Using Mel-Mapped Best Tree Encoding for Baseline-Context-Independent-Mono-Phone Automatic Speech Recognition

Amr M. Gody[*1], Rania Ahmed Abul Seoud[*2], Mai Ezz El-Din[*3]

**Electrical Engineering, Faculty of Engineering, Fayom University, Egypt*

[1]amg00@fayoum.edu.eg

[2]raa00@fayoum.edu.eg

[3]Mai.ezzeldin.89@gmail.com

Abstract*: Best-Tree Encoding (BTE) is first introduced by Amr M. Gody [1] as new features for Automatic Speech Recognition (ASR) problem. BTE is basically acting as spectrum analyzer. It relies on Wavelet packets to get projection of signal power into predefined filter banks. The feature components are encoded into digital form using certain entropy method and certain digital encoding procedure. In this research BTE is further developed by including two more key factors into the BTE process. The key factors are Mel-scale (MS) and baseband Bandwidth mapping (BM).This Research provides a baseline performance evaluation for Context-independent mono-phone recognition (Without Grammar) of English by using Vid-TIMIT database. Vid-TIMIT consists of 43 speakers (19 female and 24 male), reciting short sentences. The recording of this database was done in a noisy environment (mostly computer fan noise) and also it is not hand verified. Total of 15643 phone segments are used for testing and evaluating the newly proposed features. HMM is used as recognition engine via HTK toolkit for its popularity in ASR. Comparison to MFCC on the same database is considered to evaluate the system results. Although it gives the same recognition efficiency as MFCC on the same testing database, the proposed model saves almost 66% of the required storage than the feature vector of MFCC.*

Keywords: *Automatic Speech recognition (ASR), Arabic Phone Recognition, Wavelet packets, Mel-Scale, WPBTE, MFCC, HTK and BTE.*

Symbols:

BTE	***Best-Tree Encoding***
ASR	***Automatic Speech Recognition***
MS	***Mel-scale***
BM	***Bandwidth mapping***
BM-BTE	***Band Mapping in BTE***
MM-BTE	***Mel Mapping in BTE***
MBM-BTE	***Mel and Band Mapping in BTE***
MFCC	***Mel Freqancy Cepestral Coefficient***

1 INTRODUCTION

Human have several inherent characteristics that facilitate them differentiate from one another. Over the years, biometrics has emerged as the science which realizes and attempts to imitate and simulate the human brain via capture unrivaled personal features and consequently performing the mission of human identification.

The applications of speech processing technology may be generally classified into Speaker Recognition and Speech Recognition. The speech recognition can be defined as the capability to differentiate the spoken words. However, the speaker recognition is the capability of the differentiation between the speaker people. The feature extraction process transforms the signal properties which are significant for the pattern recognition mission to a simple format. This format simplifies the distinction of the classes. The most used approach in ASR is HMM. The most popular fundamental units are either mono phone models or context-dependent units may be called Tri-Phone or syllables. Syllable or Tri-Phone HMM needs balanced database for Tri-Phone with reasonable reparations for each model. The objective in this research is to test newly developed features for ASR. As of this objective and considering the concern of unbalanced database for Tri-Phone makes us to choose Mono-Phone based model instead to avoid database effect in the final results. Although the mono-phone is rarely used in practical application, it is used in this research for the availability of its database. Comparison results are used as measurer

for the efficiency of the proposed models. Comparison to Mel Frequency Cepstral Coefficients (MFCC) -based HMM model is considered to provide the results of the proposed models.

Mel Frequency Cepstral Coefficients (MFCCs) with various speech parameterizations are commonly used as features in speech recognition systems, especially in the HMM-based approach. There are other speech features also that have been used such as the systems which can automatically recognize numbers spoken into a telephone, Line Spectral Frequencies(LSF), Linear Predictive Coding (LPC), short-time energy, and wavelet-based.

Wavelets are recently used in many real time applications, such as automatic speech recognition applications. In [2], it focus on speech recognition and has undertaken a study to determine an appropriate set of parameter setting for first- generation phone –based systems that use Hidden Markov Models (HMMs). Their experiments are based on the open-source toolkit HTK and were restricted to English-language TIMIT corpus. Fourteen male and fourteen female speakers were chosen from each district, which created the training set. In this paper, all tests used the TIMIT speech corpus. TIMIT consists of 6300 utterances from 630 different speakers of American English, recorded with a high-fidelity microphone in a noise-free environment. The MFCC analysis performed the best and most consistently, with no real gain in performance when the number of parameters was increased. MFCC parameters are 12 and by inclusion of the delta coefficients, it gave a considerable increase in performance, but increased the number of parameters two fold. The acceleration coefficients also increased the performance measure but not as dramatically compared to the delta coefficients. With the inclusion of acceleration coefficients the amount of parameters increases three fold. The recognition accuracy of MFCC for mono-phone data varied between 41% and 44%.

In this paper, the experiments are implemented using the open-source toolkit HTK and were restricted to English-language Vid-TIMIT corpus. 43 speakers (24 male and 19 female speakers) were chosen from each district, which created the training set. In this paper, all tests used the Vid-TIMIT speech corpus which recorded of in a noisy environment (mostly computer fan noise) and also it is not hand verified. Vid-TIMIT consists of 15643 phone segments from 43 different speakers of American English, recorded in a noisy environment.

MFCC is chosen as reference for the results. MFCC is used to obtain the reference results on the available Vid-TIMIT. Then experiments are run to get the results for the proposed features and models. The results are provided as comparison to the reference MFCC results on Vid-TIMIT. All proposed models parameters are 4 and by inclusion of the delta coefficients, it gave a considerable increase in performance, but increased the number of parameters two fold. The acceleration coefficients also increased the performance measure but not as dramatically compared to the delta coefficients. With the inclusion of acceleration coefficients the amount of parameters increases three fold. The recognition accuracy of proposed model (MBM-BTE) for mono-phone data achieves 38.9 %. When applied the same conditions on the MFCC and using Vid-TIMIT database, the recognition accuracy of MFCC achieves 38.25%. (Note that Vid-TIMIT Database is superimposed with additive noise. In [2], TIMIT is used without any additive noise; it is proven that MFCC get 41% success rate for mono-phone database. This is good indication that MFCC on Vid-TIMIT with 38% is very reasonable and can be used as reference results for non-Grammar mono-phone recognition problem). In summary; the recognition accuracy of 38% using MFCC on Vid-TIMIT is very reasonable compared to 41 to 44% on cleaner and larger database (TIMIT) as mentioned above. This comparison is mentioned here to give confidence on using the MFCC on Vid-TIMIT as reference through this research.

A new design feature called wavelet Packets Best Tree Encoding (WPBTE) is designed to enhance the efficiency of Automatic Speech Recognition (ASR) by providing human like processing of speech stream. Many loops of enhancement for BTE are provided to enhance the efficiency. In [3] introduced a completely automated phone recognition system based on Best Tree Encoding (BTE) 4-point speech feature. The System identified spoken phone 57.2% recognition rate based on BTE. Another paper [4] deals with newly designed features for speech signal that can be used in Automatic Speech Recognition (ASR). In [4] the Information related to speech phoneme is encoded into 15 bits instead of 7 bits in the original version of Best Tree Encoding (BTE4) in [1]. Best Tree Encoding of 5 levels of wavelet analysis (BTE5). This feature has given 25% efficiency enhancement over the original BTE4[1] for solving the ASR problem. In addition; BTE5 feature vectors with size 15, gives 25% success rate (SR) while MFCC with size 13 gives 39% SR for the same problem. This is very promising that BTE5 is approaching 71% of the SR of the most popular feature used in the applied area of automatic speech recognition. A third paper [5] aims to enhance BTE encoder by adding two factors to BTE encoder. Analysis levels (AL) in wavelet packets becomes 7 and Energy becomes 4 components instead of single component in BTE. BTE7 gives 22% Efficiency enhancement over BTE4 and about 45% efficiency enhancement over BTE5. In addition, BTE7 indicates more

stability for recognition results over both BTE4 and BTE5. BTE7 gives more than 10% accuracy enhancements over both BTE4 and BTE5.

The following sections navigate through the details of this research. Section two illustrates the block diagram and the key parameters to improve the performance of BTE. Section three is deeply introduces the proposed models. Section four provides review analysis of HMM and HTK. Section five demonstrates the experiments parameters. Finally, the results, discussion, and conclusion on the obtained results are illustrated at sections six and seven.

2 MEL MAPPED BEST TREE ENCODING (BTE) OVERVIEW

In this section, BESTTREE ENCODING features, mel frequency mapping will be illustrated.

A. Besttree Encoding (BTE) Model

BTE is first introduced by Amr Gody in [1]. It is intended to develop Human-Perception-Like based features. The procedure of extracting BTE will be illustrated through the block diagram in Fig. 1.

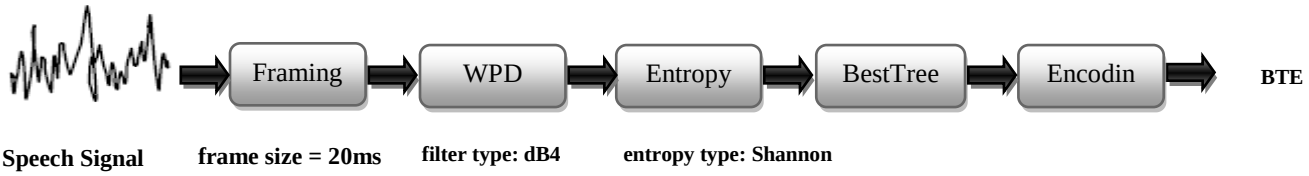


Figure 1: Block diagram of creating BTE

The first stage of creating BTE features vector is the ***framing block.*** At this step, the time waveform (samples) is segmented into small frames (20 ms). The second step of creating BTE is the ***Wavelet packet decomposition (WPD*) *block***. At this block, Daubechies wavelet filter family with four points [6] is used. The output of this step is the Wavelet packets Decomposition (WPD) of the input time frame. This is the step of spectrum extraction from the time waveform as shown in hierarchical filter bank equivalent for this decomposition wavelet filters Fig. 2.

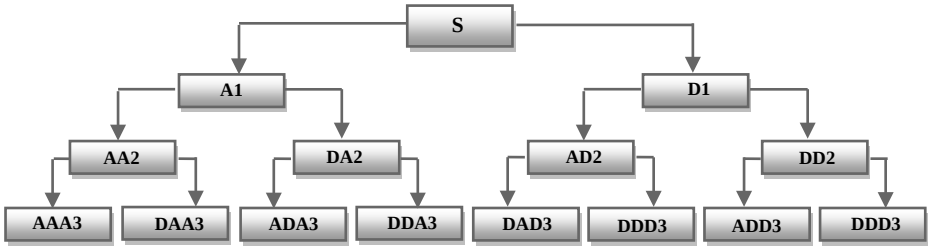


Figure 2: Signal decomposition using wavelet packets [7].

The next step is Entropy. This step is the next step after ***WPD.*** Entropy is the key step to enhance BTE. Entropy is used to measure information in each tree node in figure 2. Accordingly the best tree is decided by removing all low informative tree nodes. In the original BTE **Shannon entropy** is chosen (Schneier, Shannon & Claude E., January, 1951 [8]), as given in equation 1.

$$H(X) = -\sum_{i} p(x_i) \log p(x_i) \qquad (1)$$

where $p(x_i)$ is the probability of a given symbol.

The next Block is ***Best-Tree block.*** The Besttree function [9] uses the entropy to remove the low informative tree nodes. The final step is the ***Encoding block***. This step is the process to generate BTE features vector. Binary encoding algorithm is used to encode the structure of the best tree obtained in the previous step. Tables1 and 2 illustrate the four points encoding algorithm for BTE-4. For detailed discussion the reader may refer to paper [1]. In summary the nodes are rearranged in order to minimize the distance between the adjacent in frequency features vectors. As quick example, wavelet packets indexing system in table 1indicates that node two and node three are subsequent but they are not in frequency (V1 at low band while V2 at High band).Table 2 illustrates the proposed coding. Note that node 2 and 3 falls into the same band as well as they are consecutive numbers.

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				

B. The Pruning Process Of The Binary Tree

WPD is expressed as 4-Levels binary tree as shown in figure 3. Each node in the tree contains "left and right" children except the leaves nodes. Leaves are those nodes with no Childs. Entropy is evaluated for all Tree nodes. Each Tree node is identified by unique number identifier as shown in figure 3.

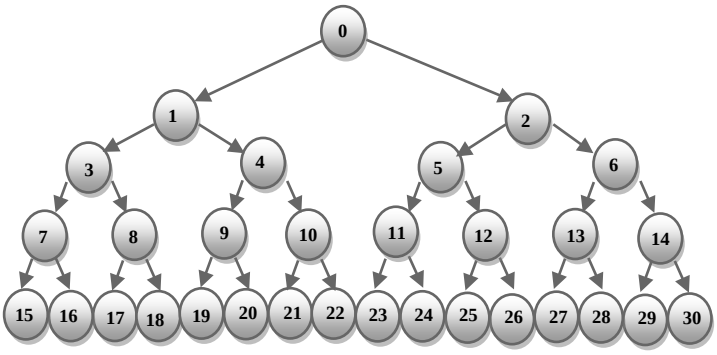


Figure 3: The tree before the cutting

The best tree is then obtained from the original binary tree by pruning of unnecessary nodes of the tree depending on Shannon entropy. Matlab is used for this process. The code of the pruning process is provided in Fig.4. The idea of the pruning process is based on comparing the entropy of the children with its parent. If the summation of the entropy of each

two children is bigger than its parent entropy, those children will be pruned. Fig. 5 defines the variables that are used in the pruning code. Fig.6 shows sample of best tree after pruning process. The work in this research modifies this pruning algorithm to remove the non-informative nodes according to Mel-Scale (Human-Perception Mechanism). This will be introduced in the next section.

```
ento   = NaN*ones(size(Ent));
rec    = 2*ones(size(An));
rec(Tn_ind)  = ones(size(Tn));
ento(Tn_ind) = Ent(Tn_ind);
J = wrev(find(rec==2));
k=1:length(J)
   ind_n   = J(k);
   node    = An(ind_n);
   child   = node*Order+[1:Order]';
   i_child = gidxsint(An,child);
   echild  = sum(ento(i_child));
if echild < Ent(ind_n)
      ento(ind_n) = echild;
      rec(ind_n) = 2;
else
      ento(ind_n)  = Ent(ind_n);
      rec(ind_n)   = rec(i_child(1))+2;
      rec(i_child) = -rec(i_child);
end
```

Figure 4: The pruning condition code

Ent: is a vector that contains the entropy of each node in the tree.
An: All Nodes
Tn: Terminal nodes
Tn_ind: The index of the terminal nodes
Order: is equal to 2 because it is a binary tree.
J: is a vector that contains from 15 to 1.

Figure 5: Variables that are used in the pruning code

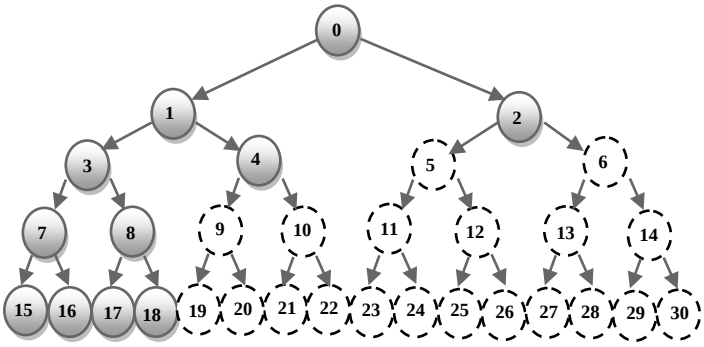


Figure 6: The pruned tree (best tree)

C. Mel Frequency Cepstral Coefficients (MFCC)

MFCC's are depending on the variation of the frequency with the human ear's critical bandwidths. Human ears cannot differentiate between different sounds in high frequency scale in the same efficiency like it can do in low frequency scale. Mel-Scale (MS) is a scale that reflects the hearing mechanism in the human ear. At MS frequency, linear response is located below 1000 Hz. At high frequencies the relation is becoming more logarithmic.

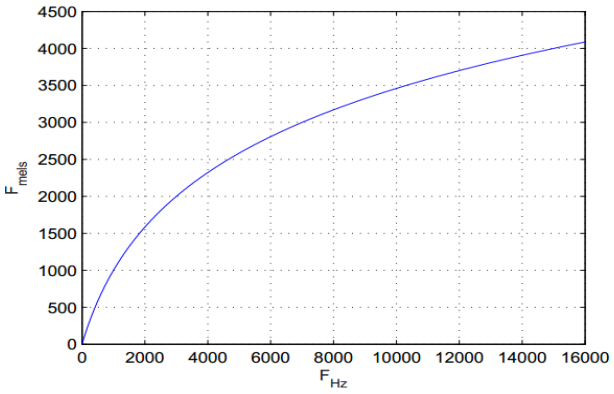


Figure 7: Relationship between the frequency scale and Mel-Scale (MS).

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

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

where, f_{Hz} is the frequency in the hertz unit.

The critical bandwidth which varies with the frequency is band-pass filter, adjusted around the center frequency. This center frequency is separate on the linear range (which be around 100, 200,… 1000 Hz) and in the logarithmic area (Above 1000Hz).

The block diagram of MFCC feature extraction algorithm is as shown in Fig. 8

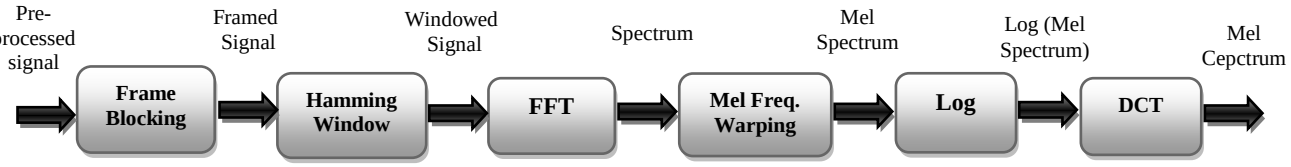


Figure 8: Block Diagram implementation of the technique

After that the signal is multiplied over short-time window. This window is used to avoid the arising problems which occur due to the truncation of the signal. The Fast Fourier Transform (FFT) is used to estimate the spectrum for each frame. Then to obtain the Mel spectrum, passing this spectrum through Mel filter bank. Finally, the logarithm followed by the Discrete Cosine Transform (DCT) that produces a set of feature vectors (one vector corresponding to each frame) which are then termed as MFCC. The Mel-Scale will be included as key factor in Entropy function in order that to enhance the pruning of leave nodes in such to keep the nodes that matches with Mel-Scale and to remove the nodes that don't match with Mel-scale.

D. Resampling Preparation Database

Re-sampling aims to reducing the sampling rate by a factor M as illustrated at Fig. 9. The output of the down-sampled signal $y(m)$ does not contain all the information about the original signal $x(m)$. Consequently, Band mapping (BM) is frequently applied in the filter banks and usually preceded by filters to extract the relevant frequency bands.

Figure 9: Block Diagram of the Down-sampling

The aim of applying the Band Mapping on the input signal is to remove the redundancy in the baseline signal. The Baseline signal is sampled at 32 KHz. This implies that the Band width is 16 KHz. Leave Tree nodes in figure 3 should cover the bandwidth of the baseline signal. This implies that node 15 at the low band (starts from 0 Hz) while node 30 will map the high band ends with 16 KHz. The count of leave nodes are 16, hence each leave node in figure 3 will contribute band width of $\frac{16}{16} = 1\ KHz$ .As the matter of fact that Human speech information lies in the range of 0 to 4 KHz, hence node 18 will be the maximum possible limit of human information. The remaining nodes from 19 to 30 are expressing frequency components higher than 4 KHz.

Band Mapping is the process of down sampling the baseline signal to 10KHZ. This will make the bandwidth is 5 KHZ. Consequently, the leave nodes in figure 3 will cover the bandwidth of 5 KHZ. Then the resolution to express the contained information will be enhanced by including all Tree nodes instead of the limitation in the previous version of BTE by ignoring the 4 KHz limit of human hearing mechanism. Using this mapping all tree nodes will be included into the process of encoding but not only part of Tree nodes as of the previous version. Note that in figure 6 all leaf over 18 are truncated in the previous version of BTE because the nodes higher than 18 all are not informative. Those nodes bear information about frequency components higher than 4 KHz which are not significant in human hearing mechanism.

3 PROPOSED MODELS

In this paper, three enhanced BTE models are proposed. The first enhanced model applies the Band-Mapped (BM) BTE. The second enhancement is achieved by combining the Mel-Scale BTE. Finally, both Mel-Scale and BM are both integrated. In the following sub-sections, the three models will be demonstrated in details.

A. Band Mapped BTE (BM-BTE)feature

At this feature 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 as indicated before in section 2-D. BM-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).

As illustrated at Fig.10.a and 10.b, the tree before applying the BM is distributed at the bandwidth from 0 to 16 KHz. According to the MS, 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.

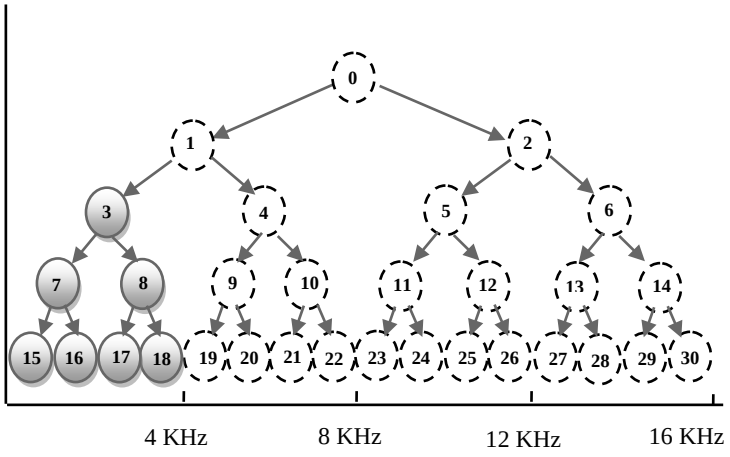


Figure 10.a: The tree before applying the Band Mapping (BM)

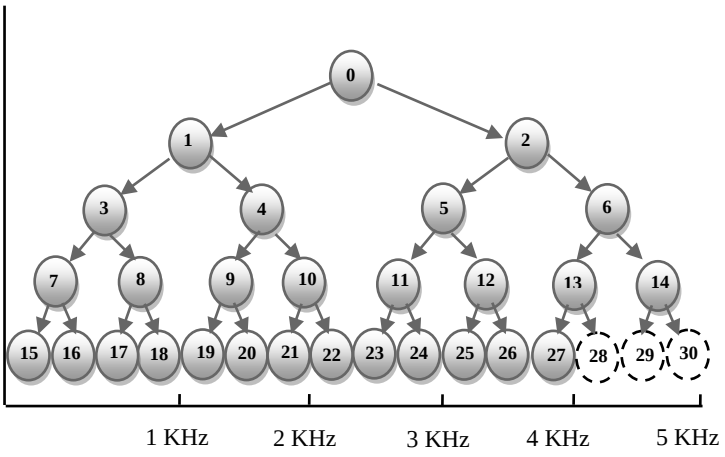


Figure 10.b: The tree after applying the Band Mapping (BM)

```
[y fs] = wavread(file);
downSampleRate = int32(fs/10000);
 y = downsample(y,downSampleRate);
 fs = fs / downSampleRate;
```

Figure 11: The Down-Sampling code

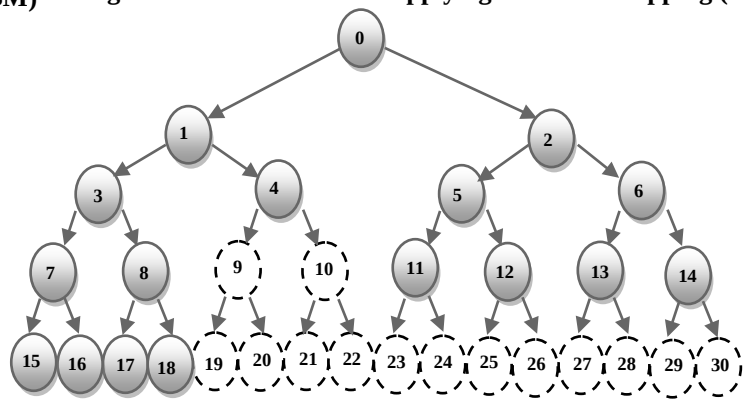


Figure 12: Sample BM Tree. It is clearly appear that all nodes are contributing in tree shape structure. Compare it to figure 6, only the first left quarter in the tree is used in building the tree.

The following code is used to limit the band to become about 10000 HZ as shown in Fig.11. A sample of a tree that is produced from the BTE model with using BM after the pruning process is illustrated in Fig.12.

B. Mel Mapping BTE (MM-BTE) feature

In this approach, weight is calculated for each node based on the position of this node on the MS curve. 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.

The weight (W) of each node can be calculated as of equation 5 where the mean frequency of the node is $f_{node}$and $delta$ is the absolute difference between node's frequency and the MS frequency of this node. The $delta$ can be determined as of equation 4. f_{mel} is evaluated for each node according to equation 3.

$$f_{Mel} = 2595 * log_{10}(1 + \frac{f_{node}}{700}) \quad (3)$$

$$Delta = |f_{mel} - f_{node}| \quad (4)$$

$$Weight\ (W) = \frac{(f_{node} - delta) * 100}{f_{node}} \quad (5)$$

Frequency range of each node (Node Bandwidth) can be determined based on the start, mean, and end frequencies of its parent and also based on which side will be taken by the node (left or right node) as illustrated in the equations number 6 to 11where f_P is the parent frequency, f_{psr} is the parent start range frequency, f_{node-s} is the child start frequency, f_{node-e} is the child end frequency, and f_{per} is the parent end range frequency.

Left side child calculations

$$f_{node} = \frac{f_P - f_{psr}}{2} + f_{psr} \quad (6)$$

$$f_{node-s} = f_{psr} \quad (7)$$

$$f_{node-e} = f_p \quad (8)$$

Right side child calculations

$$f_{node} = \frac{f_{per} - f_p}{2} + f_p \quad (9)$$

$$f_{node-s} = f_p \quad (10)$$

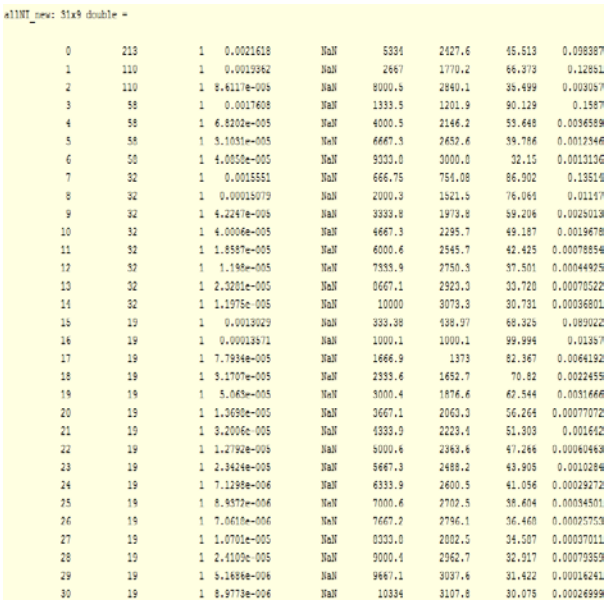
$$f_{node-e} = f_{per} \quad (11)$$

For example, node 1 and 2 are the children of node 0 as shown in Fig.14. Node 0 covers the band from 0 to 10000Hz with center frequency 5000HZ whereas Node 1 has f_c 2500Hz and covers the band from 0 to 5000Hz and Node 2 covers the band from 5000 to 10000Hz with f_c equals 7500Hz and so on for the remainder of the tree nodes. Fig.13 illustrates the Matlab script for MS-BTE model.

Also as shown in the Matlab script (Last line in Fig.13), each entropy is multiplied by its corresponding weight to get weighted entropy that is used during the pruning process. Fig.14 illustrates the output matrix from this script. Fig.15 illustrates a sample for an output tree after the pruning process,

```
child = node*order+(1:order)';
   i_c   = child+1;
for k =1:order
      parentFreq = double(freq_nodes{ind,1});
      parentStartRange = double(freq_nodes{ind,2});
      parentEndRange = double(freq_nodes{ind,3});
      childFreq = 0;
      childStart = 0;
      childEnd = 0;
if k == 1,
         childFreq = (parentFreq - parentStartRange) / 2 + parentStartRange;
         childStart = parentStartRange;
         childEnd = parentFreq;
elseif k==2,
         childFreq = ((parentEndRange - parentFreq) / 2) + parentFreq;
         childStart = parentFreq;
         childEnd = parentEndRange;
end
      freq_nodes{i_c(k),1} = childFreq;
      freq_nodes{i_c(k),2} = childStart;
      freq_nodes{i_c(k),3} = childEnd;
      mell = 2595*log10(1 + (childFreq/700));
      delta = abs(mell - childFreq);
      weight = (childFreq - delta)*100/childFreq;
      allNI_new(i_c(k),:) = [allNI(i_c(k),:) childFreq mell weight (weight *allNI(i_c(k),4))];
```

Figure 13:BTE model with Mel-scale (MS) techniques code



0	213	1	0.0021618	NaN	5334	2427.6	45.513	0.098387
1	110	1	0.0019962	NaN	2667	1770.2	66.373	0.12851
2	110	1	8.6117e-005	NaN	8000.5	2840.1	35.499	0.003057
3	58	1	0.0017608	NaN	1333.5	1201.9	90.129	0.1587
4	58	1	6.8202e-005	NaN	4000.5	2146.2	53.648	0.0036589
5	58	1	5.1051e-005	NaN	6667.3	2652.6	39.786	0.0012346
6	58	1	4.0858e-005	NaN	9333.0	3000.0	32.15	0.0013136
7	32	1	0.0015551	NaN	666.75	754.08	86.902	0.13514
8	32	1	0.00015079	NaN	2000.3	1521.5	76.064	0.01157
9	32	1	4.2247e-005	NaN	3333.8	1973.8	59.206	0.0025013
10	32	1	4.0006e-005	NaN	4667.3	2295.7	49.187	0.0019678
11	32	1	1.8587e-005	NaN	6000.6	2545.7	42.425	0.00078854
12	32	1	1.198e-005	NaN	7333.9	2750.5	37.501	0.00044925
13	32	1	2.3201e-005	NaN	8667.1	2923.3	33.728	0.00070522
14	32	1	1.1975e-005	NaN	10000	3073.3	30.731	0.00036801
15	19	1	0.0019029	NaN	333.38	438.97	68.325	0.089022
16	19	1	0.00013571	NaN	1000.1	1000.1	99.996	0.01357
17	19	1	7.7934e-005	NaN	1666.9	1373	82.367	0.0064192
18	19	1	9.1707e-005	NaN	2333.6	1652.7	70.82	0.0022455
19	19	1	5.063e-005	NaN	3000.4	1876.6	62.544	0.0031666
20	19	1	1.3698e-005	NaN	3667.1	2063.3	56.264	0.00077072
21	19	1	3.2006e-005	NaN	4333.9	2229.4	51.303	0.001642
22	19	1	1.2792e-005	NaN	5000.6	2383.6	47.266	0.00060463
23	19	1	2.3424e-005	NaN	5667.3	2488.2	43.905	0.0010284
24	19	1	7.1198e-006	NaN	6333.9	2600.5	41.056	0.00029272
25	19	1	8.8572e-006	NaN	7000.6	2702.5	38.604	0.00034501
26	19	1	7.0618e-006	NaN	7667.2	2796.1	36.468	0.00025753
27	19	1	1.0701e-005	NaN	8333.8	2882.5	34.587	0.00037011
28	19	1	2.4100e-005	NaN	9000.4	2962.7	32.917	0.00079359
29	19	1	5.1686e-006	NaN	9667.1	3037.6	31.422	0.00016241
30	19	1	8.9773e-006	NaN	10334	3107.8	30.075	0.00026999

Figure 14: Output Matrix from the code

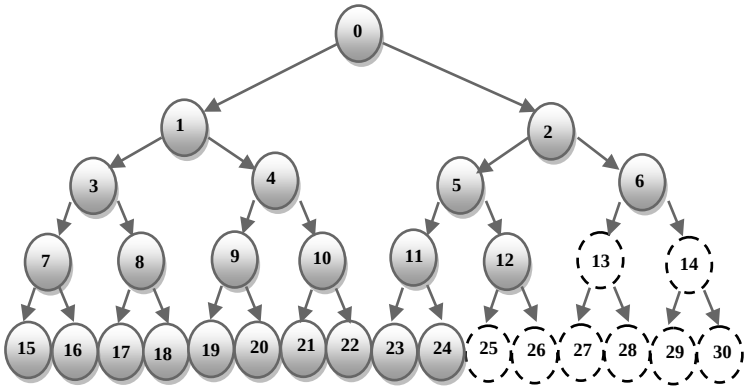


Figure 15: The tree after the pruning using MS-BTE (best tree)

C. Mel and Band Mapped BTE (MBM-BTE) feature

This is combination of BM-BTE and MM-BTE. The symbol MBM-BTE is designated to identify this model along with this paper script. Fig. 16 illustrates the tree after the pruning process of the tree using both Band mapped and Mel Scale techniques.

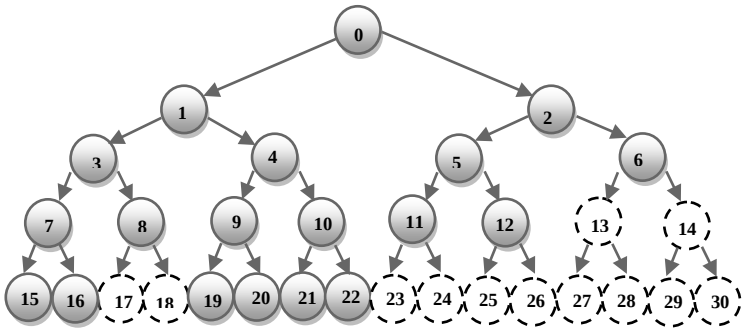


Figure 16: The tree after the pruning using MS and BM (best tree)

4 HMM MODEL

The Hidden Markov Model (HMM) is a robust statistical tool for solving the automatic speech recognition (ASR) problems and modeling generative sequences. HMM applications are exploited for many interested areas of the signal processing. In this research, Hidden Markov Model Toolkit (HTK) [10] is utilized for configuring, training and testing the proposed models. HTK will be trained and tested using BTE feature vectors for different models {BM-BTE, MS-BTE and MBM-BTE}. HMM models are adapted from the single Gaussian HMMs to multi mixture components. Hence, the process of increasing number for the components in a mixture is called *Mixture splitting.* The process occurs by increasing the number of mixtures by one or two then re-estimating. Therefore, the number of mixture is increased and re-estimated in each step until the optimal number of components is obtained.

5 EXPERIMENTS

In the proposed models framework, HTK tool is utilized to build HMM based speech processing tools, especially speech recognizers. The set of HMM models parameters could be estimated by using the training tools. This process occurred by using training utterances and their associated transcriptions. In this framework MATLAB and Microsoft C# (C sharp) language are used for building the needed logic for the initial models of HTK.

In ASR, performance can be greatly enhanced by adding Delta and Acceleration coefficients of the vector components to features vector. Where delta coefficients (D) indicate the first order regression coefficients and the acceleration coefficients (A) indicate the second order regression coefficients. HTK gives the option to take these coefficients (D and A) into consideration during the training and testing process. HTK also allows changing the number of Gaussian Mixtures in the range from one to three in HMM emitting states. This Research provides baseline performance evaluation for vocabulary-independent mono-phone recognition of English by using Vid-TIMIT database. The HMM-based recognizer was trained with not-hand-verified data from 43 speakers (19 female and 24 male), reciting short sentences. Using 35 context-independent phone models, baseline phone accuracy was obtained on an independent test set of 15634 phone segments from 20 speakers. The recording of this database was done in a noisy environment (mostly computer fan noise) and also not hand verified. In this paper 20% from database was used for testing and 80% was used for training. For reference of using Vid-TIMIT in speech recognition, reader is recommended to read the thesis by Nasir Ahmad [11]. In his work [11], MFCC-based audio features from noisy audio signal were then extracted at signal-to-noise ratios (SNR) ranging from 30dB to -10 dB and combined with each of the new motion-based visual features using Vid-TIMIT database. This database contains 420recorded wave files which are mono-phones in wav format, 335 files used for training and for testing, 85 files are used. The recording was done in a noisy environment (mostly computer fan noise). The sampling rate of the recorded wave files is 32 kHz, 16 bit.

6 RESULTS AND DISCUTION

HMM model with three emitting states and various Gaussian Mixtures in each state are used to model the English Mono-Phones. MATLAB, Visual studio, and Hidden Markov model Tool Kit (HTK) are used to implement the framework and evaluating the results. The results of the three proposed models {BM-BTE, MM-BTE and MBM-BTE} are illustrated in this section. As mentioned earlier, MFCC will be considered as reference for all models. MFCC will be used as features vectors for Vid-TIMIT, the same database used in this research, to obtain the reference results.

Fig.17 illustrates the success rate % for MFCC (The reference curve) against the Pruning threshold (HTK parameter that is being adapted for optimal results). For more details about this parameter the reader is recommended to read HTK book [12]. The recognition rates are optimized by changing the values this pruning factor. For MFCC (The reference curve) the recognition success rate % starts from 33.13% towards 38.13%.

Fig. 18 shows the effect of changing the pruning threshold on BM-BTE model with Delta only and another time with Delta and Acceleration. The maximum value of recognition rate in BM model achieves 38.33%.

Fig. 19 illustrates the effect of changing the pruning threshold on MM-BTE model with Delta only and with Delta and Acceleration in the other curve. The maximum value of recognition rates in MM-BTE model achieves 35.62%. And the effect of changing the pruning threshold on MBM-BTE model is shown in Fig. 20, the maximum value of recognition rates in MBM-BTE model achieves 38.9%. Finally, comparison chart is provided to illustrate the success rate for each model against the reference MFCC in Fig. 21.

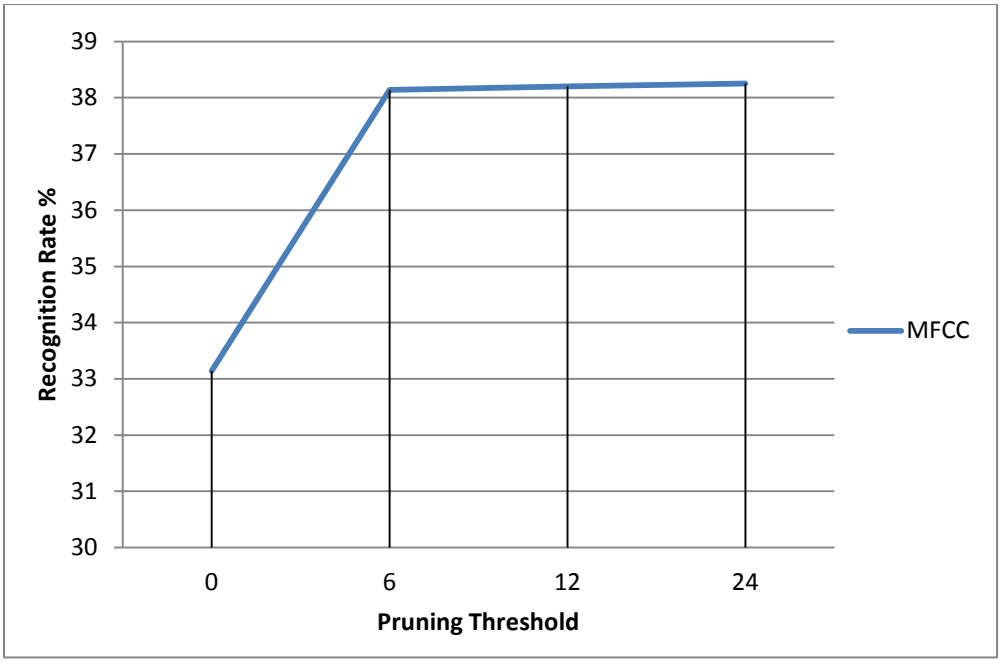


Figure 17: Effect of increasing pruning threshold on recognition rate in MFCC model

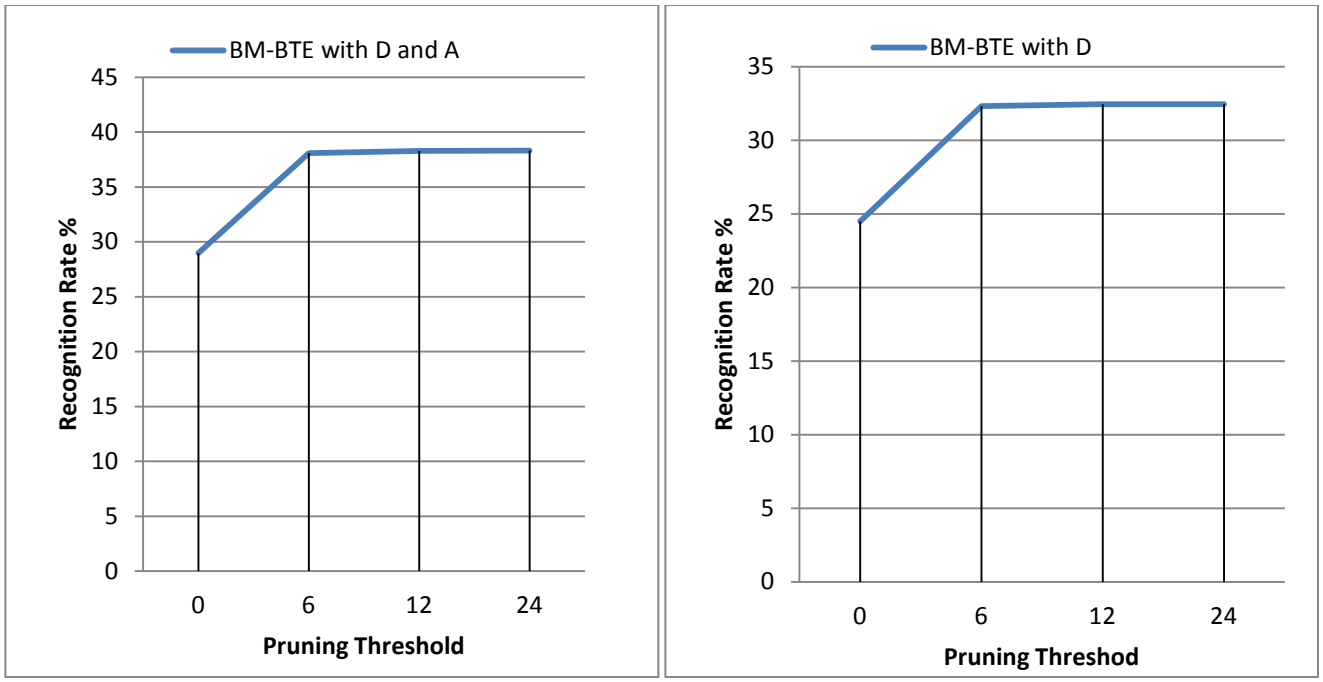


Figure 18: Effect of increasing pruning threshold on recognition rate in BM-BTE model

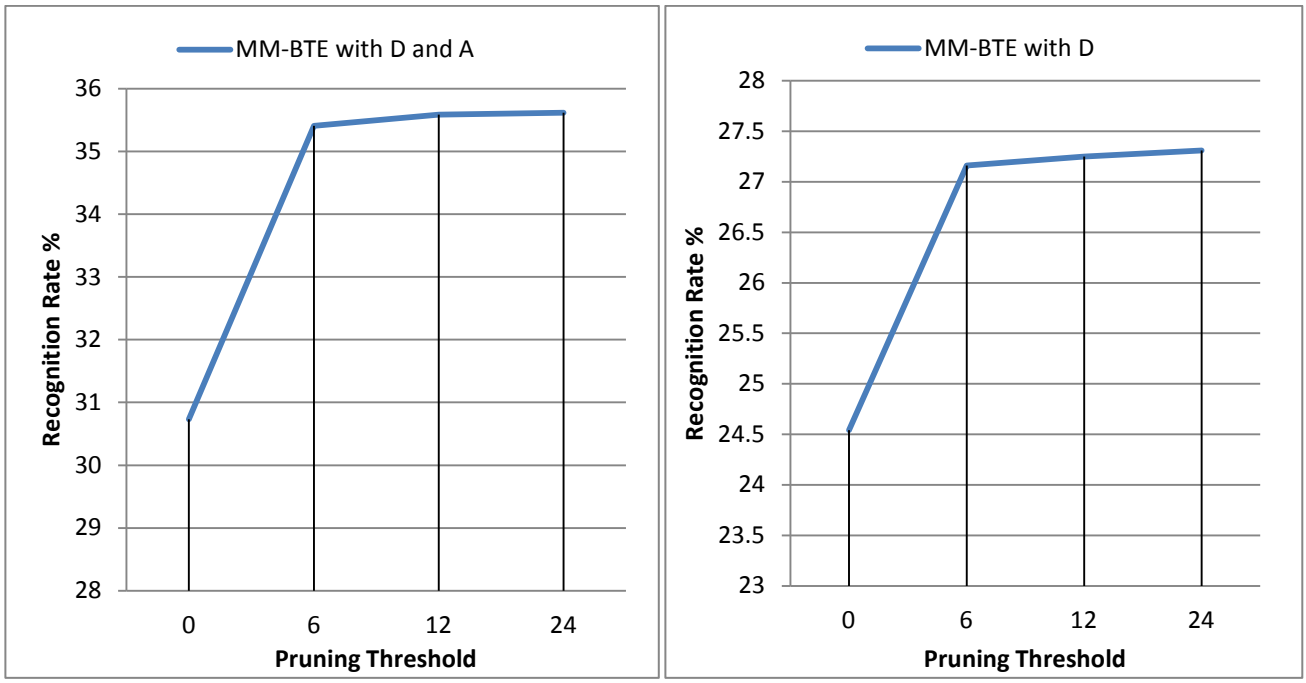


Figure 19: Effect of increasing pruning threshold on recognition rate in MM-BTE model

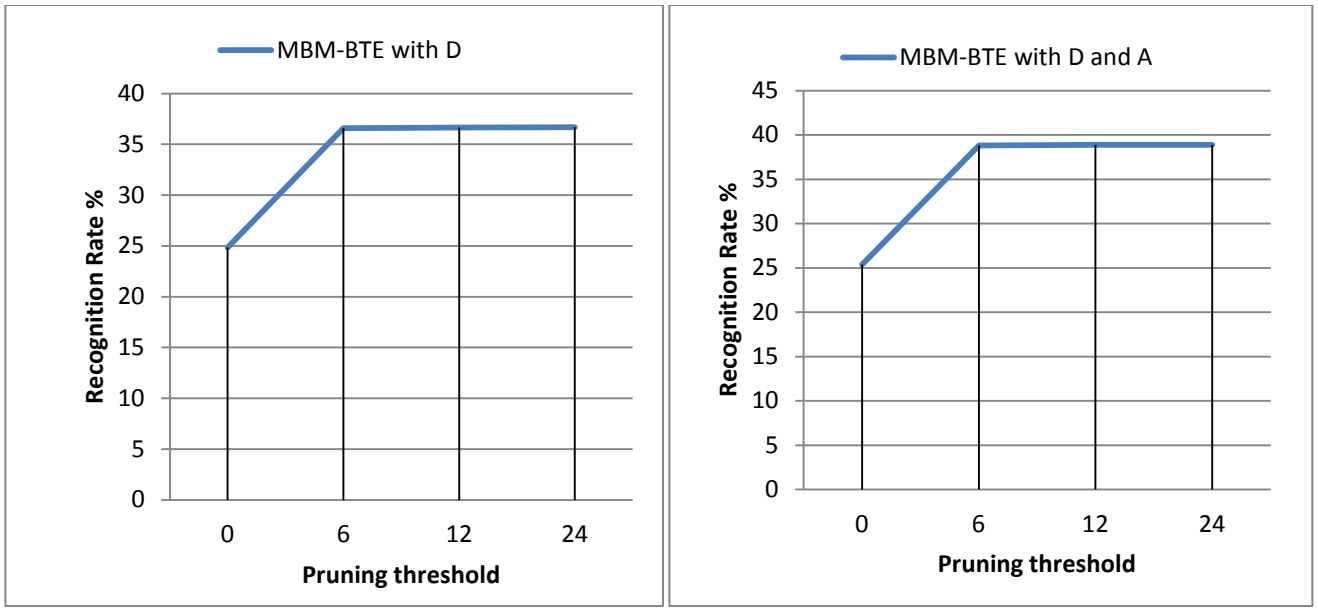


Figure 20: Effect of increasing pruning threshold on recognition rate in MBM-BTE model

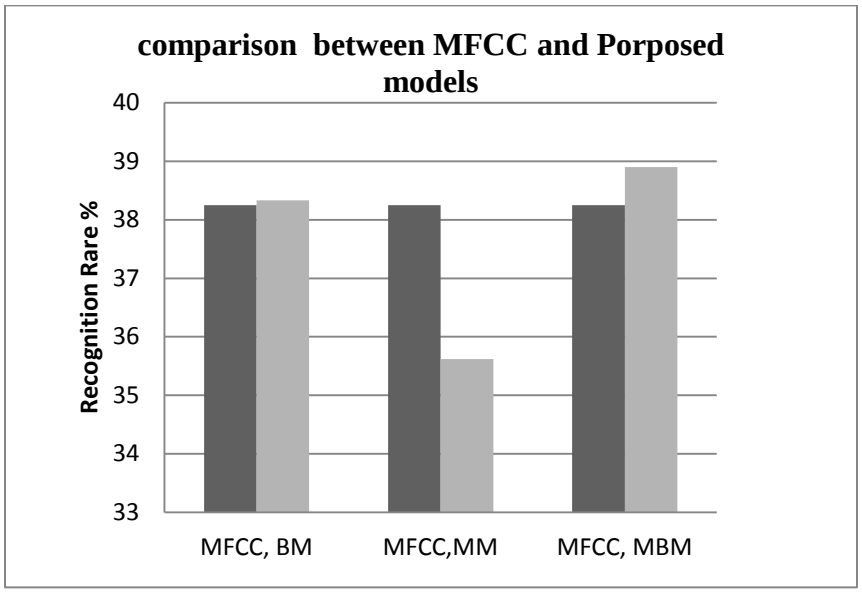


Figure 21: Comparison between the maximum value of MFCC and proposed models

CONCLUSION

The methodology of mixing BTE 4 model with the Mel Scale and Band mapping the baseline signal (MBM-BTE) is very promising to enhance the behavior of Automatic Speech Recognition problem. In the proposed model the results of ASR for English mono-phone is almost exceeding the obtained results using MFCC on the same database for the same problem. The problem of English mono-phone is chosen without grammar because the target is to prove the efficiency of the proposed model against MFCC by providing comparison results on the same database for the same problem. Adding the grammar will not change the relative efficiency; it will just elevate the obtained success rate but will keep the relative difference. As of that the complexity will not add value for the final results, it is chosen to work for mono-phone problem. Mono-phone problem has lower success rate in general than context-Dependent phone recognition

with Grammar. The values range from 41% for mono-phone compared to 70% for Context-Dependent and may exceed 96% after adding the Grammar and language model [2]. In this research, the proposed model gives 38.9% compared with 38.13% by the reference MFCC for Mono-Phone Recognition of English language. This indicates 1.02 enhancements over MFCC for the same problem. Take into account that MBM-BTE vector size is only 4 components compared to MFCC (12 components) {without adding delta and acceleration parameters). So that this research provides new promising features that gives 1.02 enhancements over the best practical features in use in the market of ASR (MFCC) with 0.33 size of MFCC (almost 66% size improvement over MFCC).

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 – Energy Besttree 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 Math Works, 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 approach for audio-visual automatic speech recognition", A Doctoral Thesis. Submitted in partial fulfillment of the requirements for the award of Doctor of Philosophy of Loughborough University.

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

BIOGRAPHY

Amr M. Godyreceived 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.

Rania Ahmed Abul Seoud 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 Computer networks.

Mai Ezz El-din received the B.Sc. degree in Electrical Engineering – Communications and Electronics Department with very good degree, from the Faculty of Engineering - Misr University for science and technology, Egypt, in 2011.She joined the M.Sc program in Fayoum University - Communications and Electronics Department in 2012.She received the Pre-Master degree in Fayoum University with very good degree, in 2012. Her areas of interest include Best-Tree Encoding model, speech recognition.

التعرف على الكلام تلقائيا في سياق أحادي الصوت بإستخدام خريطة الميل والتشفير الشجري الأمثل

عمرو م. جودي, رانيا أحمد أبو السعود, مي عز الدين

الهندسه الكهربية, كلية الهندسة, جامعة الفيوم, مصر

الملخص العربي

تم تقديم نموذج التشفير الشجري الأمثل (BTE) للمره الأولي من قبل Amr M. Gody والذي يعطي نتائج مبشرة في مشاكل التعرف علي الكلام التلقائي. ويعمل ال BTE في الاساس كمحلل للطيف. ويعتمد علي حزم الموجيات للحصول علي إسقاط الأشاره بإستخدام أنواع من الفلتر بنك تكون محددة سابقا. ويتم ترميز مكونات النموذج في شكل رقمي بإستخدام طريقة entropy معينه وإجراء ترميزي رقمي معين. وقد تم في هذا البحث تطوير ال (BTE) بإضافه إثنان من العوامل الرئيسيه في عملية إنتاج ال (BTE). وهذه العوامل هي: مقياس ميل (MS) وتخطيط النطاق الترددي القاعدي (BM). ويقدم هذا البحث تقيم أداء خط الأساس لتميز وحدة الصوت الأحاديه للسياق المستقل (بدون إستخدام قواعد اللغه) للغه الانجليزية وبإستخدام قاعدة البيانات Vid-TIMIT . وتتكون قاعدة البيانات Vid-TIMIT من 43 متحدث (19 سيدة و 24 رجل), يتحدثوا جمل قصيره. وقد تم تسجيل قاعدة البيانات في بيئة مضاف إليها ضوضاء (ومعظمها ضوضاء مروحة الكومبيوتر) بالإضافة أنها لم يتم مراجعتها يدويا. يستخدم إجمالي 15643 وحدة صوت لتقيم وإختبار النموذج الجديد المطروح. ويستخدم ال HMM كطريقة لتصنيف أدوات ال HTK وذلك لشهرتها الواسعة في مجال ال ASR. ولتقيم نتائج النموذج المطروح قد تم مقارنته بنتائج ال MFCC بعد تطبيقها علي نفس قاعدة البيانات المستخدمه في هذا البحث. وعلي الرغم من أن النموذج المطروح يعطي نفس نتائج ال MFCC عند إستخدام نفس قاعدة البيانات, إلا إنه يوفر أكثر من 66% من التخزين المطلوب مقارنة بال MFCC.