster. The energy components are evaluated for each cluster to form extra 4 components that are appended to the features vector. The encoder is illustrated through tables 5 and 6.

TABLE 5
CLUSTERING CHART TO EXPLAIN THE 7 POINTS ENCODING ALGORITHM BEFORE ARRANGEMENT

L7	L6	L5	L4	L3	L2	L1	L0
127	63	31	15	7	3		0
128							
129	64						
130							
131	65	32					
132							
133	66						
134							
135	67	33	16				
136							
137	68						
138							
139	69	34					
140							
141	70						
142							
143	71	35	17	8			
144							
145	72						
146							
147	73	36					
148							
149	74						
150							
151	75	37	18				
152							
153	76						
154							
155	77	38					
156							
157	78						
158							
….	….	….	…..	….	4		
….	…..	….	…..	…..	5	2	
….	….	…..	…..	…..	6		

TABLE 6
CLUSTERING CHART TO EXPLAIN THE 7 POINTS ENCODING ALGORITHM AFTER ARRANGEMENT

	L7	L6	L5	L4	L3	L2	L1	L0
V1	0	2	6	14	30	62	Low Signal	Base Signal
	1							
	3	5						
	4							
	7	9	13					
	8							
	10	12						
	11							
	15	17	21	29				
	16							
	18	20						
	19							
	22	24	28					
	23							
	25	27						
	26							
	31	33	37	45	61			
	32							
	34	36						
	35							
	38	40	44					
	39							
	41	43						
	42							
	46	48	52	60				
	47							
	49	51						
	50							
	53	55	59					
	54							
	56	58						
	57							
V2	…..	…..	…..	…..	…..	…..	62	
V3	…..	…..	…..	…..	…..	…..	62	High Signal
V4	…..	…..	…..	…..	…..	…..	62	

4) Mel-Scale BTE4

New trend is considered in the version of BTE. This version is detailed in [13].The algorithm of evaluating the
best tree is targeted in this version of BTE. Mel scale

is included to evaluate the best tree nodes. In addition to Mel-Scale; Resembling the original waveform is considered to map the bandwidth to 5(Khz).

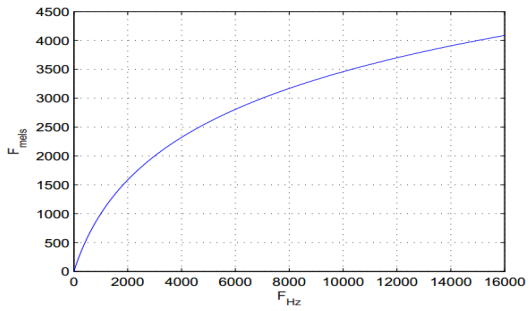

Fig 4: Relationship between the frequency scale and Mel-Scale (MS).

The formula which is used for MS (f_{Mel}) is given as following:

$$f_{Mel} = 2595 * log_{10}(1 + \frac{f_{Hz}}{700}) \quad 2$$

Where, f_{Hz} is the frequency in the hertz unit.
In this approach, weight is calculated for each node based on the position of this node on the MS curve shown in figure 4. Nodes on low frequency band will be given high weights which indicate high ability of human hearing and vice versa.

II. Hybrid BTE Model

In this paper, increasing the wavelet packet levels, considering Mel scale in best tree evaluation, Band Mapping to 5(KHz) and Energy components are mixed to constitute hybrid BTE model. The new Hybrid model is abbreviated (H-BTE).Mel Scale technique (M-BTE) was explained in more details in [13].

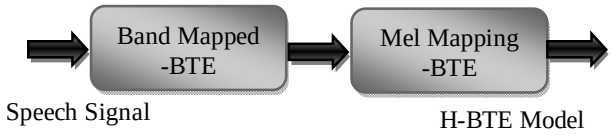

Fig 5: Block diagram implementation of the technique (H-BTE).

Fig. 5 indicates each part of the H-BTE model. First the re-sampling is applied on the input signal to Map the baseline speech signal to fill out the complete frequency band of wavelet packet analysis. Band Mapped-BTE will consider the down-sampling to 10 (kHz) of the baseband signal. In this case the signal will spread over 5 (kHz) instead of 16 (kHz). The tree before applying the BM is distributed at the bandwidth from 0 to 16 KHz. According to the MS figure 4, the human hearing can distinguish the sounds up to 4 KHz, so the nodes from 4 KHz till 16KHZ are not used. On the other hand, the tree allocated at the bandwidth from 0 to 5 KHz after applying the BM. Thus, most of the nodes in this tree are
concentrated at the bandwidth from 0 to 4 KHz which reflect the human hearing band according to the MS curve Fig 4.

After down sampling the input signal then weight is calculated for each node based on the position of this node on the MS curve Fig 4. Nodes on low frequency band will be given high weights which indicate high ability of human hearing and vice versa. Including Node weight in each node in the entropy equation is the key of Mel-Scale based tree pruning. H-BTE is combination of Band Mapped-BTE and Mel Mapping-BTE .This methodology of applying Hybrid model on BTE generations is very effective to enhance the behavior of Automatic Speech Recognition problem.

III. Design and Implementation of the HMM Recognition Engine

A. *INTRODUCTION*

As mentioned before there are many approaches to model ASR system, HMM is the most common statistical approach used in ASR. The HMM approach provides a natural and highly reliable way of recognizing speech for a wide range of applications. This section describes the experiment steps for English phone recognizer which is designed to recognize phones in continuously spoken utterance by using the HTK tool. For context independent phone recognition without grammar. At the end of this chapter the experiment results will be given in tables.

Fig 6 shows a block diagram for the experiments procedure, where MFCC is considered as a reference feature and H-BTE is the feature. HTK is used in the experiments as an engine which is well known tool kit using the same sample of TIMIT database then making analysis for both results, it will be adapted to enhance H-BTE results based on the verification against MFCC which gives a success rate about 45% for mono phone recognition for the same database as shown in fig 7.

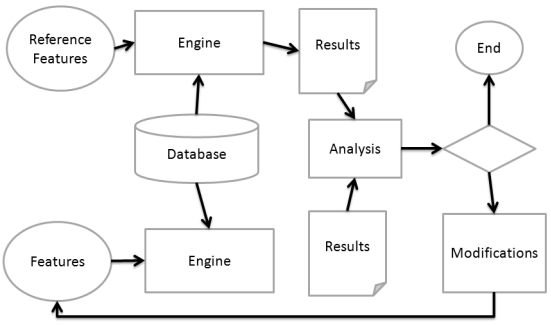

Fig 6: Experiments procedure block diagram

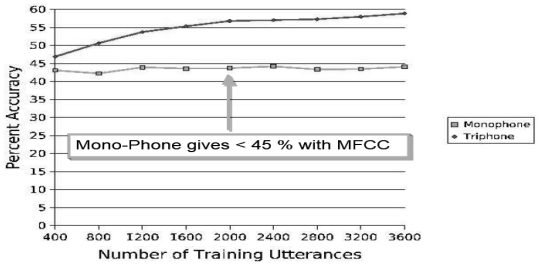

Fig 7: chart indicate MFCC results for mono phone and tri phone using TIMIT database

B. RESEARCH FRAMEWORK

In this section, the framework will be introduced through the working components, or for simplicity we can call each as worklet. Here is below the list of the worklets:

- HTK platform tools for training and testing, which is a collection of command-line options such as HERest and HVite. Each makes a special function, which is explained in detail in HTK book [15]
- Microsoft C# (C sharp) is used for building the needed logic for building initial models of HTK.

- Matlab platform is used for the feature extraction process and converting the input wave file to BTE file.

- Training and test database TIMIT, The TIMIT corpus of read speech has been designed to provide speech data for the acquisition of acoustic-phonetic knowledge and for the development and evaluation of automatic speech recognition systems.

C. THE EXPERIMENT VARIEBLES

1) Delta And Acceleration Coefficients

The performance of a speech recognition system can be enhanced by adding time derivatives to the basic static parameters. In HTK, these are indicated by attaching qualifiers to the basic parameter kind. The qualifier D indicates that first order regression coefficients (referred to as delta coefficients) are appended, the qualifier A indicates that second order regression coefficients (referred to as acceleration coefficients). The delta coefficients are computed using the regression equation given in Equation.3 as calculated in HTK book [10].

$$d_t = \frac{\sum_{\theta=1}^{\theta} \theta (C_{t+\theta} - C_{t-\theta})}{2 \sum_{\theta=1}^{\theta} \theta^2} \qquad (3)$$

Where d_t is delta coefficient at time t computed in terms of the corresponding static coefficients $C_{t+\theta}$ to $C_{t-\theta}$. The value of θ is set using the configuration parameter DELTAWINDOW (Delta window size). The same formula is applied to the delta coefficients to obtain acceleration coefficients except that in this case the window size is set by ACCWINDOW (Acceleration window size).

Adding the Delta and Acceleration terms to the feature vector will increase the vector size by the number of the vectors element for Delta and for Accelerations. If the original vector size is 4 elements after adding the energy term it will become 4x3[original 4+ Delta 4+Acceleration 4]=12 elements which increase the information of the human perception.

2) Split Energy

As discussed in the previous chapter that the Split Energy components are appended to BTE main vector components of BTE which increases the vector size by extra 4 elements. The performance of a speech recognition system can be enhanced by adding Split Energy components to the basic static parameters. These are indicated by attaching "_ES" qualifier to the basic parameter kind and when we add it the feature vector will be "BTE_ES".

3) Number Of Gaussian Mixture in HMM

We will study the effect of increasing the starting number of Gaussian Mixture (GM) on the recognition results and determining what will be the optimal number of Gaussian mixture that will give better results.

D) Procedure of the Experiments

In brief, we can say that process of building an isolated words speech recognizer using HTK tools consists of the following steps:
1. Constructing a Dictionary for the models
2. Building the Grammar (a "Language Model")
3. Creating Transcription files for Training data
4. Encoding the data (Feature Extraction/Processing)
5. (Re-)training the Acoustic Models
6. Evaluating the recognizer against the Test data
7. Reporting recognition results

Success Rate

This section gives standard execution assessment to vocabulary-free mono-phone recognition of English by utilizing a part of TIMIT database. The HMM-based recognizer was prepared with not-hand-checked information from 630 speakers (70% female and 30% male), recounting short sentences. This database contains 130 recorded wave documents which are mono-phones in wav group, 26 records utilized for testing (20% of the total).

IV. RESULTS AND DISCUSSION

By applying HTK steps on the part of TIMIT database using BTE feature with frame size 25(ms) and take into consideration the effect of Gaussian Mixture number on the success rate in the basic form and after applying the Hybrid model. Table 7 indicates all the results before using the new technique and after applying the Hybrid mode. MFCC is taken as a reference to all the results which consider MFCC result equals 100%.

$$result = \frac{Success\ rate\ of\ BTE}{Success\ rate\ of\ MFCC} \qquad (4)$$

The following chart indicates the success rate in BTE4 which clarify the effect of Gaussian Mixture number and also the different qualifiers on H-BTE4. We note that the results range from 24.4 to 85.6 related to MFCC.

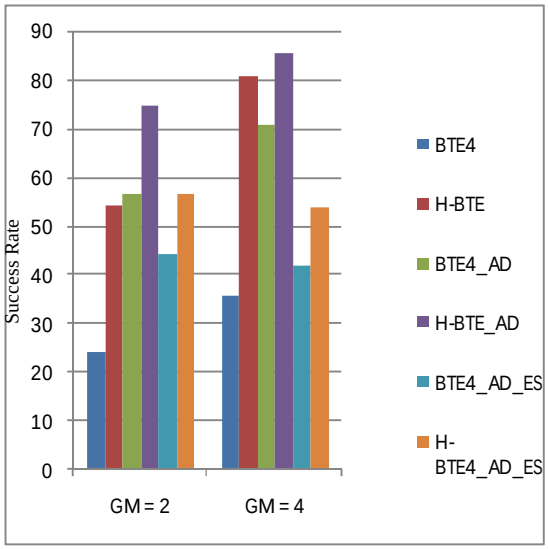

Fig 8: chart indicate the effect of GM No on the success rate of H-BTE4

starting value of GM is 4 is 76 adding the qualifiers delta and acceleration, after applying Hybrid model the best result happened at the same condition with result 94.1 the highest success rate.

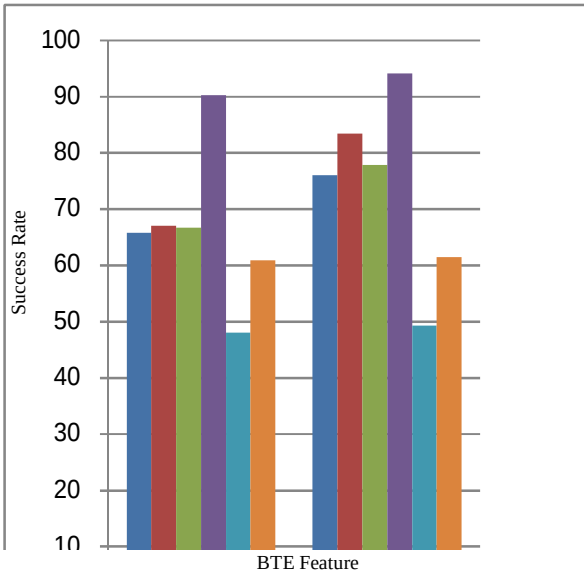

Fig 9: chart indicate the effect of GM No on the success rate of H-BTE5

In BTE7 the best value of result achieved when the starting value of GM is 4 without internal increment is 49.8 adding the qualifiers delta and acceleration, after applying Hybrid model the best result happened at the same condition with result 61.5

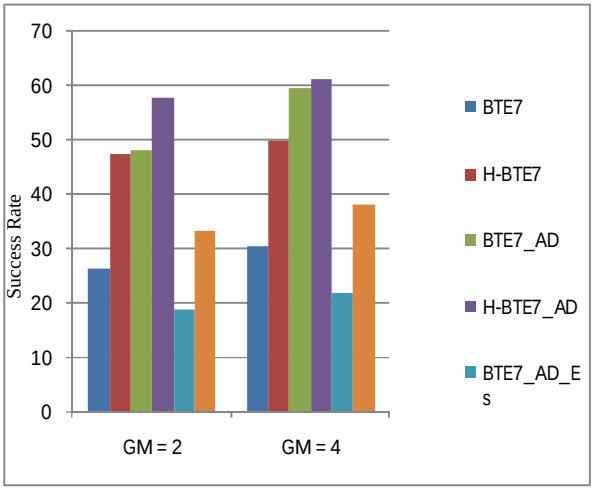

Figure 10: chart indicate the effect of GM No on the success rate of H-BTE7

TABLE 7
THE EFFECT OF GAUSSIAN MIXTURE NO ON THE BTE GENERATIONS BEFOR AND AFTER APPLYING HYBRID MODEL

Base Model	Vector size	Base SR (%)	Hybrid model				SR of the Hybrid model (%)
			AD	SE	GM-2	GM-4	
BTE4	4	24.4			✓		36.1
BTE4	4	54.6				✓	81.2
BTE4	12	56.7	✓		✓		71.2
BTE4	12	75	✓			✓	85.6
BTE4	15	44.3	✓	✓	✓		56.8
BTE4	15	41.98	✓	✓		✓	53.9
BTE5	4	65.8			✓		76
BTE5	4	67.1				✓	83.5
BTE5	12	66.7	✓		✓		77.8
BTE5	12	90.3	✓			✓	94.1
BTE5	15	48.1	✓	✓	✓		60.9
BTE5	15	49.35	✓	✓		✓	61.5
BTE7	4	26.3			✓		30.4
BTE7	4	47.6				✓	49.8
BTE7	12	48.3	✓		✓		59.5
BTE7	12	57.9	✓			✓	61.5
BTE7	15	18.71	✓	✓	✓		33.33
BTE7	15	21.86	✓	✓		✓	38.1

Approximately the results of H-BTE are the same of MFCC but there is another advantage of H-BTE over MFCC is the Q factor or the cost which calculated in Equation.5

$$Q = \frac{Sucess\ Rate\ (SR)}{Vector\ size} \qquad (5)$$

The following chart Fig 10 indicate the values of H-BTE models versus MFCC, as Q factor increases it makes the model is more trusted.

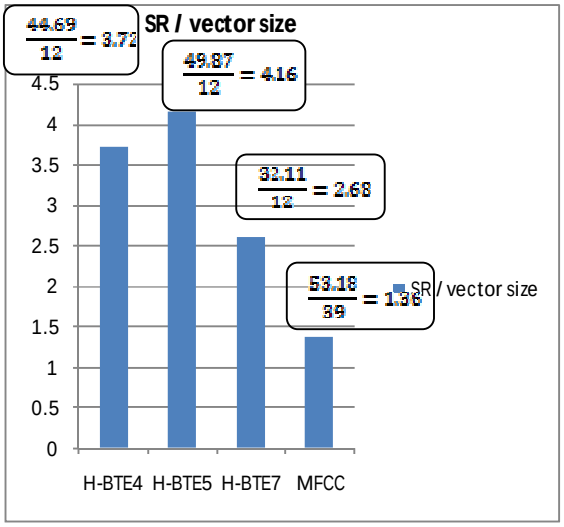

Fig 12: chart indicate the Q factor of H-BTE generations versus MFCC

V. CONCLUSION

Using Mel based entropy function significantly enhance the BTE results where BTE5 plus A&D gives success rate equals to 96%with respect to the success rate of the reference MFCC in solving the same problem but vector size 33% of MFCC vector size which is the highest for all BTE generations and by considering MFCC. H-BTE7 indicates non expected results as it gives decrease in success rate with respect to H-BTE5. Although the resolution is increased the results is degraded. This is an indication that the encoding process should be reviewed. Some other work is done for validating the encoding process. It indicates that the bits should be reordered for optimal results [16]. This work is done on BTE4. It is expected that the reordering technique should be implemented over BTE5 and BTE 7 to enhance the encoding. Hence it is expected that the results will be corrected for H-BTE7. The hybrid model will be continually developed by adding more optimizations for the encoding as well as the Mel scale.

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, 17-18 December 2008.

[2] Barnard, E, Gouws, E, Wolvaardt, K and Kleynhans, N. 2004. "*Appropriate baseline values for HMM-based speech recognition*". 15th Annual Symposium of the Pattern Recognition Association of South Africa, Grabouw, South Africa, 25 to 26 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, PP. 153-159 ,December 2011, Cairo, Egypt.

[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,PP. 126-134, December 2012, Cairo, Egypt.

[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] Michel Misiti, Yves Misiti, Georges Oppenheim, Jean-Michel Poggi, "*Wavelet Toolbox for Use with MATLAB: User's Guide*", The MathWorks, Inc., Version 1, 1996.

[7] MatLab,http://www.mathworks.com/access/helpdesk/help/toolbox/wavelet/ch06_a11.html.

[8] http://en.wikipedia.org/wiki/A_Mathematical_Theory_of_Communication

[9] R.R. Coifman, M.V. Wickerhauser, *"Entropy-based Algorithms for best basis selection,"* IEEE Trans. on Inf.Theory, vol. 38, 2, PP. 713-718, 1992.

[10] Steve Young, Mark Gales, Xunying Andrew Liu, Phil Woodland,et al. ,2006 The HTK Book, Version 3.41, Cambridge University Engineering Department, http://www.htk.eng,cam.ac.uk.

[11] Nasir Ahmad, "*A motion based approachfor audio-visual automatic speech recognition*", A Doctoral Thesis. Submitted

in partial fulfillment of the requirementsfor the award of Doctor of Philosophy of Loughborough University.
[12] HTK Book documentation, "http://htk.eng.cam.ac.uk/docs/docs.shtml".
[13] 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*",2015.
[14] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, The HTK Book (for HTK Version 3.4). Cambridge, U.K.: Cambridge Univ. Eng. Dept., 2006
[15] Amr M. Gody, Rania Ahmed AbulSeoud,Mohamed Hassan "Automatic Speech Annotation Using HMM based on Enhanced Wavelet Packets Best Tree Encoding (EWPBTE) Feature", PESCT 2013, FayoumUniveristy, 2013
[16] Barnard, E, Gouws, E, Wolvaardt, K and Kleynhans, N. 2004. "*Appropriate baseline values for HMM-based speech recognition"*. 15th Annual Symposium of the Pattern Recognition Association of South Africa, Grabouw, South Africa, 25 to 26 November 2004
[17] MatLab,http://www.mathworks.com/access/helpdesk/help/toolbox/wavelet/ch06_a11.html.
[18] http://www.researchgate.net/publication/251754208_An_HMM_based_speakerindependent_continuous_speech_recognition_system_with_experiments_on_the_TIMIT_database
[19] Carla Lopes and Fernando Perdigao (2011). "*Phoneme Recognition on the TIMIT Database"*, Speech Technologies, Prof. IvoIpsic(Ed.),ISBN:978-953-307-996-7,InTech,Availablefrom: http://www.intechopen.com/books/speech-technologies/phoneme-recognition-on-the-timit-database

First Author: **Amr M. Gody** received the B.Sc. M.Sc., and PhD. from the Faculty of Engineering, Cairo University. Egypt, in 1991, 1995 and 1999 respectively. He joined the teaching staff of the Electrical Engineering Department, Faculty of Engineering, Fayoum University, Egypt in 1994. He is author and co-author of about 40 papers in national and international conference proceedings and journals. He is the Acting chief of Electrical Engineering department, Fayoum University in 2010, 2012 , 2013 and 2014. His current research areas of interest include speech processing, speech recognition and speech compression.

Second Author: Rania Ahmed AbulSeoud received the B.S. degrees in Electrical Engineering-Communications and Electronics Department at Cairo University – EL Fayoum Branch in 1998 and M.S.E. degrees in Computer Engineering at Cairo University in 2005. Her Ph.D. degree was from the Biomedical Engineering department, Cairo University in 2008. She worked as a Demonstrator and a Teaching Assistant in Electrical Engineering Department of Misr University for Science and Technology, Egypt since 1998. She is currently an Associate Professor of the Electronics and Communications Engineering Department, Fayoum University, Egypt.. Her areas of interest in research are Artificial Intelligence, Natural Language Processing, and computational linguistics, and machine translation, application of artificial Intelligence to computational biology and bioinformatics and Computernet works.

Third Author: **Marian M.Ibraheem** received the B.Sc. degree in Electrical Engineering – Communications and Electronics Department with very good degree, from the Faculty of Engineering - Fayoum University in 2008.She joined the M.Sc program in Fayoum University - Communications and Electronics Department in 2012.She received the Pre-Master degreeinFayoum University with very good degree, in 2012. Her areas of interest include Best-Tree Encoding model, speech recognition.