Novel Image PreprocessingApproach for Automatic Speech Recognition

Amr M. Gody[1], Yossra A. Emam[2], Nashaat M. Hussein[3]

Electrical Engineering Department, Faculty of Engineering, Fayoum University
El-FayoumEgypt
[1]amg00@fayoum.edu.eg
[2]eng.ussraemam@yahoo.com
[3]Nmh01@fayoum.edu.eg

Abstract:***This research is intending to provide a novel approach of manipulating automatic speech recognition using image recognition approach. This research introduces hybrid 2D-Image-Hidden Markov Model(2DI)-(HMM) approach to handle preprocessing classification task in Automatic Speech Recognition System (ASR). The focus in this research is in the classification task. Due to that the proposed approach is novel and is a task in the whole ASR, it is evaluated using relative comparison to other popular approaches to run the same task on the same database. The relative comparison with hybrid Gaussian Mixture (GMM)-HMM with Mel Frequency Cepstral (MFCC) features is considered as reference results. This research introduces a new method of mapping speech signal into two-dimensionalspace. Speech stream is segmented and then the frequency contents are projected into frequency domain using a balanced tree structure filter. The wavelet packets technique is used to implement the filtering. The tree structure is captured into image. Database is constructed of encoded images. The imagesthenare segregated into speech classes. Hybrid Discrete Cosine Transform (DCT) based featuresare used for image encoding with (HMM) as Class model is evaluated against MFCC-HMM for the same classification problem. The proposed hybrid model indicates better balanced results over MFCC-HMM for handling the different classes. The considered classes in this research are vowels, consonants, plosives and speech silence.***
KED-TIMITCorpus is used in this research as source of speech information. This approach is indicating promising results especiallyin Silence and vowels detection.

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

1 INTRODUCTION

Automatic speech Recognition (ASR) is the task to convert the speech utterance into a text script. ASR is a challenging task. This research is intending to provide a preprocessing task to enhance the successor task of ASR.
The research Goals in this research paper are:

1- To figure out the use of Two Dimension Image Encoded (2D image) approach in speech recognition.
2- To evaluate the proposed Hybrid model Two Dimension Image Encoded 2DI-HMM with respect to GMM-HMM for handling the same preprocessing classification task.

A. LiteratureReview

Speech processing tasks may be classified intothree categories;speech synthesis, speech encoding and speech recognition. Speech recognition is the process of converting the speech signal into sequence of words or classes. Spokenlanguageconsists of units like words or sub-words calledSyllables. Mono phone and tri phoneare considered examples of sub-word units, recognizing the language unit is the objective of automatic speech recognition.

In this researchthe workis oriented toward speech classification intovowels, consonants, silent and closures sounds. In this research speech duration (frame) is transformed into 2-dimensional image using technique called Best Tree. Best tree is an algorithm that visualizesthe best locations in frequency domain that contains information. This technique is basically relying on the entropy of wavelet packet tree nodes. The entropy is utilized to select the best nodes that represent the signal.This trend was previously introduced by the main author to develop new speech features called Best Tree Encoding (BTE). BTE is best illustrated by the main's author research team in reference paper [1]. The researchintroduced a study and evaluation of context independent phone recognition using BTE.The research providesa comparison against MFCC as evaluation technique.The archived results show that the recognition rateusingthatproposed new features (BTE)is almost approaching the popular MFCC's but it is better than MFCC in memory space needed to store the features vector byaverage saving of 66%.

This promising achievement makes it worthy to try boosting the results by trying to modify the Best Tree Encoding technique. In this researchit is intended to directly apply the encoding technique on the Best tree shape (as an image).
The modern trends in speech recognition, speech stream are being manipulated differently according to its class. The popular classes are Vowels, Consonants, Silence and closures. There is a lot of good research in this area. In the following paragraphs some of the good efforts in this direction will be introduced.

Jinjin Ye in [2] introduceda study in classifying speech phonemes into isolated fricatives,nasal phonemes and vowels, in this research, TIMIT corpus is used. Histogram is introduced to reconstruct phase spaces. To calculate the probability mass estimation of the classifier he uses a classifier called Naïve Bayes which was tested on three males and was trained on six males;maximum achieved success rate for fricatives was 94.44% at phoneme'sh', 57.14% at phoneme 'nx' for nasals and 50.00% at phoneme 'ay' for vowels.These results show that the nasals and vowels need more manipulation using GM.

Jan Macek in [3] introduced a study comparing between Machine learning techniques and HMM method using fricative and vocalic acoustic features.In this research TIMIT corpus was used. HMM shows better results for less skewed data as vocalic features. The accuracy of classification of vocalic feature using HMM classifier reaches 81.4%.It shows thatmachine learning techniques need more manipulation on less skewed data as vocalic where its accuracy is better on more skewed data as fricatives. Its accuracy is 88 % for fricatives.

Jun Wang in [4] introduced a study quantifying the articulatory distinctiveness using a data-driven technique for eleven consonants and eight major vowels based on the movement time series data of lip and tongue for English language. The classification was obtained by using support vector machine and Procrustes analysis techniques. Then the articulatory distinctiveness between consonants and vowels was measured using Procrustes distance. To derive articulatory consonants and vowel spaces,the distance metrics of consonant pairs and vowel pairs were used. The accuracies using support vector machine and Procrustes analysis for consonant classification were 88.94% and91.37% and for vowel were 89.05% and 91.67% respectively.

Ying-Yee Kong in [5] introduced a study investigating for a set of classification methods and acoustic features for three sets of fricative consonants. These three sets are different in articulation's place.TIMIT corpus was used for this research; MFCC was used as a feature. The classification was 85% or greater at +10 dB SNR using 14 or 24 Gammatone filter and 13 MFCC coefficients, and using 14 Gammatone filter with SNRs from +20 to +5 dB SNR, the classification accuracy was greater than 80%.
There is as well a new trend of using more powerful hybrid models to handle ASR. Those models are hybrid of Deep Neural network and HMM or Recurrent Neural network. Those models indicate enhancements over GMM-HMM to handle ASR. Ossama Abdel-Hamid in [6] introduces a concise description of the basic Convolutional Neural Network(CNN) by explaining how it can be used for speech recognition, then a limited-weight-sharing scheme was proposed to better model speech features. Experimental results show that compared with deep neural network (DNNs), CNNs reduce the error rate by 6%-10% on the TIMIT phone recognition . Using hybrid deep neural network (DNN)- hidden Markov model (HMM) shows significant improvement on speech recognition performance over the conventional Gaussian mixture model (GMM)-HMM.
Alex Graves in [7] by using deep recurrent neural networks which combine the multiple levels of representation with the flexible use of long range context that empowers RNNs, deep Long Short-term Memory RNNs achieve a test set error of 17.7% on the TIMIT phoneme recognition. The proposed hybrid 2Dimage -HMM approach is not compared to those promising approaches in the present research. It will be a future work to extend this research by comparing it to both models. In this research the comparison is only run over the ordinary GMM-HMM model.

By the aid of the open-source toolkit Hidden Markov model Tool Kit(HTK), the implementation of machine learning in this paper was processed.All the experiments were limited to KED-TIMIT of English language corpus (453 utterances) which was groupedinto two equal parts as training set and testing set.
Navigating through the sections of this paper,section two discusseddiscrete cosine Transform (DCT) technique, the definition of Mel Frequency Cepstral Coefficients(MFCC), vector quantization technique and best tree and entropy, section three discussed theTest cases and experimental model, section four discussed the results and section five discussed the conclusions.

B. DiscreteCosineTransform(DCT)

In seventies, Ahmed and Rao introduced Discrete Cosine Transform (DCT), after that several versions of DCT have been discussed and it has become popular [8]. To reduce the redundancy, some DCT coefficients were taken. The original image was recovered from these coefficients only. DCT technique is mainly based on converting [9] the image's data to its frequency components. The coefficients in left upper area (upper left corner) in DCT matrix have low frequency(largest magnitude), this area is related to smooth regions and illumination variation.The coefficients in the right bottom area (right bottom area)of the DCT matrix have high frequency (lowest magnitude); this area is related to details information of the image edges and noise. At the middle area the coefficients of medium frequency (medium magnitude) are found, it represents the image general structure, Fig. 1, 2.

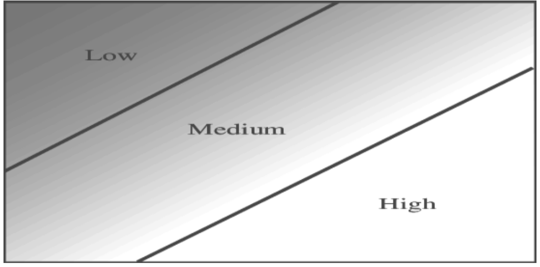


Figure 1: Three regions of DCT coefficient matrix

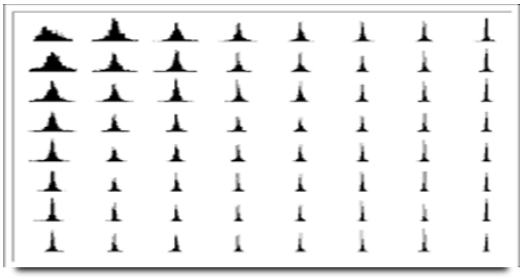


Figure 2: Histogram of DCT coefficients of a 'bridge' image

C. MelFrequencyCepstralCoefficients (MFCC)

Sounds in high frequency scale and low frequency scale have different effectin human's ears response. There is a scale that shows the human ear hearing mechanism called Mel-scale (MS). The linearity in this scale is below 1000 Hz. But at high frequencies the relation is logarithmic as shown in Fig. 3.
The formula which is used for MS (f_{Mel}) is given as following:

$F_{Mel} =2595*\log_{10} (1+(f_{HZ}/700))$ (1),

where:$f_{HZ}$is the frequency in hertz.

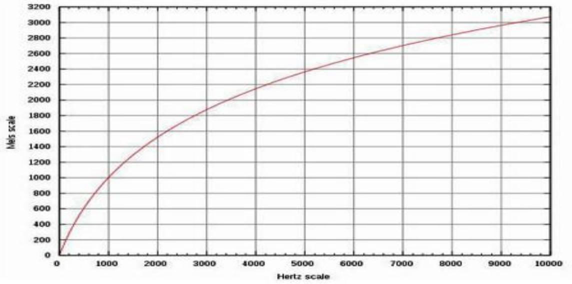


Figure 3: Relationship between the Frequency Scale and Mel-Scale (MS) [10].

By applying the above formula on each frame, Y matrix will be converted to Y_{mel}as follows:
Y_{mel} = mel (Y) = $\{\hat{X}_1 \ldots\hat{X}_M\}$ (2),
The Y_{mel}matrix is applied on wavelet packet decomposition (WPD) filter to have the tree that expresses the speech signal after passing the tree on the entropy function to extract the informative nodes only and have the best tree.

D. VectorQuantizationTechnique

Vector quantization technique is the process of mapping infinite vector quantities with finite vector quantities. This technique is mainly used in speech processing, image processing and signal processing. It is useful in the field of speech coding where the sample is represented by less number of bits so the memory used , bit-rate and complexity gets reduced. Other result of vector quantization is the loss of quality so a great balance must be done to avoid much loss in quality and having a great reduction in bit-rate. Vector quantization and scalar quantization are the two types of vector quantization. The quantization of samples on the basis of sample by sample is called scalar quantization, while quantizing samples in groups called vectors is called vector quantization. This type of quantization increases the quantized optimality by increasing memory requirements and computational complexity.

Vector quantization is considered more effective than the scalar quantization according to Shannon theory and this theory also focuses on increasing the performance of the vector quantization by choosing the best dimension of the vector, where the vectors of larger dimensions get better performance than vectors of smaller dimensions.

As in Fig. 4, Let $S_k$be a set of "N" dimensional input vectors with samples in the range 1≤k≤M is matched with the real valued "N" dimensional code words of the codebook $L = 2^b$. Where "M" is the count of samples, "N" is the vector size of each sample and "b" is the bit count to address "L"code words into the code book"CB". The code word that best matches the input vector and has the lowest distortion is taken and the input vector is replaced by it,where the codebook has a length of L and has a finite set of code words [11].

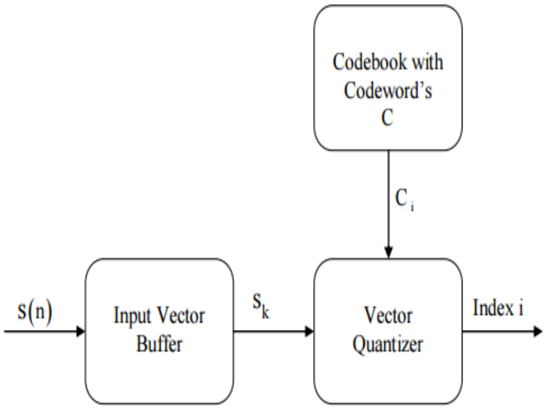


Figure 4: Block diagram of vector quantization

Vector quantization is considered a lossy data compression technique based on block coding and it may be considered also as an approximation technique. The basic definition of vector quantizationcan be illustrated in Fig. 5.

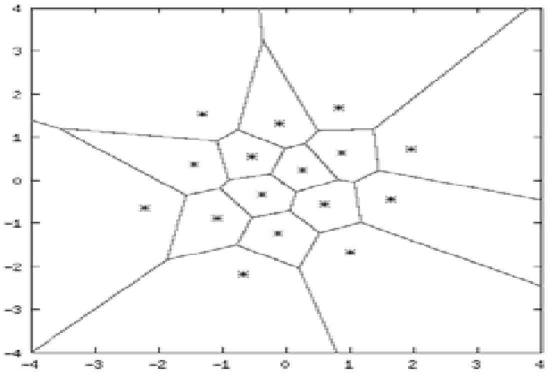


Figure 5: Two dimensional vector quantization example

The numbers that fall in a defined region are signed by a star, these stars are called codevectors. The codebook is all the sets of codevectors and the encoded regions are those regions that are defined by borders. The vector quantization is used here for its high accuracy and its simplicity of implementation [12].

In this research, vector quantization is used in two cases when using MFCCexperiment: (Table 5, case 5) and (Table 4, case 4).

1) *Vector Quantization in MFCC*

When applying vector quantization on MFCC experiment, with codebook size equal to 10 bit, good results were obtained. The results of MFCC experiments indicate that vector quantization achieves the required modification in the result with this high codebook value.

2) *Vector Quantization in Best Tree Encodinginto 2D mono color JPG image (BTEI)*

The resulted feature vector components of each frame in htk file may have some trivial values such as 0 or 1compared to its other values like 3290. These trivial values were removed and the large values were subtracted by 3274 to scale the numbers in the proper value range in htk formatted file. When making vector quantization on the resulted components with codebook size equal to 3 bit, good results were achieved but the increase in the results using vector quantization is not high in its value. This indicates that Vector quantization process achieves little modification in the results in BTEI experiment with low codebook value.

E. Best Treeand Entropy

The proposed model maps the single dimension time waveform speech signal into a two-dimensional image. It is inherited from the Best Tree of the wavelet packets. For continuity, the reference Best Tree Encoding procedure [13]is altered in such that replacing the encoding task with image capturing task. Fig. 6 illustrates the new task procedure after excluding the encoding task from the tasks procedure.

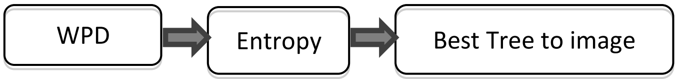


Figure 6: Wavelet packets Best Tree to image tasks procedure[14]

By using Daubechies wavelet filter with four points, in this step the extraction of the spectrum from the time waveform is done as shown in Fig. 7.

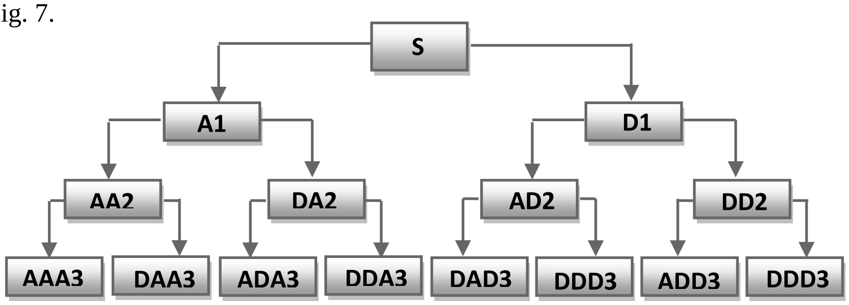


Figure 7: Signal decomposition using wavelet packets [15].

The Entropy is the key step to enhance the tree. The entropy is considered as a measure for the information in each tree node in Fig.8. All low informative nodes will be removed. The type of entropy that is used in this model is called Shannon entropyas given in equation 3[16].

$H(X) = -\Sigma_i p(x_i)\log p(x_i)$ (3),

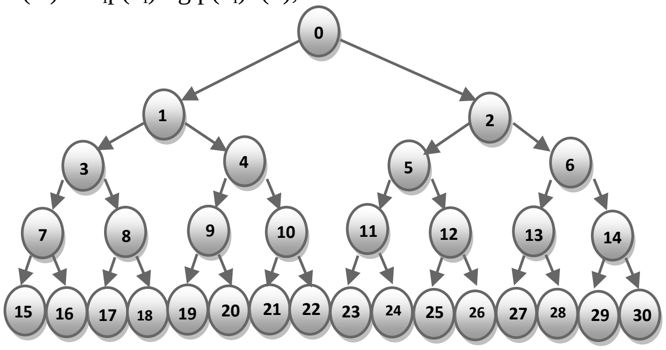


Figure 8: The tree before the cutting

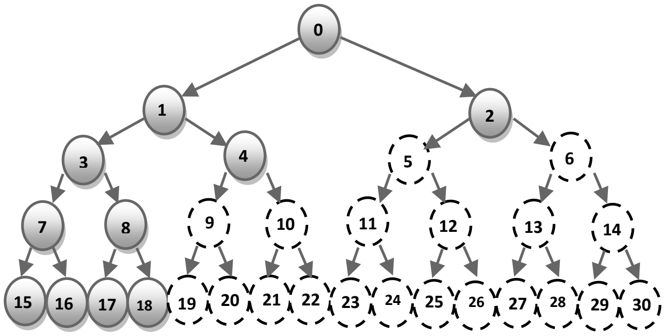


Figure 9:The tree after removing the

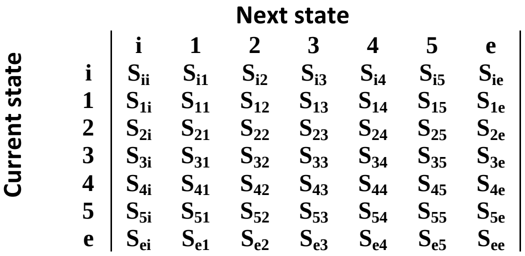
where: p $(x_i)$is the probability of the symbol $x_{i..}$A tree with four levels is the output of WPD process. Every node in the tree can be expressed as a child or a parent. This tree has nodes which have no children. A node that has children isa parent node. By evaluating each node in the tree with entropy, each node has a unique number as an identifier. As shown in Fig.9.

The process neglecting the unnecessary nodes in the binary tree using Shannon entropy is based on comparing the entropyof the parent and its children. If the summation of the entropy of the parent is higher than the two children entropy summation, these two children will be removed from the tree.

F. Machine Learning using HMM

1) Hidden Markov model(HMM)

HMM is an acoustic model that is used in speech recognition process for extracting the best results. Its concept is to estimate the probabilities for a sequence of state events. HMM can be described easily using it transition matrix as shown in Fig. 10[17].



		Next state						
		i	**1**	**2**	**3**	**4**	**5**	**e**
Current state	**i**	S_{ii}	S_{i1}	S_{i2}	S_{i3}	S_{i4}	S_{i5}	S_{ie}
	1	S_{1i}	S_{11}	S_{12}	S_{13}	S_{14}	S_{15}	S_{1e}
	2	S_{2i}	S_{21}	S_{22}	S_{23}	S_{24}	S_{25}	S_{2e}
	3	S_{3i}	S_{31}	S_{32}	S_{33}	S_{34}	S_{35}	S_{3e}
	4	S_{4i}	S_{41}	S_{42}	S_{43}	S_{44}	S_{45}	S_{4e}
	5	S_{5i}	S_{51}	S_{52}	S_{53}	S_{54}	S_{55}	S_{5e}
	e	S_{ei}	S_{e1}	S_{e2}	S_{e3}	S_{e4}	S_{e5}	S_{ee}

Figure 10: The transition matrix

The probability of each state transition to the next state and its state is inserted in this matrix. Assume having transition matrix 7*7 The probability of moving from the state 3 to state 5 is the probability value that can be found in the fourth row and the sixth column in the transition matrix to find the value S_{35} . The first and the last rows represent the initial and the final states respectively.The first state is a starting point and the last state is an exit point. HMM does not have the ability to remain inthe initial state or return again once left it so the first column probabilities are all equal to zero. HMM cannot also transit to any other state once it goes to the final state so the last row probabilities are all equal to zero. So the first and the last states are called non emitting states. The remaining states are called emitting states; fig 11 shows the state diagram for 5 state HMM which illustrates the emitting and non-emitting states.

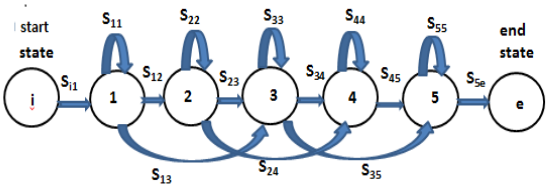


Figure 11: The state diagram for 5 states HMM

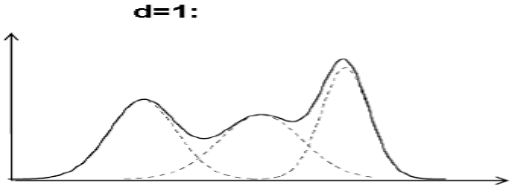


Figure 12: One dimensional Gaussian Mixture Model

In this research all the phones are represented with five states hidden Markov model.

2) Gaussian mixture GM

GM is an estimator model used for the calculation of the probability density function for the statistical system. It builds an arbitrary distribution with many models.GMM is used in many fields such as speech recognition and in musical instruments. Number of mixtures is varying from two to eight in the presentresearch. HMM and Gaussian mixture model are both used for representing the emitting states in speech recognition system using HTK tools[18]. Fig. 12 shows one dimensional Gaussian mixture model.

3) Training and evaluation using HTK

To evaluate HTK steps we must have some important files like having the grammar, the recorded data and the dictionary file. Fig 13 shows the grammar file and Fig 14 shows the dictionary file.

Master label file (MLF) and the label file are also needed files in the experiment. Fig 15 shows the MLF file; Fig. 16 shows the label file. The prototype file is also an important file that illustrates the no. of states used in the experiment. Fig. 17

shows the prototype file. There are some HTK tools that are used for training and recognition processes such as the following tool:

The HTK tool HCompV will scan a set of data files, compute the global mean and variance and set all of the Gaussians in a given HMM to have the same mean and variance.

```
HCompV -C config -f 0.01 -m -S train.scp -M hmm0 proto
```

Hence, assuming that a list of all the training files is stored in train.scp,the command will create a new version of proto in the directory hmm0 in which the zero means and unit variances have been replaced by the global speech means and variances. Hmm0 is usedto re-estimate other HMM models using HERest tool.

```
HERest -A -D -T 1 -C config -I phones0.mlf –S train.scp -H hmm0/macros -H
hmm0/hmmdefs -M hmm1
monophones0
```

HVite tool is used also in recognizing the test data. In the given example,the output file from HVite command is calledrecout.mlf.

```
HVite -A -D -T 1 -H hmm7/macros -H hmm7/hmmdefs -C config–Stest.scp -l '*' -irecout.mlf -w phnet -p 0.0 -s 5.0 dict
monophones1
```

To extract the result and the confusion matrix the following tool will be used:

```
HResults -p -I testref.mlf monophones1 recout.mlf
```

```
 $Phone =  SIL |  P |  C |  V  ;
(SIL < $Phone > SIL)
```

Figure 13:The grammar file

```
SIL         SIL
P           P
C           C
V           V
SENT-END        []   SIL
SENT-START    []   SIL
```

Figure 14: the dictionary file

```
#!MLF!#
"*/kdt_001.lab"
0             3990280          SIL
3990280    4814580          C
4814580    5560380          V
5560380    6041220          C
6041220    6669250          V
6669250    6796820           P
```

Figure 15: The MLF file

```
0              3990280         SIL
3990280   4814580          C
4814580   5560380          V
5560380   6041220          C
6041220   6669250          V
6669250   6796820          P
6796820   6875330          C
6875330   7705120          C
7705120   8525920          C
8525920   8895280          C
8895280   9049180          P
9049180   9264630          C
```

Figure 16:The Label file

```
~h "proto0"
<BeginHMM>
          <VecSize> 9
<user>
          <NumStates> 7
<STATE> 2

<MEAN> 9
0.0 0.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0
<VARIANCE> 9
1.0 1.0 1.0 1.0 1.0 1.0 1.0
1.0 1.0
<STATE> 3
<MEAN> 9
0.0 0.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0
<VARIANCE> 9
1.0 1.0 1.0 1.0 1.0 1.0 1.0
1.0 1.0
<STATE> 4
<MEAN> 9
0.0 0.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0
<VARIANCE> 9
1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0
1.0
<STATE> 5
<MEAN> 9
0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
0.0
<VARIANCE> 9
1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0
1.0
<STATE> 6
<MEAN> 9
0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
0.0
<VARIANCE> 9
1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0
1.0
<TRANSP> 7
0.0 1.0 0.0 0.0 0.0 0.0 0.0
0.0 0.7 0.3 0.0 0.0 0.0 0.0
0.0 0.0 0.7 0.2 0.1 0.0 0.0
0.0 0.0 0.0 0.7 0.3 0.0 0.0
0.0 0.0 0.0 0.0 0.7 0.2 0.1
0.0 0.0 0.0 0.0 0.0 0.7 0.3
0.0 0.0 0.0 0.0 0.0 0.0 0.0
<ENDHMM>
```

Figure 17:the prototype file

2 ENCODING WAVELET PACKETS BEST TREE INTO 2D MONO COLOR IMAGE (BTEI)

The proposed model is illustrated in Fig. 18. The waveform was resampled into 10000 Hz and framed into 20ms frame size. The frequency scale was transformed to Mel scale and applied to WPD (4db filter) followed by Shannon entropy to have the best tree images.Each image was normalized to gray scale and was split into nine parts. Each part was transformed its DCT and then the maximum two values in the resulting DCT coefficients were sorted as columns and normalized to have the feature vector with 18 components. This feature vector was manipulatedby applying it to machine learning using HMM, where all zero or one component values are removed and the remaining components were subtracted by 3274.After this manipulation the vector size will be 9 components only. All these steps will be discussed in details. BTEIexperiment is applied on case1, case 2, case 3 and case 4. In case 1, case 2 and case 3; 2, 4 and 8 Gaussian mixtures were used respectively while in case 4, vector quantized data was used without using Gaussian mixture.

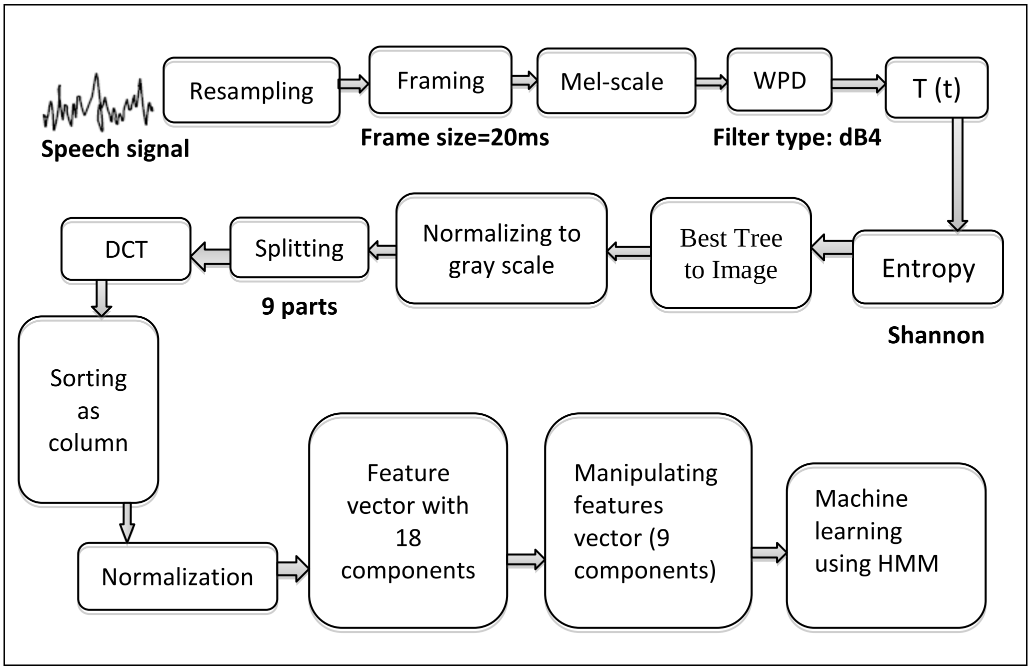


Figure 18: The proposed model

All the images of the best tree which result from the entropy function are normalized to gray scale black and white images and split into 9 parts as shown in Fig. 19.

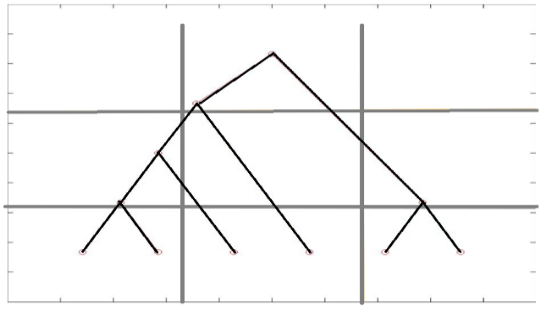


Figure 19: The nine parts of the image

By using discrete Fourier transform equation, each part in the image is treated individually to get the maximum two absolute values as illustrated in the next equation:

V_i= $\max_2 \|DCT\ (A_i)\|$(4),

So the Matrix A will transferred to Matrix V as follows:

$$V = \left\{\begin{matrix} V1\ V2\ V3 \\ V4\ V5\ V6 \\ V7\ V8\ V9 \end{matrix}\right\} \ (5) \ V = \left\{\begin{matrix} V1 \\ . \\ . \\ . \\ V9 \end{matrix}\right\} \qquad (6)$$

Sorting V matrix as a single column vector can be achieved as following:

Because Viis a2×1 vector, so V will be 18 × 1vector.For the image size in pixels of [xy], Where x for rows and y for columns, the coefficients of DCT of each image is computed by the function f (u, v):

$$f(u,v) = \alpha(\mathrm{u})\,\alpha(\mathrm{v}) \sum_{\mathrm{y=0}}^{\mathrm{M-1}} \sum_{\mathrm{x=0}}^{\mathrm{N-1}} \mathrm{f(m, n)} \cos\left({}^{(2\mathrm{m}+1)\Pi\mathrm{u}}\!/_{2\mathrm{M}}\right) \cos\left({}^{(2N+1)\Pi v}\!/_{2M}\right) \qquad (7),$$

where:

$$\alpha(\mathrm{u}) = \begin{cases} {}^{1}\!/_{\sqrt{M}}, & \mathrm{u} = 0 \\ \sqrt{{}^{2}\!/_{M}}, & 1 \le \mathrm{u} \le \mathrm{M}\text{-}1 \end{cases}$$

and

$$\alpha(\mathrm{v}) = \begin{cases} {}^{1}\!/_{\sqrt{N}}, & \mathrm{v} = 0 \\ \sqrt{({}^{2}\!/_{N})}, & 1 \le \mathrm{v} \le \mathrm{N}\text{-}1 \end{cases}$$

Normalization is changing the range of the image pixels; it is called sometimes histogram stretching, contrast stretching or dynamic range expansion [19]. The main idea of normalization process is to transfer the gray scale image of n-dimensional which has intensity values from Min to Max into a new range from new Min to new Max. The first step in calculating the normalization is calculating the norm of the vector $\hat{V}$.

$n = \| \hat{V} \|_2 = \sqrt[2]{\sum_{i=1}^{9} \sum_{j=1}^{2} V_{ij}^2}$ (8), then calculate the minimum value of all elements in $\hat{V}$.

$m = \min \hat{V}$ (9),

Then apply the normalization equation on $\hat{V}$ to extract the features vector as following:

$$Q_i = \left\{ \left(\tfrac{V_{ij}}{n} - m \right) \times 10^4 \right\} \forall_{i,j} = \begin{Bmatrix} F_1 \\ . \\ . \\ . \\ F_{18} \end{Bmatrix} \quad (10),\ F_K = \left(\frac{V_{ij}}{n} - m \right) \times 10^4 \quad (11),\ \text{where: } k = (i-1) \times 2 + j$$

So the complete observation set (Q) of M vectors is $Q = \{Q_1 .....Q_M\}$ (12),

Then this Observation sample will be stored into Hidden Markov Model Tool Kit(HTK) format file for further processing using the HTK.

3 TEST CASES AND EXPERIMENTAL MODEL

A. Case 1 (2D image-HMM)

- Units classification using BTEI, HMM and Gaussian mixture.
- Classified units: Silent (SIL), Vowel (V) such as (iy- ae- ow- uh), closures (P) such as (tcl- gcl- pcl- dcl- bcl- sp- xxx- kcl- glottal-stop), all the remaining phones with letter (C)Consonants[20].
- HMM design: 5 states for all classified units.
- Gaussian Mixtures (GM): 2 in all states.
- Features: BTEI.
- Features vector size: 9 components.

Summary: all speech units are handled the same way. They all are assumed 5 states. Each state is modeled using 2 GMs. The design is described in Fig. 20.

Drawbacks: this model may not be suitable for consonant units. Single state model may be more accurate. But this is not considered in this case.

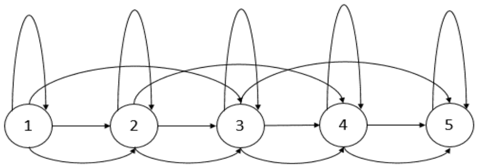


Figure 20: 5-states model for representing classified units

B. Case 2 (2D image-HMM)

- Units classification using BTEI, HMM and Gaussian mixture.
- Classified units: Silent (SIL), Vowel (V) such as (iy- ae- ow- uh), closures(P) such as (tcl- gcl- pcl- dcl- bcl- sp- xxx-kcl- glottal-stop), all the remaining phones with letter (C) Consonants.
- HMM design: 5 states for all classified units.
- Gaussian Mixtures (GM): 4 in all states.
- Features: BTEI.
- Features vector size: 9 components.

Summary: all speech units are handled the same way. They all are assumed 5states;each state is modeled using 4 GMs. The design is described in Fig. 20.

Drawbacks: this model may not be suitable for closures and consonants units.

C. Case 3 (2D image-HMM)

- Units classification using BTEI, HMM and Gaussian mixture.
- Classified units: Silent (SIL), Vowel (V) such as (iy- ae- ow- uh), closures (P) such as (tcl- gcl- pcl- dcl- bcl- sp- xxx-kcl- glottal-stop), all the remaining phones with letter (C) Consonants.
- HMM design: 5 states for all classified units.
- Gaussian Mixtures (GM): 8 in all states.
- Features: BTEI.
- Features vector size: 9 components.

Summary: all speech units are handled the same way. They all are assumed 5 states;each state is modeled using 8 GMs. The design is described in Fig. 20.

Drawbacks: this model may not be suitable for closures.

D. Case 4 (2D image-HMM)

- Units classification using BTEI, HMM and vector quantization
- Classified units: Silent (SIL), Vowel (V) such as (iy- ae- ow- uh), closures (P) such as (tcl- gcl- pcl- dcl- bcl- sp- xxx-kcl- glottal-stop), all the remaining phones with letter (C) Consonants
- HMM design: 5 states for all classified units
- Gaussian Mixtures (GM): single Gaussian in all states
- Vector Quantization (VQ): 3bit codebook
- Features : BTEI
- Features vector size : 1 components

Summary: all speech units are handled the same way. They all are assumed 5 states. Vector quantization was made on all htk files with 3 bit codebook. The design is described in Fig. 20.

Drawbacks: this model may not be suitable for consonants and vowels.

E. Case 5 (MFCC-HMM), The reference results

- Units classification using MFCC, HMM and vector quantization.
- Classified units: Silent (SIL), Vowel (V) such as (iy- ae- ow- uh), closures (P) such as (tcl- gcl- pcl- dcl- bcl- sp- xxx-kcl- glottal-stop), all the remaining phones with letter (C) Consonants.
- HMM design: 5 states for all classified units.
- Gaussian Mixtures (GM): single Gaussian in all states.
- Vector Quantization (VQ): 10 bit codebook.
- Features: MFCC.
- Features vector size: 1 component.

Summary: all speech units are handled the same way. They all are assumed 5 states; Vector quantization was made on all htk files with 10 bit codebook. The design is described in Fig. 20.

Drawbacks: this model may not be suitable for consonants.

F. Case 6 (MFCC-HMM), The reference results

- Units classification using MFCC and HMM.
- Classified units: Silent (SIL), Vowel (V) such as (iy- ae- ow- uh), closures (P) such as (tcl- gcl- pcl- dcl- bcl- sp- xxx-kcl- glottal-stop), all the remaining phones with letter (C) Consonants.
- HMM design: 5 states for all classified units.
- Gaussian Mixtures (GM): single Gaussian in all states.

- Features: MFCC.
- Features vector size: 13 components.

Summary: all speech units are handled the same way. They all are assumed 5 states. The design is described in Fig. 20.
Drawbacks: this model may not be suitable for vowels and closures.

4 RESULTS

The experiments were made on 50% from the database for training and 50% from the database for testing, with five emitting states at each phone in the English KED-TIMIT database[20].

TABLE 1
Confusion Matrix For Five State Model GM=2
(Success Rate= 70.9%) Case 1

symbol	**SIL**	**C**	**V**	**P**	**DEL**	**Total**	**SR**
SIL	239	0	0	0	2	241	99.2
C	549	1180	2037	463	2130	6359	18.6
V	15	0	526	5	20	566	92.9
P	40	2	178	974	142	1336	72.9
INS	88	5	219	37			
SubTotal							70.9

TABLE 2
Confusion Matrix For Five State Model GM=4
(Success Rate= 56.6%) Case 2

symbol	**SIL**	**C**	**V**	**P**	**DEL**	**Total**	**SR**
SIL	237	0	0	0	4	241	98.3
C	318	1039	2796	23	2183	6359	16.3
V	0	0	551	1	14	566	97.3
P	46	1	546	194	549	1336	14.5
INS	28	0	40	2			
Subtotal							56.6

TABLE 3
Confusion Matrix For Five State Model GM=8
(Success Rate=56.6%)Case 3

symbol	**SIL**	**C**	**V**	**P**	**DEL**	**Total**	**SR**
SIL	237	0	1	0	3	241	98.3
C	205	1598	2957	10	1589	6359	25.1
V	1	0	564	0	1	566	99.6
P	27	0	637	46	626	1336	3.4
INS	19	0	46	2			
Subtotal							56.6

TABLE 4
Using Vector Quantization
(Success Rate=43.8 %) Case 4

symbol	**SIL**	**C**	**V**	**P**	**DEL**	**Total**	**SR**
SIL	240	0	0	1	0	241	99.6
C	184	368	0	2573	3234	6359	5.79
V	17	0	0	261	288	566	0
P	7	0	0	932	397	1336	69.8
INS	12	0	0	856			
SubTotal							43.8

TABLE 5
CONFUSION MATRIX FOR MFCC FIVE STATE MODEL, USING VECTOR QUANTIZATION (SUCCESS RATE=72.3 %)CASE5

symbol	**SIL**	**C**	**V**	**P**	**DEL**	**Total**	**SR**
SIL	238	0	0	0	3	241	98.76
C	225	2463	991	0769	1911	6359	38.73
V	15	1	419	26	105	566	74
P	25	5	78	1038	190	1336	77.69
INS	42	9	247	119			
SubTotal							72.3

TABLE 6
CONFUSION MATRIX FOR MFCC FIVE STATE MODEL, NO VECTOR QUANTIZATION (SUCCESS RATE=41.69 %)CASE 6

symbol	**SIL**	**C**	**V**	**P**	**DEL**	**Total**	**SR**
SIL	238	0	0	0	3	241	98.76
C	363	4319	0	0	1677	6359	67.92
V	32	2	0	0	532	566	0
P	58	2	0	1	1275	1336	0.08
INS	13	0	0	0			
Subtotal							41.69

Table 7compares all cases for each class, and Fig. 21visualizes the data in Table 7.

TABLE 7
COMPARISON BETWEEN ALL CASES FOR EACH CLASS, CASE 5 AND CASE 6 ARE GMM-HMM APPROACH

Class	**Case1**	**Case2**	**Case3**	**Case4**	**(Case5*)**	**(Case6*)**	**Best SR**	**Best Case**
SIL	99.2	98.3	98.3	99.6	98.76	98.76	99.6	Case4
C	18.6	16.3	25.1	5.79	38.73	67.92	67.92	Case6
V	92.9	97.3	99.6	0	74	0	99.6	Case3
P	72.9	14.5	3.4	69.8	77.69	0.08	77.69	Case5
Average	70.9	56.6	56.6	43.8	72.3	41.69	72.3	Case5

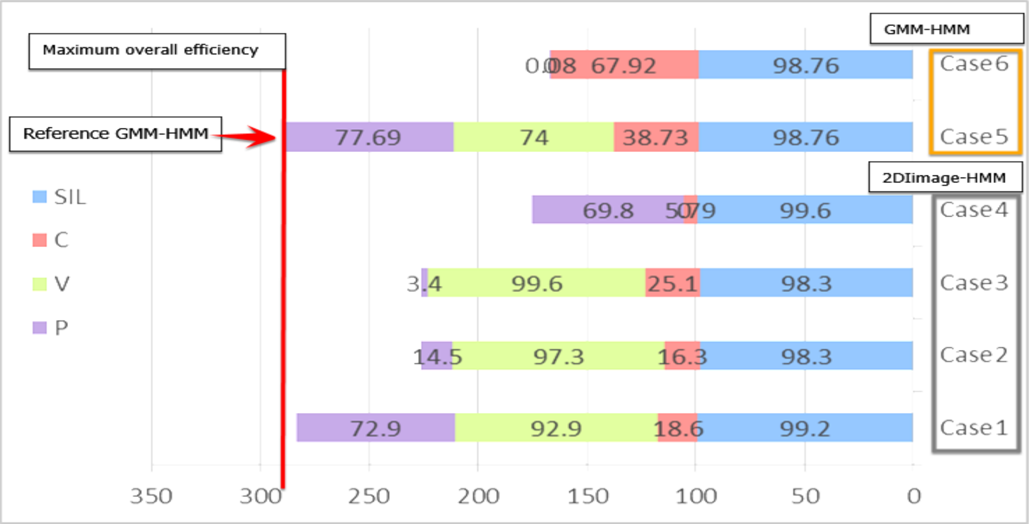


Figure 21: A visualization of the data in Table 7.

In figure 21, case 5 and case 6 are representing GMM-HMM approach. The better is case 5 which gives the best-balanced results over all classes. This case will be considered the reference case in the subsequent comparison and evaluation. The case that has the biggest area for a certain class phone is the better case for this class phone. All cases supplied acceptable results for silence, but the silence's biggest area is found in case 4. The cases that supplied acceptable results for vowelsare case1, case2,case 3 and case5but the vowel's biggest area is found in case 3.Also acceptable results for closuresare found in case 1, case 4 and case 5 but the closure's biggest area is found in case 5. The only good result for consonants is found in case 6. Case 2 and case 3 show the same success rate. The biggest success rate value is found in case 5 as it has the longest bar and the worst one is found in case 6. In addition to that,Using vector quantization gives better results of MFCC experiment because ofgood code book due to many pole vectors so the result of the features is discriminative as in case 5 which gives the bestsuccess rate and the bestresult forclosures and better result for silence and vowels. With increasing of GM vowels gives the best result at GM=8in case 3, while silence and closures give good results at GM=2 in case 1 and the best consonant result was at case 3 with GM =8.

5 CONCLUSIONS

This paper introduces novel approach for preprocessing task that is intending to enhance the overall automatic speech recognition. It introduces hybrid speech-image model. The model uses Discrete Cosine Transform (DCT) and Hidden Marcov Model (HMM). KED-TIMIT database is used in allexperiments. Using five states model for closures such as (gcl- tcl –kcl- pcl- bcl-dcl- glottal-stop- xxx- sp), for silence, for each vowels such as (iy- ae- ow- uh) and for each phone in all residual Englishphones; the proposed modelachieves success rate of70.9%with good recognition for silence, closures and vowels by using GM= 2, success rateof 56.6% when using GM=4 with good recognition for vowels and silence but when using GM=8 the model achieves the same success rate as in case 2 (GM=4) with good recognition for silence and vowels.It is concluded that silent is best detected using minimum count of Gaussian mixtures but vowels are best detected using higher count of Gaussian mixtures.

By applyingvector quantization on (BTEI), the result was significantly degraded from 70.9% to43.8%.But this is not the case whenapplying vector quantization on MFCC. Applying VQ using MFCC featuresenhancesthe achieved overall success rateto72.3% compared with an overall success rate of 41.69% using MFCC without using vector quantization. Two conclusions can be mentionedin this regard. First: VQ enhances the results in MFCC but not in BTEI. This is indicating that the features of BTEI are not good discriminating the classes as of the MFCC does. Comparing the codebook size of 3 bits in case of BTEI to 10 bits in MFCC is an evidence of the concluded statement that BTEI features are less class discrimination than MFCC features. Second: The degradation in success ratewithout using VQ in MFCC is indicating that the Gaussian mixture count set in MFCC experiment is not the best fit or not the best statistical function for MFCC vector of 13 components. HMM model is highly sensitive to Gaussian mixture count. When using VQ this count is set to 1 but is not the case when using the complete features vector of 13 components.

The overall comparison between BTEI and the most popular features MFCC using the same database for the chosen speech unit classificationindicates that BTEI is a promising feature. There are many parameters that can be altered to enhance the efficiency of the proposed BTEI model by modifying the entropy function, changing the vector size, increasing number of maximum selected discrete cosine transform components, increasing the number of divisionsin each image, changing the number of statesthat represent each classand using additional features like delta and acceleration.

REFERENCES

[1] Amr M. Gody, Rania AbulSeoud, and Mai Ezz," Using Mel-Mapped BestTree Encoding for Baseline-Context-Independent-Mono-Phone Automatic Speech Recognition “, *the Egyptian Society Of Language Engineering (ESOLE)*, journal,vol. 2, no.1, pp. 10-24, Month April,2015.

[2] Jinjin Ye, Richard J. Povinelli, Michael T. Johnson.,"Phoneme classification using naïve bayes classifier in reconstructed phase space",*IEEE 10th Digital Signal Processing Workshop and the 2nd Signal Processing Education Workshop*, DOI: 10.1109/DSPWS.2002.1231072,USA, 2002.

[3] Macek, Jan & Kanokphara, Supphanat & Geumann, Anja., "Articulatory-acoustic feature recognition: Comparison of machine learning and HMM methods",*Proceedings of the 10th International Conference on Speech and Computer*, vol. 9 , pp. 99-102, Patras, Greece, 17-19 October 2005.

[4] Wang, J., Green, J. R., Samal, A., & Yunusova, Y., "Articulatory Distinctiveness of Vowels and Consonants: A Data-Driven Approach",*Journal of Speech, Language, and Hearing Research (JSLHR),*vol. 56, no. 5, PP.1539–1551, 2013.

[5] Kong Y-Y, Mullangi A, Kokkinakis K,"Classification of Fricative Consonants for Speech Enhancement in Hearing Devices", *Public Library of ScienceONE (PLoS ONE*), vol.9, no. 4,pp. 80-84, 2014.

[6] Ossama Abdel-Hamid, Abdel-rahman Mohamed, Hui Jiang, Li Deng, Gerald Penn, and Dong Yu, " Convolutional Neural Networks for Speech Recognition", IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 22, NO. 10,pp. 1533-1545, OCTOBER 2014.

[7] Alex Graves, Abdel-rahman Mohamed and Geoffrey Hinton, “SPEECH RECOGNITION WITH DEEP RECURRENT In Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP2013), pp. 6645–6649, 2013.

[8] N. Ahmed, T. Natarajan, K. Rao. , " Discrete Cosine Transform", *IEEE transactions on speech and audio processing,*vol. 23, no.1, pp. 90–93, 1974.

[9] K. Rao, P. Yip, *Discrete Cosine Transform: Algorithms, Advantages, Applications*, 1st ed.,publisher: Academic Press,United Kingdom, 1990.

[10] Appleton, Perera, eds., *The Development and Practice of Electronic Music*,2nd ed., USA, Prentice-Hall, 1975.

[11] A.D. Subramaniam, B.D. Rao, "PDF optimized parametric vector quantization of speech line spectral frequencies", *IEEE transactions on speech and audio processing,*vol. 11, no. 2, pp. 130-142, 2003.

[12] Md. RashidulHasan, Mustafa Jamil, Md. GolamRabbani Md. Saifur Rahman, "Speaker identification using Mel frequency cepstral coefficients", *3rd International Conference on Electrical & Computer Engineering,(ICECE)* , vol. 28, pp. 566-567, Dhaka, Bangladesh, 2004.

[13] Othman Lachhab, Joseph Di Martino, El Hassan Ibn Elhaj, Ahmed Hammouch, "Real Time Context-Independent Phone Recognition Using a Simplified Statistical Training Algorithm", *3rd International Conference on Multimedia Computing and Systems (ICMCS')12*, Morocco, 2012, hal-00761816.

[14] Amr M. Gody, "Wavelet Packets Best Tree 4 Points Encoded (BTE) Features", *The Eighth Conference on Language Engineering*,pp.189-198,Ain-Shams University, Egypt,17-18 December 2008.

[15] Amr M. Gody, Tamer M. Barakat, SayedZaky, "Context Dependent Tri-Phone Automatic Speech Recognition using Novel Spectrum Analysis Approach", *International Journal of Engineering Trends and Technology (IJETT)*, vol.30 , no. 5, pp. 217-222, December 2015.

[16] Jie Wu, Jiasen Sun, Liang Liang, YingchunZha, "Determination of weights for ultimate cross efficiency using Shannon entropy",*Expert Systems with Applications*, vol. 38, no. 5, pp. 5162-5165, 2010.

[17] Mark Hasegawa-Johnson,Hao Tang and Thomas Huang, "A Novel Vector Representation of Stochastic Signals Based on Adapted Ergodic HMMs",*IEEE transactions on speech and audio processing,*vol. 17, no. 8, pp. 715-718, 2010.

[18] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, et al.,*The HTK book*,9nd ed., USA,2006.

[19] A. Al-Haj, "Combined dwt-dct digital image watermarking",*Journal of computer science*, vol. 3, no. 9, pp.172-184,2007.

[20] Alan W Black (1997), KED-TIMIT, Available from:http://festvox.org/dbs/dbs_kdt.html, (Accessed May 2018).

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 the Acting chief of Electrical Engineering department, Fayoum University in 2010, 2012, 2013, 2014 and 2016. His current research areas of interest include speech processing, speech recognition and speech compression. He is author and co-author of many papers in national and international conference proceedings and journals such as Springer(International Journal of Speech Technology),the Egyptian Society of Language Engineering (ESOLE) journal and conferences, International Journal of Engineering Trends and Technology (IJETT), Institute of Electrical and Electronics Engineers (IEEE), International Conference of Signal Processing And Technology (ICSPAT), National Radio Science Conference(NRSC), International Conference on Computer Engineering &System (ICCES) & Conference of Language Engineering(CLE).

Yossra A. Emamreceived the B.Sc. degree in Electrical Engineering – Communications and Electronics Department with very good degree, from the Faculty of Engineering - Fayoum University in 2010. She joined the M.Sc program in Fayoum University - Communications and Electronics Department in 2013 .She received the Pre-Master degree from Fayoum University with very good, in 2014. Her areas of interest include Automatic Speech Segmentation.

N.M.Hussain Hassanreceived his B.Sc.in communication and electronics engineering from Al-Azhar University-Egypt in 2002. In 2005 he received his M. Sc.degree in communication and electronics engineering from(C.N.M.)National Center of Microelectronics, Sevilla University-Spain. In 2009, he received his Ph.D. in Digital Integrated Circuit Design for the Application of Image processing from (C.N.M.) National Center of Microelectronics, Sevilla University- Spain. Currently, he is working as a Lecturer at Fayoum University-Egypt. His research interest includes optimization of digital Image processing techniques such as image compression, enhancement, pattern recognition analysis, edges detections, and image hiding data, Application of these techniques such as artificial vision, smart vision and SLAM system, Biomedical image processing, digital signal processing and hardware implementation such as VHDL, Xilinx and FPGA.

ARABIC ABSTRACT

طريقة مبتكرة للمعالجة الإستباقية للإشارة لغرض التعرف التلقائي على الكلام بإستخدام نموذج يعتمد على معالجة الصور

عمرومحمدجودي[١] , يسرا عبد المنعم إمام[٢] ,نشأت محمد حسين[٣]

قسم الهندسه الكهربية , كليةالهندسة , جامعةالفيوم , مصر

[1]amg00@fayoum.edu.eg

[2]eng.ussraemam@yahoo.com

[3]Nmh01@fayoum.edu.eg

ملخص

يهدف هذا البحث إلي تقديم طريقة مبتكرة للتعرف علي الكلام تلقائيا باستخدام الصور . وهو يقدم موديل لصور ثنائية الأبعاد المدمج مع نظام HMMلعمل نظام تصنيف يستخدم في التعرف علي أصناف الأحرف تلقائيا . يتم تصنيف الأحرف إلي (حروف ساكنة(.Sil)، حروف متحركة(V)، حروف لا تحتوي كلام (P) و حروف ذات طبيعة إنفجارية (C)).وحيث أن التصنيف هو الهدف الرئيسي في هذا البحث ، فإن هذا المنهج يعتبر منهج جديد ومهم في مجال التعرف علي الكلام التلقائي. ويتم تقييم البحث من خلال مقارنته بأبحاث أخري عالمي ة لها نفس الهدف وتستخدم نفس قاعد ة البيانات المستخدم ة أو قواعد بيانات شبيه ة. وتعتبر قاعدة البيانات المسما ةKED-TIMIT هي قاعد ة البيانات المستخدم ة في هذا البحث كمصدر للمعلوم ة الكلامية. وتعتبر مقارنة نظام GMMالمدمج مع نظام HMMبنظام MFCCهدف رئيسي من أهداف هذا البحث. ويقدم البحث طريقة لتحويل إشارة الكلام إلي مجال ثنائي الأبعاد.يتم تقطيع موج ة الكلام ومن ثم يتم إسقاط الموجة إلي مجال التردد باستخدام فلتر لنظام الشجره الناتج . يتم تطبيق نظام Wavelet packetلتطبيق الفلتر. تخزن شكل الشجرة إلي صور لعمل قاعدة البيانات التي سوف يتم إستخدامها فيما بعد. تستخدم هذه الصور فيما بعد في نظام DCTالمدمج مع نظام HMM والذي سوف يتم مقارنته فيما بعد بنظام MFCC-HMM. وقد أعطي نظام DCT-HMM نتائج واعدة عن نظام MFCC_HMM في أصناف الحروف الساكنة والحروف المتحركة.