Broad Phonetic Classification of ASR using Visual Based Features

Doaa A. Lehabik [*1], Mohamed H. Merzban[*2], Sameh F. Saad [**3], Amr M. Gody [*4]

[*]*Electrical Engineering Department, Fayoum University, Fayoum, EGYPT*

[1]da1174@ fayoum.edu.eg

[2]mhm00@fayoum.edu.eg

[4]amg00@fayoum.edu.eg

[**]*Modern Sciences and Arts University, 6 October City, Giza, Egypt*

[3]dr.sam.far@gmail.com

Abstract*: This paper presents a novel method of classifying speech phonemes. Four hybrid techniques based on the acoustic-phonetic approach and pattern recognition approach are used to emphasize the principle idea of this research. The first hybrid model is constructed of fixed state, structured Hidden Markov Model, Gaussian Mixture, Mel scaled Best Tree Image, Convolution Neural network, Vector Quantization (FS-HMM-GM-MBTI-CNN-VQ). The second hybrid model is constructed of variable state, dynamically structured Hidden Markov Model, Gaussian Mixture, Mel scaled Best Tree Image, Convolution Neural network, Vector Quantization (VS-HMM-GM-MBTI-CNN-VQ). The third hybrid model is constructed of fixed state, structured Hidden Markov Model, Gaussian Mixture, Mel scaled Best Tree Image, Convolution Neural network (FS-HMM-GM-MBTI-CNN). The fourth hybrid model is constructed of variable state, dynamically structured Hidden Markov Model, Gaussian Mixture, Mel scaled Best Tree Image, Convolution Neural network (VS-HMM-GM-MBTI-CNN). TIMIT database is used in this paper. All phones are classified into five classes and segregated into Vowels, Plosives, Fricatives, Nasals, and Silences. The results show that using (VS-HMM-GM-MBTI-CNN-VQ) is an available method for classification of phonemes, with the potential for use in applications such as automatic speech recognition and automatic language identification. Competitive results are achieved especially in nasals, plosives, and silence high successive rates than others.*

Keywords: *ASR, Classification technique, HTK, Wavelet Packet Decomposition, Convolution Neural Network, Vector Quantization, Hidden Markov Model*

1 INTRODUCTION

Speech is the most competent and popular means of human communication which is produced as a sequence of phonemes. From these phonemes, we extract features vector which is necessary for the classification method. This classification of sounds is implemented for more applications like speech recognition and language recognition. The broad phone classes are usually known as vowels, plosives, fricatives, nasals, and silence. This categorization can improve speech recognition and hence categorization techniques were attempted. The current research presents new classification techniques. There are hybrid features models used in this technique that consist of MBTI, CNN and or without VQ. A fixed and variable state Hidden Markov model with various Gaussian mixture numbers is used to get a higher rate of recognition. The subsequent sections will explain the details of this research. Section 2 discusses previous related work. Section 3 illustrates each type of the proposed model structure. Section 4 shows an Experiment environment that contains database and experiment procedure. The results would be presented in section 5 and conclusions would be presented in section 6.

2 RELATED WORK

The essential mission of the acoustic model in speech recognition is to know the exact (phone, phoneme, sub-word or word) class for any frame of the voiced signal. The phoneme can be defined as the smallest phonetic unit in a language that is capable of conveying a distinction in meaning. Phonemes that may be within the same class include very similar temporal properties and can be simply confused.

J. Ye et. al. [1] proposed a novel method for classifying the speech phonemes based on histograms. The proposed method classifies phonemes to fricative, vowel, and nasal in TIMIT database. The results showed that a reconstructed phase space approach is a specific method for classification and it achieved overall recognition rates of 61.59%, 34.49% and 30.21% for fricative, vowel, and nasal phonemes respectively.

T. Jeff Reynolds et al. [2] introduced research into the classification of the speech signal into seven classes. These classes are fricatives, semi-vowels, diphthongs, plosives, nasals, closures, and vowels in TIMIT database. MFCC, PLP and LPC were three feature extraction techniques that were gathering to perform this work. HMM and Multi-Layer Perceptron (MLP) were performed in this survey. The highest rate of recognition was achieved by gathering MFCC, PLP, and LPC. The achieved phone classification rate was 84.1%.

P. Scanlon et al. [3] investigated expert classifiers specific to each broad phonetic class and performed phonetic classification by combining scores from the different experts. Classifying the speech signal into vowels, stops, fricatives, and nasals was done in TIMIT database. PLP with first and second derivatives was used in this system. The classification was performed using MLP and HMM. The highest obtained phone accuracy was 74.2%.
H. B. Kekre et al. [4] proposed a technique for isolated word recognition depending on zero-crossing features and energy with vector quantization. In feature extraction, noise is deleted by using the endpoint detection algorithm and also endpoints are detected. The database consists of ten words and the number of sample utterances per word is 20. The maximum recognition rate is obtained for a codebook size of 4 as 85%, while recognition rate drops as codebook size is increased above 4.
G. Kiss et al. [5] introduced research to segment and classify the speech signal into nine classes. MRBA, KIEL, and TIMIT were used as the corpus in this study. Feature vectors were determined using Bark-scale spectral resolution. TIMIT corpus achieved the highest average of classification accuracy with 80% and the confusion matrix showed that 90% success is in low-middle and high vowel recognition.
Deekshitha et al. [6] presents a new classification for broad phonemes by features that are obtained immediately from a speech at the level of this signal. Broad phoneme classes comprise vowels, nasals, fricatives, stops, approximants, and silence. This classification is applied to three systems, each system is applied to three tests and results are 54%, 61%, and 46% for the combination on TEST 1, TEST 2 and TEST 3 respectively.
A. Chittora et al. [7] proposed a novel method to classify the phonemes in Gujarati language by utilizing a modulation spectrogram to extract features. These phonemes were divided into vowels, semivowels, affricates, fricatives, stops, and nasals that have been classified utilizing support vector machine (SVM). Mean classification accuracy is 95.70 % when using a combination model of Phoneme with a fusion of MFCC and proposed features as a features vector.
Nasereddin et al. [8] proposed research for classifying speech signals into 4 classes using HMM, Dynamic time warping (DTW) and Dynamic Bayesian Network (DBN) with MFCC feature extraction. DBN outperformed in recognizing one class while HMM is achieving higher recognition rate for the other classes.
S. Salim et al. [9] presented a two-stage system for spotting the boundaries of vowels, nasals, and approximants in Malayalam [10] speech signal. In the first stage, ANN is used to classify a speech signal that is classified into six broad phoneme classes. For the second stage, the frequency domain parameter named spectral peak frequency is suggested for accurate verification of nasals. Sonorant and non-syllabic features are used for verifying approximants and the syllabic feature is used for locating vowels.
Deekshitha G. et al. [11] presented a novel method for spotting the fricative and plosive regions from continuous speech. A two-stage recognition system is designed for spotting and verifying the fricative and plosive region. In the first stage, a DNN is used and Broad classifiers are Silence, Vowel, Nasal, Fricative, and Plosive. In the second stage, a spectral centroid is used to verify the fricative regions and the difference in the spectral spread is used to verify the plosive regions. The verification results in TIMIT database are 78.37% for Fricative and 68.75% for Plosive.

3 PROPOSED MODEL

This research submits a new classification technique in real-time that depends on creating new hybrid features and developing HMM according to the property of each class. There are four hybrid techniques that consist of (MBTI, CNN, VQ) and (HMM-GMM) to improve the performance of automatic speech recognition. In this model presented in Fig. 1, the input speech signal was resampled into 10 kHz to best distribute the wavelet tree structure through the significant bands as in Fig. 2, where the speech signal frequency band reaches to 4 kHz bandwidth (BW) (8 kHz sampling by Nyquist rate). If the signal is resampled by 32 kHz (16 kHz BW), the signal will suffer from noise as in Fig. 3 which indicates that the right side of this figure will be the same for all features (noise) and left side inside rectangle will change for each feature. So, we resample the signal to 10 kHz (5 kHz BW). Then framing it into small frames 20ms. Wavelet packet decomposition is used to extract the features from speech signal. The information is expressed as a two dominations image using best tree algorithm to keep the leave nodes of wavelets of high informative. All images are normalized as undying both the background color and image size. The tree image is drawn to fill the space of the image rectangle. Mel Best Tree image (MBTI) features would be normalized to vectors by Convolution Neural Network (CNN) then they enter on Vector Quantization (VQ) to extract the final features. These features enter on HMM with various GMM to analyze speech signals into five classes; Vowels (V), Plosives (P), Fricatives (F), Nasals (N) and Silences (Si). The state's number of HMM is fixed/variable for each class. Now, we will explain the details of each block of this model in the following subsection.

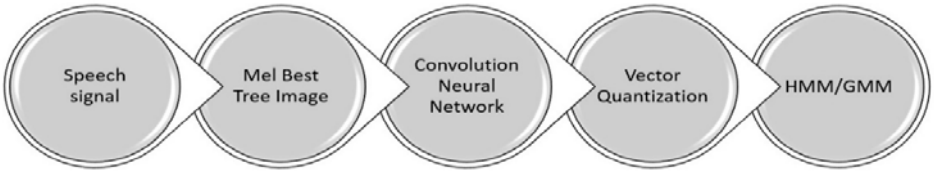


Figure 1: Block diagram of the proposed model

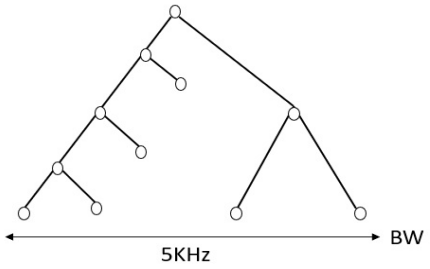


Figure 2: Mel Best Tree Image Feature that resampled in 10 kHz

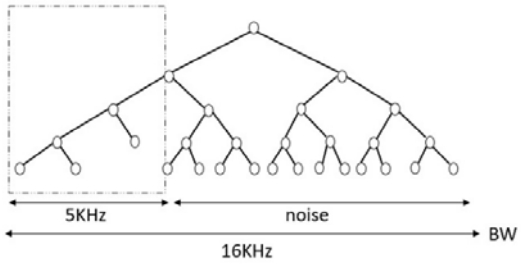
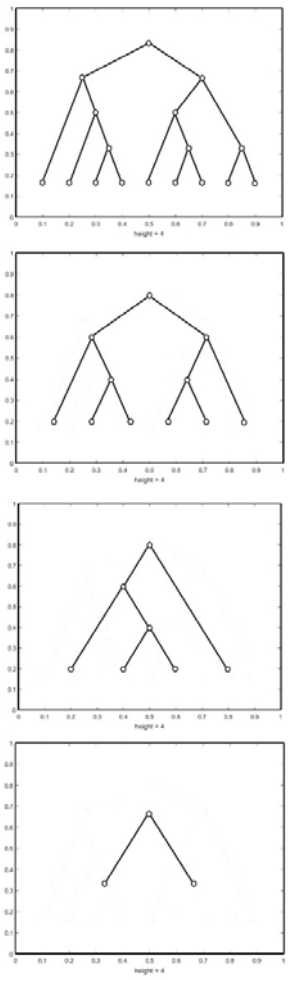


Figure 3: Mel Best Tree Image Feature that resampled in 32 kHz

TABLE I

EXAMPLE OF FEATURES OF ONE PHONE

class	Features from MBTI	Features from CNN				Features from VQ
Fricatives (sh)		-2.3569067	-2.3714445	-2.0820644	-2.1679559	801.000 646.000 209.000 194.000
		-1.0763184	-1.3239728	-1.4214518	-1.1263984	
		-2.0224409	-1.8523995	-1.8592288	-1.8750216	
		-4.3738341	-3.9820654	-3.8391368	-4.2727218	
		-0.35275137	-0.3899605	0.15108292	-0.3492497	
		1.8212616	1.6568279	1.6321151	1.0864879	
		-0.97591358	-0.8605017	-0.60308141	-1.5194747	
		-2.4071643	-1.7821561	-2.6295366	-2.0174067	
		-0.77361774	-0.8566076	-1.3336716	-1.1307474	
		-1.8448237	-1.2801373	-1.7043303	-1.5256805	
		-2.4274793	-2.0159822	-2.1782598	-2.0840702	
		-3.0980186	-2.1938372	-2.6586254	-2.6794515	
		-1.4106346	-1.3592287	-1.5418364	-1.4445210	
		-2.1357713	-1.7336098	-2.1866767	-2.1582818	
		-1.4004182	-1.0478606	-1.4029909	-1.3728735	
		0.53507209	0.73674589	0.67278725	0.55570894	
		-0.72800702	-0.9399060	-1.2511407	-0.9930871	
		-1.4832917	-1.3540589	-1.4416648	-1.3991598	
		0.052903991	0.15688454	0.080595233	-0.0328635	
		-1.9791033	-1.7657653	-1.4535840	-1.8700010	
		.	.	.	.	
		.	.	.	.	
		.	.	.	.	
		.	.	.	.	
		.	.	.	.	
		.	.	.	.	
		.	.	.	.	
		.	.	.	.	
		.	.	.	.	
		2.4473979	2.4940410	2.2598472	2.4556794	

I. Mel Best Tree Image (MBTI)

Best Tree Encoding feature (BTE) was first introduced by Amr M. Gody in [12] that depends on wavelet packet decomposition (WPD). BTE simulates the human hearing mechanism by handling the received speech. The procedure of extracting Best Tree Image (BTI) features would be illustrated through the block diagram in Fig. 4.

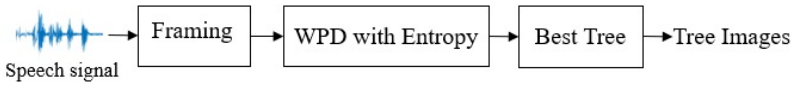


Figure 2: Block diagram of creating BTI

1) Framing and Windowing

Framing is converting the speech signal into a collection of short frames in range 20 to 40ms as the speech is a non-stationary signal. The frame length is set to 20ms (200 samples) to guarantee stationarity (Signals whose frequency content does not alter in time) inside the frame. Then applying a Hamming window that is a rectangular pulse whose width is equal to the frame length to make a smooth transition to the signal to be continuous.

2) Wavelet packet decomposition with Entropy

Wavelet is a short duration waveform that can express any function by scaling and shifting the certain original signal. Then apply the signal into a high pass filter and low pass filter. Then do the same thing again on each portion of the original signal. This operation is called decomposition. This process continues to level 4 and the output is a tree structure.

Entropy is the key step to improve BTI. It is used to measure the information in each tree node. Accordingly, the best tree is done by eliminating all low information on tree nodes. the entropy of the Wavelet Packet Decomposition (WPD) coefficients is applied as a projection of these frames of the speech signal power into defined filter banks. In the original BTI Shannon entropy is chosen. This was implemented using MATLAB as the following command.

$$t = wpdec(frame, 4, 'db4', 'shannon');$$

3) Best Tree

The Best tree function in [13] uses the entropy to calculate the low information tree nodes. The Best tree selection model in details is found in [13]. Starting from the higher-level tree nodes which every 2 nodes have one parent node. If the entropy of the parent node was higher than the sum of entropies of both Childs, then Childs would be removed. This process would repeat until the end. Keep in mind that each tree node represented by a single frequency band and the signal projection on the essential frequency band is considered as the component of each node.

There is another version of BTE that is MBTE introduced in [14] but in this research, we used MBTI. First, the input speech signal is resampled at 10 kHz. Second, converting these frequencies to Mel frequencies by using the Mel scale curve. Third, Generate EWPBTE matrix from Mel scaled data vectors by using a filter bank matrix with 50 linearly spaced filter banks which are overlapped by 50%. The filter's shape is a rectangle window. The output features are MBTIs as in Fig. 5.

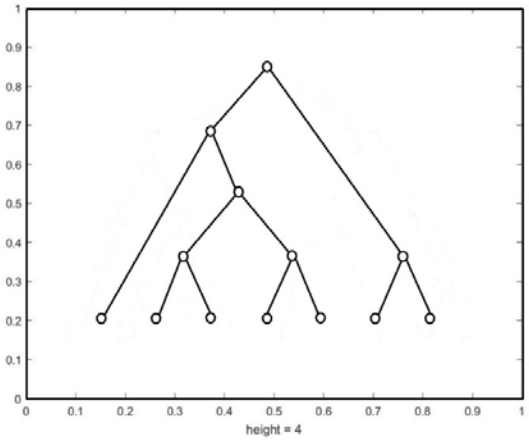


Figure 3: Example on Mel Best Tree Image feature

II. Convolution Neural Network

Convolutional neural network works based on basic neural networks [15] whereas all experiments are conducted under the concept of deep neural network and hidden Markov model where a DNN or CNN. In this research, we choose the deep residual network (Resnet50) that acts as a type of CNN with 50 layers as in Fig. 6. This network can organize image into 1000 object classifications. It has a size for an input image of [224 224 3]. The output feature vector is included in 1000 features of this image. We can run this network on the Graphics processing unit (GPU) by using the tools in MATLAB.

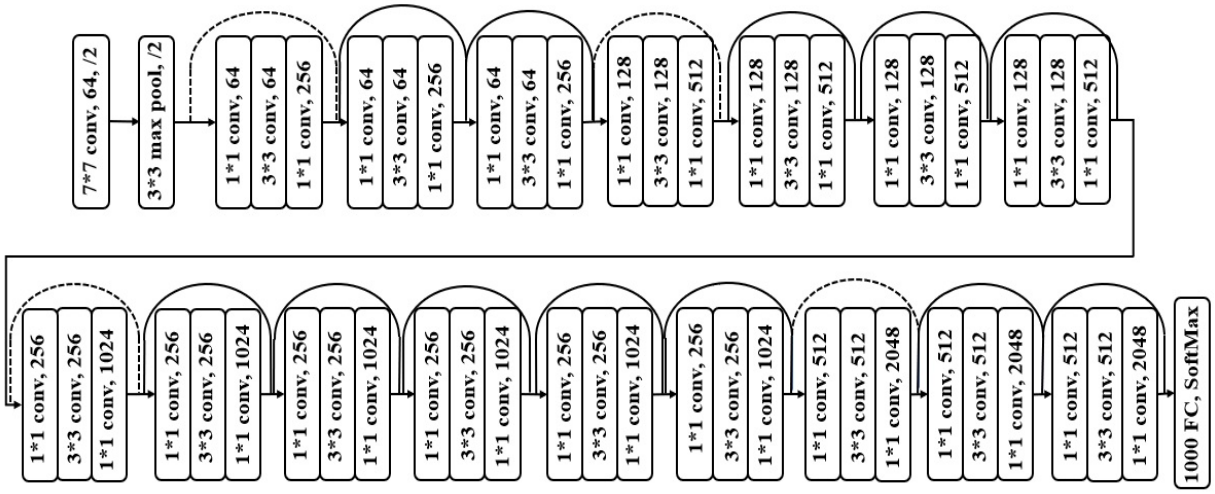


Figure 4: Architecture of Resnet 50

III. Vector Quantization

Vector Quantization (VQ) is a technique of mapping vectors from a general vector space to limited regions in that space. Each region was recognized as a cluster centered by a codeword. A codebook is gathered from codewords. After the feature vectors extracted from input speech provide a set of training vectors. These training vectors are used to create the VQ codebook. There is a popular algorithm that is LBG algorithm [Linde, Buzo, and Gray]. To cluster a combination of L training vectors into a combination of M codebook vectors. The algorithm is introduced by the subsequent steps [16]. First, make a one vector codebook. Second, splitting the current codebook to duplicate the size of the codebook. Third, find the codeword for each training vector in the current codebook that is closest. Fourth, renew the codeword in any cell by utilizing the centroid of the training vectors allocated to the cell. Fifth, duplicate 3 and 4 until the average distance drops under a preset threshold. Sixth, repeat 2, 3 and 4 until a codebook size of M is designed.

B. Hidden Markov Model with Gaussian Mixture Model

Hidden Markov Model (HMM) is the strongest method used in automatic speech recognition. This system is produced for the Markov process with private parameters and we want to distinguish the hidden parameters from the observation. The states are hidden, and the probability distribution for each is known as the variable which affects the states. Temporal data and states are usually identified as separate GMMs [17] in the HMM model as in Fig. 7. The transition matrix learns from training data and it is known to the transition of state to another [18].

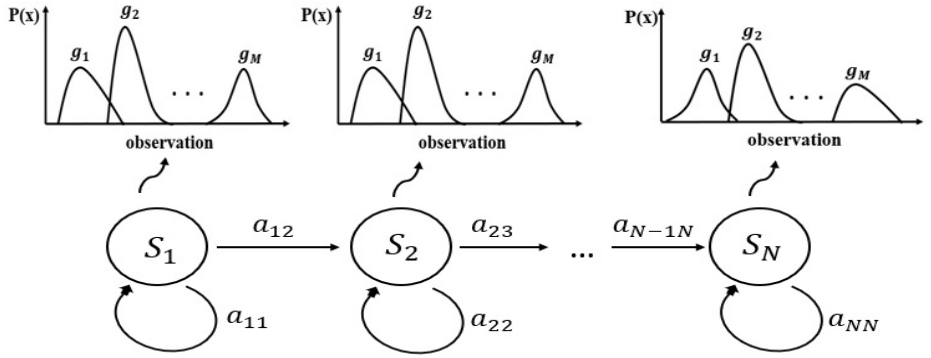


Figure 5: Architecture of HMM with GMM

4 EXPERIMENT ENVIRONMENT

The following subsections explain the type of database that is used in this paper. The process of converting raw speech into features used in this classification and verification. MATLAB 2018b and visual studio 2015 are used as lab environment. The specifications of the laptop that is used in the experiment are 8.00 GB RAM, 64-bit operating system, Intel(R) Core (TM) i5-8250U CPU @1.60 GHz 1.80 GHz and NVIDIA GeForce MX150 with 8061 MB Memory @4 GHz.

A. Database

The continuous Corpus of TIMIT [19] is an acoustic-sounding speech made of English, recorded by a microphone at 16 kHz and 16-bit resolution. This database holds 6300 sentences (5.4 hours) in 630 speakers from 8 regional dialects of the United States (US). Each speaker articulated 10 sentences and all the sentences were identified with its phone level. The main version of TIMIT includes 61 phones. The database is prepared to modify transcription files for the character recognition objective of this research. Vowels (V), Plosives (P), Friction (F), Nasal (N) and Silence (Si) as in [11]. The Table 2 shows each classifier with phones assigned to it.

TABLE II

PHONES CLASSIFIERS

Classifiers	TIMIT Labels
Vowels (V)	aa, ae, ah, ao, ax, ax-h, axr, ay, aw, eh, el, er, ey, ih, ix, iy, l, ow, oy, r, uh, uw, ux, w, y
Plosives (P)	p, t, k, b, d, g, jh, ch, bcl, dcl, gcl, pcl, tcl, kcl, q, dx
Fricatives (F)	s, sh, z, zh, f, th, v, dh, hh, hv
Nasals (N)	m, em, n, nx, ng, eng, en
Silences (Si)	h#, epi, pau

B. Procedure of proposed model

This model aims to provide a new approach to the manipulation of automatic identification of speech using an image recognition method in real-time. This research presents a unique way to produce the speech signal in a two-dimensional space by using one type of neural network for extracting the features. Using Graphic processor units (GPU) to train the neural networks containing large data sets in less time and with less computing infrastructure. Now, we will introduce the steps of this model

- TIMIT database is used as an input speech signal and enters in the MBTI block as in Fig. 1.
- In MBTI block; reads the speech signal and resamples it into 10 kHz. These samples are framed into 20ms which each frame contains 200 samples. Then apply the Mel-scale curve to convert all frequencies to Mel frequencies [20]. Then apply WPD with Shannon entropy to extract the best tree features [12]. The output of this block Mel Best Tree Images (MBTI) features.
- MBTIs are entered as input into the CNN block. In this block; we used (Resnet-50) Network in MATLAB to extract features from images. We apply this in each wav-file and images are the frames of this wav-file that are extracted from MBTI block. The size of these images is [244 244 3] (this size from the properties of resnet-50 network). The output features vector consists of 1000 components.
- These features vectors (for one wav-file) enter on vector quantization block which converts 1000 component features vector into the best one component feature vector. This algorithm was applied to all vectors.
- After that, save these features vectors of wav-file into HTK file format.
- Now the output is HTK file features. Then HTK toolkit will be used for building HMM-GMM based acoustic model for automatic speech recognition.

There are two models of HMM: in the first model; all classifiers are trained using the same fixed state HMM as shown in Fig. 8. This is contained in six states of which the first and the end are non-emitting states. The non-emitting states are needed to identify the entry and exit states in HMM model. We choose four emitting states because of the best implementation for the Fricatives phones that have a long duration. In the second model; each classifier is trained by its own number of states. Vowels and nasals have a duration that is less than fricatives so three emitting states HMM are suitable to be applied as in Fig. 9. Plosives have a short duration (little number of frames) so they are modeled using one emitting state HMM as in Fig. 10. Fricatives are modeled by four emitting states HMM as in Fig. 8. Silences are modeled using two emitting states HMM with the back transition because silence phones can take more/small duration that depends on the phone. The back transition is found to allow long or small time. If we are now in state 3, there is a chance to move to state 4 by 0.1 probability, remain in state 3 by 0.7 probability or move back to state 2 by 0.2 probability to make long duration as in Fig. 11.

Gaussian mixtures with different counts are considered to construct the observation symbol probability function. The system is tested against different Gaussian mixture counts (1, 2, 4, 8, 16 and 64). There is another two model, that is designed without using vector quantization. The feature vector that was obtained from CNN (Resnet50) contains 1000 components. Then this vector enters on two designs of HMM to make recognition. The first design used the same fixed state HMM for all classifiers as shown in Fig. 9. Whereas we used three emitting states instead of four as in the first model. There is a difference between the first and third models. Now, the features vector consists of 1000 instead of one so, three emitting states are enough. In the second design, we used a variable state for each classifier. Vowels and nasals are modeled using three emitting states as in Fig. 9. Plosives are modeled using one emitting state as in Fig. 10, fricatives are modeled by four emitting states as in Fig. 12. The change in the values of the transition matrix would obtain accurate values of the classification. Silences are modeled using two states as in Fig. 13.

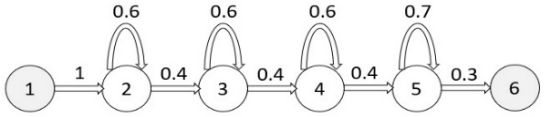


Figure 8: Fixed states HMM for all classifiers model in first model

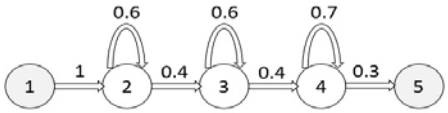


Figure 9: Vowel and Nasals design in second model

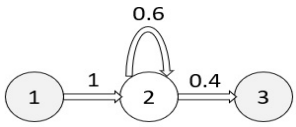


Figure 10: Plosive design in second model and fourth model

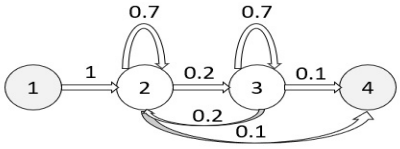


Figure 11: Silence design in second model

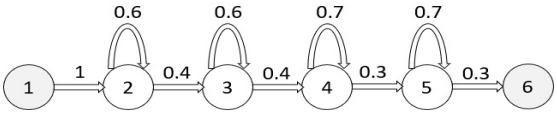


Figure 12: Fricatives design in fourth model

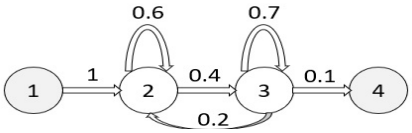


Figure 13: Silence design in fourth model

5 RESULTS AND DISCUSSIONS

There are two models of features vector that are implemented in this research. First, MBTI with CNN and VQ is used and this vector consists of one component. Then case 1: apply it in fixed HMM states and case 2: apply it in dynamic HMM states. Second, MBTI with CNN is used and this vector consists of 1000 components. Then case 3: apply them in fixed HMM states and case 4: apply them in dynamic HMM states. The comparative study is implemented to show the details and the key power in each specific feature set. We calculate the success rate (SR) in each case as in table 2 The success rate can be defined by equation 1 and the result is shown as in Fig. 14 that shows the value of each SR against the Gaussian mixture model (GMM) In this equation: (D denotes deletions), (S denotes substitution) and (N denotes the number of phones in the expected transcription). The following tables represent the result of five classifiers in four cases with different numbers of Gaussian mixture (1, 2, 4, 8,16 and 64).

- Case 1: Using MBTI, CNN and VQ features in a fixed number of states for HMM.
- Case 2: Using MBTI, CNN and VQ features in a dynamic number of states for HMM.
- Case 3: Using MBTI and CNN features in a fixed number of states for HMM.
- Case 4: Using MBTI and CNN features in a dynamic number of states for HMM.

$$SR = \frac{N - D - S}{N} \qquad (1)$$

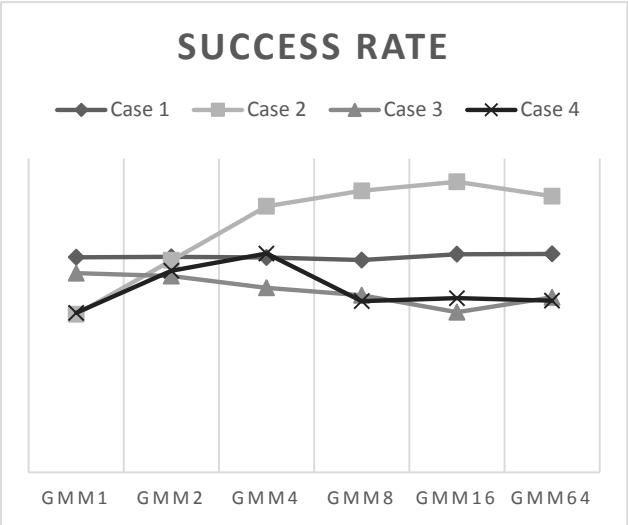


Figure 14: Success rate results

TABLE III

EXPERIMENT RESULTS OF SUCCESS RATE

Mixture Count	Case 1 SR%	Case 2 SR%	Case 3 SR%	Case 4 SR%
1	54.87	40.31	**50.79**	40.69
2	54.96	54.05	50.06	51.41
4	54.80	67.85	47.07	**55.81**
8	54.14	71.81	45.18	43.66
16	55.64	**74.11**	40.87	44.44
64	**55.72**	70.43	44.63	43.81

The results in Table 3 shows that in case 2 (VS-HMM-GM-MBTI-CNN-VQ) using Gaussian mixture number 16 achieved 74.11% success rate. This success rate is the highest rate obtained using the two techniques of features extraction by fixed and variable states of Hidden Markov Model. Table 4 shows the confusion matrix and the highest success rate for each classifier when using vector quantization as in case1 and case 2. In case 1: vowels achieved 81.3% using gaussian number 16, plosives achieved 62.3% using gaussian number 64, fricatives achieved 40.6% using gaussian number 8, nasals achieved 94.9% using gaussian number 8 and silences achieved 87.7% using gaussian number 64 as in Fig 15. In case 2: vowels achieved 82.4% using gaussian number 64, plosives achieved 92.5% using gaussian number 4, fricatives achieved 86.9% using gaussian number 64, nasals achieved 99.3% using gaussian number 1 and silences achieved 84.3% using gaussian number 2 as in Fig 16.

Table 5 indicates the confusion matrix for all classifiers when not using vector quantization in feature extraction as in case 3 and case 4. In case 3: vowels achieved 73.4% using gaussian number 1, plosives achieved 89% using gaussian number 1, fricatives achieved 60.6% using gaussian number 16, nasals achieved 80.9% using gaussian number 4 and silences achieved 99% using gaussian number 16 as in Fig 17. In case 4: vowels achieved 74.6% using gaussian number 4, plosives achieved 69% using gaussian number 4, fricatives achieved 83.3% using gaussian number 1, nasals achieved 89.2% using gaussian number 2 and silences achieved 98.6% using gaussian number 1 as in Fig 18.

TABLE IV

CONFUSION MATRIX AND SUCCESSIVE RATE FOR EACH CLASSIFIER IN CASE 1 AND CASE 2

GMM		Confusion Matrix											
		Fixed states HMM and using vector quantization						Dynamic states HMM and using vector quantization					
		V	P	F	N	Si	SR	V	P	F	N	Si	SR
1	V	**18006**	1320	52	5462	0	72.5	**10818**	306	464	10460	0	49.1
	P	1439	**9207**	38	4242	0	61.7	247	**4576**	450	8776	0	32.6
	F	812	870	**285**	2981	0	5.8	117	182	**2404**	3590	0	38.2
	N	184	95	6	**4337**	0	93.8	6	20	7	**4697**	0	**99.3**
	Si	101	143	1	345	**3360**	85.1	15	25	16	602	**3361**	83.6
2	V	**18408**	988	253	5297	0	73.8	**13539**	1139	626	7718	10	58.8
	P	1579	**7906**	240	4585	0	55.2	343	**10376**	441	4313	3	67
	F	784	649	**1291**	2639	0	24.1	211	577	**2688**	2859	0	42.4
	N	232	76	17	**4288**	0	93	29	57	13	**4698**	0	97.9
	Si	118	92	10	352	**3360**	85.5	38	85	43	461	**3368**	**84.3**
4	V	**18927**	766	391	5118	0	75.1	**15692**	4673	1180	3345	39	62.9
	P	1689	**6638**	437	4888	0	48.6	295	**15977**	308	681	7	**92.5**
	F	766	429	**1946**	2464	0	34.7	295	1402	**4357**	1101	8	60.8
	N	190	71	30	**4279**	0	93.6	74	354	94	**4118**	1	88.7
	Si	128	73	21	352	**3360**	85.4	63	361	49	223	**3379**	82.9
8	V	**19571**	498	460	5040	12	76.5	**17765**	4315	1542	1897	63	69.4
	P	1752	**5139**	591	5583	15	39.3	456	**16019**	448	453	24	92.1
	F	752	291	**2339**	2377	4	**40.6**	385	1038	**5360**	597	11	72.5
	N	191	36	28	**4311**	1	**94.9**	209	515	264	**3527**	5	78
	Si	137	55	23	360	**3371**	85.4	84	384	78	132	**3393**	83.3
16	V	**20852**	595	297	3788	122	**81.3**	**21796**	1876	1924	864	41	82.2
	P	2157	**5671**	379	4638	118	43.7	1547	**13659**	1264	498	24	80.4
	F	987	404	**1567**	2313	67	29.4	691	533	**6071**	294	6	79.9
	N	248	67	23	**4161**	4	92.4	596	527	540	**2627**	3	61.2
	Si	130	55	18	315	**3438**	86.9	187	246	142	66	**3387**	84.1
64	V	**18383**	1431	174	4232	170	75.4	**21767**	1279	2550	786	22	**82.4**
	P	1438	**8801**	167	3597	128	**62.3**	2086	**11043**	2400	686	22	68
	F	747	799	**990**	2321	87	20	541	283	**6701**	181	5	**86.9**
	N	181	109	10	**4107**	12	92.9	668	365	838	**2282**	3	54.9
	Si	95	115	12	263	**3463**	**87.7**	200	134	219	80	**3383**	84.2

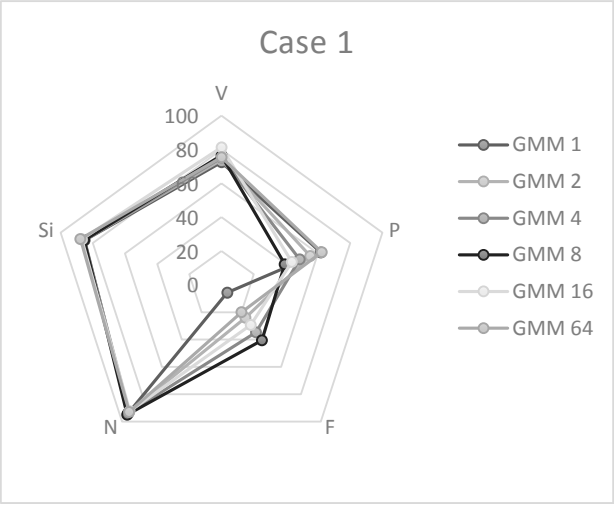


Figure 15: SR of classes in fixed states of HMM and using VQ

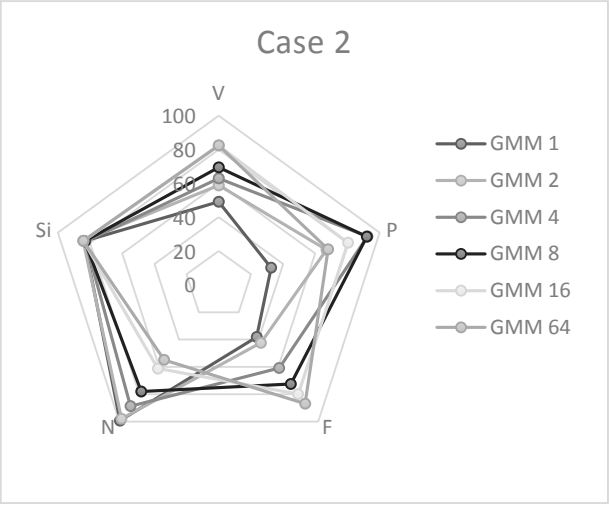


Figure 16: SR of classes in dynamic states of HMM and using VQ

TABLE V

CONFUSION MATRIX AND SUCCESSIVE RATE FOR EACH CLASSIFIER IN CASE 3 AND CASE 4

GMM		Confusion Matrix											
		Fixed states HMM and without using vector quantization						Dynamic states HMM and without using vector quantization					
		V	P	F	N	Si	SR	V	P	F	N	Si	SR
1	V	**13574**	3232	678	633	388	**73.4**	**8732**	663	2281	1991	3759	50.1
	P	388	**12428**	542	331	271	**89**	162	**4967**	1683	1397	2143	48
	F	207	1301	**1843**	232	111	49.9	55	141	**5284**	419	443	**83.3**
	N	119	678	176	**1236**	111	53.3	26	62	256	**3033**	345	81.5
	Si	23	164	16	24	**3496**	93.9	3	16	17	20	**4083**	**98.6**
2	V	**12457**	2526	1136	1662	363	68.7	**17173**	1203	318	2921	1712	73.6
	P	388	**11361**	822	836	236	83.3	1045	**7019**	206	2732	1287	57.1
	F	188	966	**2382**	556	75	57.2	538	535	**1223**	1512	767	26.7
	N	86	380	198	**2448**	58	77.2	127	143	30	**3677**	146	**89.2**
	Si	30	147	27	58	**3466**	93	46	34	1	96	**3883**	95.6
4	V	**10551**	2596	579	2538	1610	59	**17959**	2074	677	828	2522	**74.6**
	P	245	**11766**	462	1142	687	82.3	1093	**9488**	583	754	1842	**69**
	F	156	1032	**1271**	829	463	33.9	516	930	**2559**	391	1160	46.1
	N	52	303	128	**2851**	190	**80.9**	262	520	170	**1735**	673	51.6
	Si	10	103	7	52	**3755**	95.6	25	24	12	12	**4059**	98.2
8	V	**12256**	1128	803	2047	2757	64.5	**8005**	2028	2340	3347	3251	42.2
	P	335	**7660**	495	1439	1847	65	149	**8073**	1527	1968	1376	61.7
	F	141	414	**2404**	700	814	53.7	58	671	**4820**	553	581	72.1
	N	74	177	133	**2701**	335	79	24	425	188	**3119**	356	75.9
	Si	10	28	12	34	**3960**	97.9	7	22	30	52	**3989**	97.3
16	V	**10193**	822	902	1404	4395	57.5	**9082**	2431	847	4333	4303	43.3
	P	213	**6747**	550	890	2829	60.1	148	**9304**	486	2554	1962	64.4
	F	89	295	**2886**	401	1091	**60.6**	80	1044	**2424**	1269	1111	40.9
	N	28	148	127	**2309**	630	71.2	38	514	48	**3578**	348	79.1
	Si	8	11	5	18	**4080**	**99**	4	20	4	33	**4119**	98.5
64	V	**13551**	775	759	1597	2878	69.3	**10082**	2524	805	4420	3835	46.5
	P	389	**6171**	509	1189	2022	60	217	**7841**	500	2928	1936	58.4
	F	192	322	**2418**	493	822	56.9	112	1073	**2479**	1313	983	41.6
	N	80	135	99	**2530**	426	77.4	25	553	48	**3622**	299	79.7
	Si	19	21	6	23	**3956**	98.3	9	14	6	52	**4081**	98.1

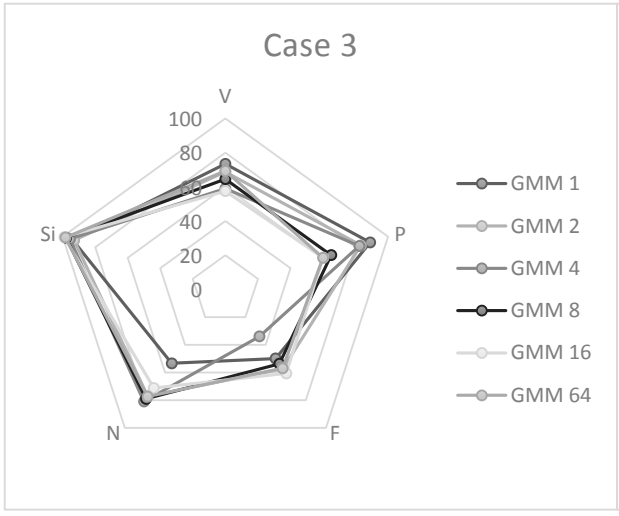


Figure 17: SR of classes in fixed states of HMM and without using VQ

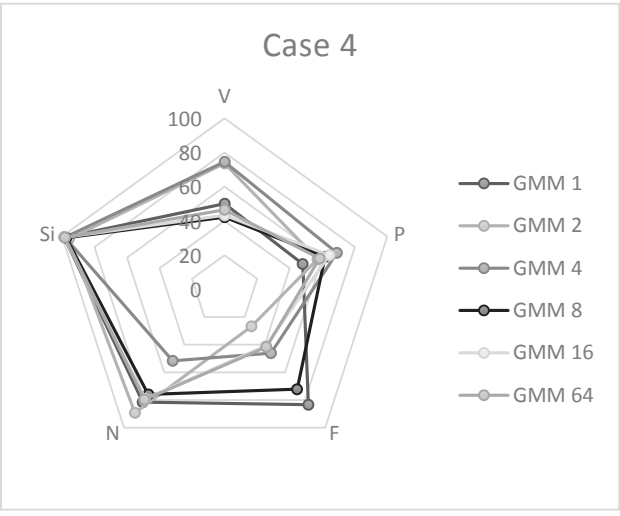


Figure 18: SR of classes in dynamic states of HMM and without using VQ

We conclude from the above tables that in the first model, using the high number of Gaussian mixtures (GM) gives the high prediction of all classifiers. In the second model, using high Gaussian mixtures give a high prediction for vowels and fricatives but low numbers of GM are prefect for other classes. In the third model, using a high number of GM is perfect for fricatives and silence but a low number of GM is good in others. In the fourth model, using low numbers of gaussian mixtures give a high prediction for all classifiers. The variance in the prediction of all classifiers between four techniques depends on the nature of features that are applied, the number of states fixed or dynamic that are applied and the number of Gaussian mixtures.

Recognition is highly sensitive to change in the transition matrix. For using the same database (TIMIT), better results have been achieved using the proposed model than those provided in [11]. Fricatives class success rate of 86.9% has been achieved compared to 78.37% in [11]. Plosives class success rate of 92.5% has been achieved compared to 68.75% in [11]. These results were achieved from (VS-HMM-GM-MBTI-CNN-VQ) model.

Also, similar work had been introduced by G. Deekshitha in [21]. The work-focused to provide a classification of speech into the same classes covered by this proposed model. The work entitled "Speech Signal Based Broad Phoneme Classification (SSBBPC) "makes use of Deep Neural Network (DNN) as an engine of the classifier for Broad Phoneme Classifier (BPC). TIMIT database had been utilized in system evaluation.

TABLE 5

RESULTS FOR DIFFERENT CLASSES FOR DIFFERENT MODELS ON THE SAME DATABASE

Models	V	P	F	N	Si
SSBBPC (REF)	91.754%	84.141%	83.825%	87.953%	89.988%
Case 1	81.3%	62.3%	40.6%	94.9%	87.7%
Case 2	82.4%	92.5%	86.9%	99.3%	84.3%
Case 3	73.4%	89%	60.6%	80.9%	99%
Case 4	74.6%	69%	83.3%	89.2%	98.6%

Table 5 indicates the results of our hybrid models and the reference model (SSBBPC). Reference model outperforms our models in vowels classification. But for plosives, fricatives and nasals; the variable states HMM with (MBTI-CNN-VQ) features (Case 2) has exceed the higher success rate. Silence achieved high success rate by using fixed states HMM with (MBTI-CNN) features (Case 3).

6 CONCLUSIONS

It has been noted that the automatic speech recognition success rate is improved using hybrid techniques of acoustic-phonetic approach and pattern recognition approach. The first acoustic hybrid model is called Fixed state, Hidden Markov Model, Gaussian Mixture (FS-HMM-GM). The second hybrid acoustic model is called Variable Structure, Hidden Markov Model, Gaussian Mixture (VS-HMM-GM). The second model gives a higher rate of correctness than the first model. Adapting both models for best overall success rate; by changing the Gaussian mixture counts to 64 mixtures are considered. There are two types of hybrid feature extraction used to enhance automatic speech recognition. The first hybrid features consist of Mel Best Tree image, Convolution Neural Network, Vector Quantization (MBTI-CNN-VQ). The second hybrid features consist of Mel Best Tree image, Convolution Neural Network (MBTI-CNN). The methodology of mixing (MBTI-CNN-VQ) gives a higher success rate of correctness than the second model (MBTI-CNN). The vector quantization technique also plays a good role in achieving good results. Results indicate that the improvement of the overall success rate is noticeable using (MBTI-VQ-CNN) features into the hybrid model (VS-HMM-GM). This indicates that the variable state structure can be utilized to increase the overall success rate due to the variation in period action of any speech class. Using vector quantization provided a high success rate because we make recognition in one feature vector for each frame instead of 1000 features vector. That will enhance training in these features to make the recognition. To be specified in terms of specific class classification performance, the highest success rates are achieved, using (FS-HMM-GM-MBTI-CNN) at (GM=16), as of almost 99% for silence class. Using (VS-HMM-GM-MBTI-CNN-VQ), the highest success rates are accomplished as of 82.4% at (GM=64) for vowel class, as of 92.5% at (GM=4) for plosive class, as of 86.9% at (GM=64) for fricative class and as of 99.3% at (GM=1) for nasals class. This implies that using the Variable Structure HMM engine and Vector Quantization is more efficient in the case of vowels, plosives, fricatives, and nasals but a Fixed Structure HMM engine without using Vector Quantization is more efficient in case of silence detection. The highest overall success rate (74.11%) is achieved using (VS-HMM-GM-MBTI-CNN-VQ). It is concluded that VQ is indicating more efficiency. In the future, some portions can be added to obtain better results. Examples of these portions are choosing various entropy functions and preparing it in MBTI, using smaller parts of syllables and choosing the best HMM to represent it. Also, increasing the number of GMM can improve the recognition rate. Using Recurrent Neural Network (RNN) instead of CNN.

REFERENCES

[1] J. Ye, R. J. Povinelli, and M. T. Johnson, "Phoneme classification using naive bayes classifier in reconstructed phase space," in *Proceedings of 2002 IEEE 10th Digital Signal Processing Workshop, 2002 and the 2nd Signal Processing Education Workshop.*, 2002: IEEE, pp. 37-40.

[2] T. J. Reynolds and C. A. Antoniou, "Experiments in speech recognition using a modular MLP architecture for acoustic modelling," *Information Sciences,* vol. 156, no. 1-2, pp. 39-54, 2003.

[3] P. Scanlon, D. P. Ellis, and R. B. Reilly, "Using broad phonetic group experts for improved speech recognition," *IEEE transactions on audio, speech, and language processing,* vol. 15, no. 3, pp. 803-812, 2007.

[4] H. Kekre, A. A. Athawale, and G. Sharma, "Speech recognition using vector quantization," in *Proceedings of the International Conference & Workshop on Emerging Trends in Technology*, 2011, pp. 400-403.

[5] G. Kiss, D. Sztahó, and K. Vicsi, "Language independent automatic speech segmentation into phoneme-like units on the base of acoustic distinctive features," in *2013 IEEE 4th international conference on cognitive infocommunications (CogInfoCom)*, 2013: IEEE, pp. 579-582.

[6] G. Deekshitha and L. Mary, "Broad phoneme classification using signal based features," *International Journal on Soft Computing,* vol. 5, no. 3, p. 1, 2014.

[7] A. Chittora and H. A. Patil, "Classification of phonemes using modulation spectrogram based features for Gujarati language," in *2014 International Conference on Asian Language Processing (IALP)*, 2014: IEEE, pp. 46-49.

[8] H. H. Nasereddin and A. A. R. Omari, "Classification techniques for automatic speech recognition (ASR) algorithms used with real time speech translation," in *2017 Computing Conference*, 2017: IEEE, pp. 200-207.

[9] S. Salim, G. Deekshitha, A. George, and L. Mary, "Automatic Spotting of Vowels, Nasals and Approximants from Speech Signals," in *2018 International CET Conference on Control, Communication, and Computing (IC4)*, 2018: IEEE, pp. 272-277.

[10] A. V. Anand, P. S. Devi, J. Stephen, and V. Bhadran, "Malayalam Speech Recognition system and its application for visually impaired people," in *2012 Annual IEEE India Conference (INDICON)*, 2012: IEEE, pp. 619-624.

[11] G. Deekshitha, K. H. Hathoon, and L. Mary, "A Novel Two-Stage System for Spotting Fricative and Plosive Regions from Continuous Speech," in *2018 International Conference on Communication and Signal Processing (ICCSP)*, 2018: IEEE, pp. 0760-0764.

[12] A. M. Gody, "Wavelet Packets Best Tree 4 Points Encoded (BTE)," *8th Conference on Language Engineering ,Ain-Shams University, Cairo, Egypt,* pp. 189-198, 2008.

[13] R. R. Coifman and M. V. Wickerhauser, "Entropy-based algorithms for best basis selection," *IEEE Transactions on information theory,* vol. 38, no. 2, pp. 713-718, 1992.

[14] A. M. Gody, M. Shabaan, and A. J. T. E. J. o. L. E. Saleh, "Automatic Speech Segmentation Using Hybrid Wavelet Features and HMM," *The Egyptian Journal of Language Engineering,* vol. 3, no. 2, pp. 1-13, 2016.

[15] K. O'Shea and R. Nash, "An introduction to convolutional neural networks," *arXiv preprint arXiv:1511.08458,* 2015.

[16] C. Ramaiah and V. S. Rao, "Speech samples recognition based on MFCC and vector Quantization," *International Journal on Computer Science and Emerging Trends (IJCSET),* vol. 1, no. 02, pp. 1-7, 2012.

[17] G. Xuan, W. Zhang, and P. Chai, "EM algorithms of Gaussian mixture model and hidden Markov model," in *Proceedings 2001 International Conference on Image Processing (Cat. No. 01CH37205)*, 2001, Thessaloniki, Greece, Greece, vol. 1: IEEE, pp. 145-148.

[18] J. C. Brown and P. Smaragdis, "Hidden Markov and Gaussian mixture models for automatic call classification," *The Journal of the Acoustical Society of America,* vol. 125, no. 6, pp. EL221-EL224, 2009.

[19] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," *NASA STI/Recon technical report n,* vol. 93, 1993.

[20] A. Gody, R. Abul Seoud, and M. Ezz El-Din, "Using Mel-Mapped Best Tree Encoding for Baseline-Context-Independent-Mono-Phone Automatic Speech Recognition," *The Egyptian Journal of Language Engineering,* vol. 2, no. 1, pp. 10-24, 2015.

[21] G. Deekshitha and L. Mary, "Speech Signal Based Broad Phoneme Classification and Search Space Reduction for Spoken Term Detection," in *TENCON 2018-2018 IEEE Region 10 Conference*, 2018: IEEE, Jeju, Korea (South), Korea (South), pp. 1601-1606.

BIOGRAPHY

Doaa A. Lehabik received the B.Sc. degree in Electrical Engineering – Communications and Electronics Department with very good with honor degree, from the Faculty of Engineering - Fayoum University in 2014. She joined the M.Sc. program in Fayoum University - Communications and Electronics Department in 2015. She received the Pre-Master degree from Fayoum University with very good with honor degree, in 2016. Her areas of interest include Speech recognition.

Mohamed H. Merzban received the B.Sc. from Fayoum university, M.Sc. from Cairo university, and PhD. from Egypt-Japan university for science and technology (E-JUST). He joined the teaching staff of the Electrical Engineering Department, Faculty of Engineering, Fayoum University, Egypt in 2006. He is now a lecturer at Fayoum university. His current research areas include computer vision and its applications to Robotics.

Sameh F. Saad received the B.Sc. degree in mechatronics engineering from High Institute of Engineering, Giza, Egypt, in 2000, the M.S. degree in mechatronics engineering from Mechatronics Laboratory, Paderborn University, NRW, Germany, in 2004 and the Ph.D. degree in electrical engineering at Cairo University, Giza, Egypt in 2012. From 2016 to 2018, he was a Researcher and a Lecturer with the October University for Modern Sciences and Arts, Giza, Egypt. His research interest includes the development of mobile robots and autonomous vehicle, automation of aquaponics ecological system, and automation of mechatronics systems.

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 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).

التصنيف الصوتي الواسع للتعرف التلقائي على الكلام بأستخدام المميزات المرئية

دعاء أحمد لحيبك [*1] , محمد حمدي مرزبان [*2], سامح فريد [**3], عمرو محمد جودي [*4]

[*] قسم الهندسة الكهربية, جامعة الفيوم , مصر

[1]da1174@fayoum.edu.eg

[2]mhm00@fayoum.edu.eg

[4]amg00@fayoum.edu.eg

[**] جامعة العلوم والآداب - جمهورية مصر العربية

[3]dr.sam.far@gmail.com

ملخص

إن هذه الدراسة تقدم أربع طرق مختلفة لعملية الكلام إلى مجموعات مختلفة حيث أن هذا التقسيم مفيد جدا في الحياه اليومية. تعتمد عملية التقسيم على الخصائص المستخرجة من الأصوات المختلفة. أول طريقة في هذا البحث تتكون من دمج مجموعة الخصائص التى تتكون من (تطبيق نظام الميل على أفضل شجرة من ال WPD مع شبكة التداخل العصبية مع VQ) مع استخدام نموذج "ماركوف الخفي" ذو الهيكلة الثابتة. ثاني طريقة تتكون من (تطبيق نظام الميل على أفضل شجرة من ال WPD مع شبكة التداخل العصبية مع VQ) مع استخدام نموذج ماركوف ذو الهيكلة المتغيرة حيث أن كل هيكلة لها عدد معين يختلف عن الإخر حسب الوقت التى تستغرقه كل مجموعة. ثالث طريقة هي التى تتكون من (تطبيق نظام الميل على أفضل شجرة من ال WPD مع شبكة التداخل العصبية) مع إستخدام نموذج ماركوف الخفي ذو الهيكلة الثابتة. رابع طريقة هي التى تتكون من (تطبيق نظام الميل على أفضل شجرة من ال WPD مع شبكة التداخل العصبية) مع إستخدام نموذج ماركوف الخفي ذو الهيكلة المتغيرة. وهنا تم تصنيف المقاطع الصوتية إلى 5 مجموعات وهي حروف متحركة (V) و حروف لا تحتوي على كلام (P) و حروف إحتكاكية (F) و حروف أنفية (N) وصامت الذي لا يحتوي على أي كلام (Si) وهذا على قاعدة بيانات من نوع TIMIT. كما أنه تم إستخدام عدد متغير من GMM الذي يتكون من (1 أو 2 أو 4 أو 8 أو 16 أو 64). وهذا لنحصل على أفضل نتائج لعملية التصنيف. ونحن نلاحظ أن الخصائص التى تحتوي على VQ تعطي أعلى نتائج حيث أن VQ يلعب دور مهم في تحسين نتائج التعرف على المجموعات. إستخدام عدد متغير من هيكلة نموذج ماركوف يلعب دورا مهما في تحسين النتائج حيث أن كل مجموعة لها وقت معين تختلف عن الأخرى. وأخيرا أحسن نموذج من الأربعة هو (تطبيق نظام الميل على أفضل شجرة من ال WPD مع شبكة التداخل العصبية مع VQ) مع إستخدام نموذج ماركوف ذو الهيكلة المتغيرة الذي يعطينا افضل نتائج التى تصل إلى 74.11% . وتم إستخدام ال HTK وذلك لشهرتها الواسعة في مجال ال ASR.

الكلمات الدالة

التعرف التلقائي على الكلام ، تقنية التصنيف ، تحليل حزمة الموجيات (WPD) ، شبكة التداخل العصبية، المتجهات الكمي (VQ) ، نموذج ماركوف الخفي.