Automatic Database Segmentation using Hybrid Spectrum - Visual Approach

Manar O. Gbaily *1, Amr M. Gody*2, Gamal A. El-Sheikh **3, Ahmed A. Nashaat*4

*Electrical Engineering Department, Faculty of Engineering, Fayoum University, Fayoum Egypt

1 noragbaily@yahoo.com

2 amg00@fayoum.edu.eg

4 aan01@fayoum.edu.eg

*** Electronics and Communication Department, PHI , 6th of October, Giza, Egypt

3 gaelsheikh@gmail.com

Abstract: ***Nowadays automated segmentation of speech signals has been attracted many of the researchers all-over the world. Many speech processing systems require segmentation of speech waveform into principal acoustic units. In this research, TIMIT DataBase (DB) is utilized to carry on this process and justify its operation or results. Thus, this paper presents a novel method of segmentation of speech phonemes, where the proposed strategy helps in the selection of appropriate feature extraction technique for speech segmentation. There are three main techniques of feature extraction used in our research; the first technique is the Mel Frequency Cepstral Coefficient (MFCC), the second technique is known by Best Tree Encoding (BTE), while the third is Image Normalized Encoder (INE), which is a hybrid technique between the Best Tree Image (BTI), and the Convolution Neural Network (CNN) ResNet-50. Then, data are trained using a hybrid model that consists of Hidden Markov Model (HMM), and Gaussian Mixture Model (GMM) to improve the performance of automatic speech recognition. The proposed model is tested and verified against the most widely used feature MFCC plus delta and delta-delta coefficients (39 parameters) to evaluate its performance. This approach has the potential to be used in applications such as automatic speech recognition and automatic language identification. The experimental results show that BTE technique achieved the highest success rate (η) (92.64%) than using the INE technique. However, the INE technique gives confusion success rate for Transition (Tr) and Non-Transition (NTr) of values 97.1% and 99.1%, respectively.***

Keywords: ***ASR, Segmentation techniques, HTK, WPD, BTE, MFCC, CNN, HMM, INE.***

1 INTRODUCTION

Speech is the primary and simple way used by the human for interaction/communication with others. It is a kind of pressure waves called acoustic waves which have features that can be used in recognizing a certain speaker or word. Speech recognition is a special case of speech processing, where it deals with the analysis of the linguistic contents of a speech signal. It is a method that uses an audio data entry to a computer or a digital system instead of a keyboard. Automatic Speech Recognition (ASR) is an important technology to enable and improve the human–human and human–computer interactions. The ASR is a task to automatically transcribe an unknown utterance from speech signal into text form, where the recognition of ASR can be enhanced (improving the success rate) by using specific database that is pertinent to a specific field. Speech segmentation is the process of determining the boundaries between words, syllables, or phonemes in uttered speech. The term speech segmentation is applicable to human mental processes as well as to natural language artificial processing. Speech segmentation is one of the branches of speed perception is representing a major sub problem of the speech recognition field. This makes it difficult to resolve the speech segmentation process properly in isolation from the speech recognition problem. Speech recognition can be split into two processes: feature extraction and pattern recognition. Feature extraction is taking charge of searching the speech characteristics and keeping them for the second process of pattern recognition. As well as the in most natural language processing problems, one must consider context, grammar, and semantics, and even so the result is often a probabilistic division (statistically based on likelihood) instead of a categorical one. 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 of sounds that is implemented for more applications like speech recognition and language recognition.

In this research broad phone classes are usually known as Transition (Tr), and Non-Transition or stable period (NTr) categories and can be used to improve speech recognition and hence categorization techniques were attempted. The research presents a new automatic segmentation technique which consists of three approaches for feature extraction; the first is the MFCC, where plus delta and acceleration coefficients (39 parameters) have been used. The second one is the BTE, while the third is INE which is a hybrid technique between the Best Tree Image (BTI), and the CNN ResNet-50. To

get higher success rate of recognition, data are trained using a hybrid model that consists of HMM and GMM to improve the performance of ASR.

This research provides a novel method of automatic segmentation used to increase the success rate of ASR. The subsequent sections explain the details of this research, started by a literature survey of relevant topics in section 2. Section 3 presents the proposed models and features for syllable classification. The DB and the experiments are presented in section 4. Results are presented and discussed in section 5. Finally, the conclusion is in section 6 complemented with some future work.

2 LITERATURE REVIEW

For the past six decades, the discipline of speech recognition has been one of the most productive study areas. HMMs [1], Feed Forward Neural Networks (FFNNs) [2], hybrid systems combining HMMs and FFNNs [3], and Deep Neural Networks (DNNs) [4], [5], [6] are the most frequent ways for enhancing ASR systems. According to the literature review, various studies are investigating into ways to improve the success rate of phone recognition classification.

In [7], the authors describe approaches for improving automatic phonetic segmentation accuracy using HMM acoustic-phonetic models. They found that applying more powerful statistical models for boundary correction that are conditioned on phonetic context and duration features enhanced test results. They discovered that combining multiple acoustic front-ends improved success rate and that conditioning the combiner on phonetic context and side information improved results, which reduced segmentation errors on the TIMIT corpus by nearly half, from 93.9 % to 96.8 % boundary success rate with a 20-ms tolerance.

Paper [8], proposes an algorithm for speech/music segregation in the presence of background noise. The proposed model represents the combination of a layer model separation method for noise removal and MFCC features for audio contextual information retrieval, which is supported by the Deep Belief Network (DBN) model for accurately segregated feature classification. A layered separation approach is applied using Recurrent Neural Network (RNN) and DNN techniques that retrieve contextual information. The separated layers are processed as MFCC features for segregation of the desired audio information. MFCC features resulted in speech segregation with a success rate of up to 91.60% by using the DBN classification model. Deep learning models decrease processing while increasing data size. After removing audio noise and performing speech segregation, applications could be modified to predict the occurrence of speech in the presence of audio noise.

In [9], the authors employed this method for phonetic segmentation and speech analysis at the phonetic level. The success rate percentage was approximately 78.14 %. The researchers discovered that the higher the number of states, the greater the alignment and precision obtained during modeling.

In [10], a phoneme classification approach using a modulation spectrogram as a feature extraction has been presented for classifying the phonemes of Gujarati language. Support Vector Machine (SVM) was used as a classification model for the phonemes. Six classes of phonemes have been established (vowels, semivowels, affricates, fricatives, stops, nasals). When the proposed features were combined with MFCC features, the best success rate was 95.70 %.

In [11], automatic segmentation of speech is about identifying boundaries of phonemes in a given utterance. The proposed technique is evaluated on Classical Arabic dataset. Extensive experiments are made to compare the proposed technique with state-of-the-art techniques, including the HMM-based forced alignment procedures. The results show that proposed technique has total error rate of 14.48%, while the success rate is 85.2% within 10 ms alignment error. When compared with the existing state-of-the-art technique, the proposed technique outperforms by 12.29% and 22.73% in terms of error rates and alignment success rates, respectively, which signifies the potential of using novel combination of Forward and Inverse Characteristics of Vocal tract (FICV) in speech segmentation.

In [12], the authors initiated research into the ten-class classification of speech signals. This study employed a set of ten Arabic numbers ranging from zero to nine. This study was implemented by MFCC. To keep the size of the feature vector constant, they used bi-directional Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU). Multi-Layer Perceptrons (MLP) were used to classify the data. There are 8800 sound waves in the DataBase, including 44 male and female recordings. In this study, 75% of the data was trained from a DataBase, while the remaining 25% was used in the testing phase. The recall, precision, and f-measure were used to evaluate this study (F1). The overall classification success rate achieved was 98.77 %.

In [13], the paper presents their work on building a speaker independent, Large Vocabulary Continuous Speech Recognition (LVCSR) system has been built for Sanskrit using HMM Toolkit (HTK). This is the maiden attempt on a

Sanskrit ASR, to our knowledge. A Graphical User Interface (GUI) for the speech recognizer has also been built using Java Swings. The developed system provides a phoneme level success rate of 62.3% and a word level success rate of 89.6% on the test set. 161 out of the 274 sentences in the test set were decoded correctly, yielding a sentence level success rate of 58.8%. A Sanskrit speech corpus with orthographic, word and phoneme level transcriptions has also been built. This corpus has 1360 sentences and 8370 words covering 46 phonemes. We plan to extend this work to develop a Very Large Vocabulary Continuous Speech Recognizer (VLVCSR) using deep learning framework.

In [14], the authors introduced a syllables classification method for ASR that includes the use of dynamic states of HMM. The MFCC and Mel Best Tree (MBT) feature techniques were applied. A subset of TIMIT databases is used in this research. The involved classes in this research are vowel, liquid, nasals, consonants, stops, and plosives. The overall success rate for MBT features was 81.01%, and 72.66 % for MFCC features.

In [15], the authors present a novel approach for classifying speech phonemes. Four hybrid approaches based on the acoustic-phonetic approach and the pattern recognition approach are used to identify the main concept of this study. The first hybrid model is Fixed State, structured HMM, Gaussian Mixture (GM), Mel-scaled Best Tree Image (MBTI), CNN, Vector Quantization(VQ) (FS-HMM-GM-MBTI-CNN-VQ), The second hybrid model is Variable State, dynamically structured HMM, GM, MBTI, CNN, VQ (VS-HMM-GM-MBTI-CNN-VQ), The third hybrid model is constructed (FS-HMM-GM-MBTI-CNN). The fourth hybrid model is constructed of (VS-HMM-GM-MBTI-CNN). The TIMIT DataBase was used in this research. All phones are classified into five classes Vowels, Plosives, Fricatives, Nasals, and Silences. The results showed that the highest overall success rate (74.11%) is achieved using (VS-HMM-GM-MBTI-CNN-VQ).

3 PROPOSED MODEL

This research proposes a novel method of automatic segmentation, where the task flow is illustrated in Figure 1. The first block at the left represents the corpus DB that has been used through this research, then the dotted rectangle is surrounding the different techniques of feature extraction that will be used in this work. The first technique is the MFCC which utilizes 13 coefficients that are extended to 39 coefficients by taking first (Δ) and second order derivatives (ΔΔ). The second technique is the BTE [16], while the third technique is Image Normalized Encoder (INE). The INE is hybrid technique that consists of Best Tree Image (BTI) and CNN ResNet-50 [15]. Then, the obtained features are applied to the recognition engine of hybrid model. The recognition engine consists of HMM and GMM with the objective to train TIMIT DB. The state's count of HMM is assumed to be fixed for each class. Speech signals are analyzed with two classes; Tr and NTr. With the goal to improve the ASR performance. Finally, the obtained results are evaluated with pertinent conclusions.

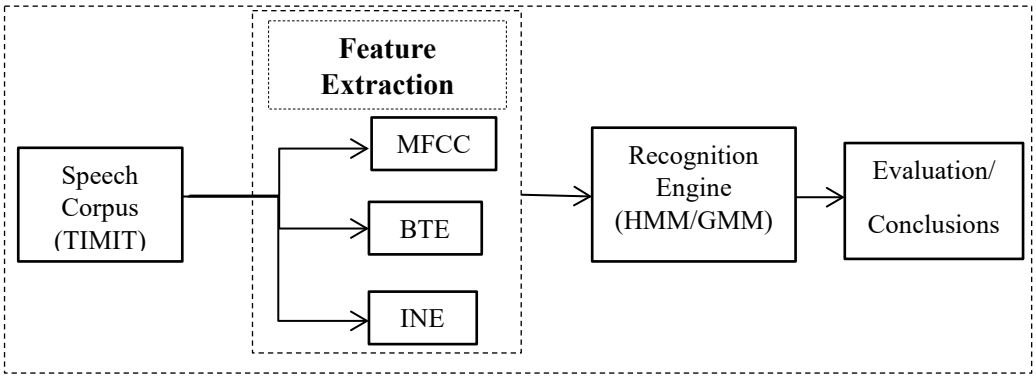

Figure 1: Task flow overview of this research

A. Speech Corpus (TIMIT)

The TIMIT speech corpus is designed to provide speech data for acoustic-phonetic studies and the development with evaluation of ASR systems. It contains a total of 6300 sentences, consisting of 10 sentences spoken by each of 630 speakers from 8 major dialect regions of the United States. All sentences were manually segmented at the phone level. It includes time-aligned orthographic, phonetic and word transcriptions as well as a 16-bit, 16 kHz speech waveform file for each utterance. The training and test of TIMIT files include 4620 and 1680 audio sentences, respectively. TIMIT DB used as an input speech signal and enters the feature extraction block, is considered as a reference due to its credibility and accuracy. The DB is prepared to modify transcription files for the character recognition objective of this research. This model includes Transition Engine {Tr, and NTr} as shown in Figure 2, where the TIMIT DB is segmented into phones and verified. The phones are replaced with another labels Tr, and NTr that will fit in this research, where Tr is for the transition part from one phone to another, while NTr is the phone itself. The Tr and NTr are processed to HTK format features file in the form of segments (labels) as shown in Figure 3 a, b. All data are pre-processed using different feature

extraction methods, and they are trained via the HMM\GMM to enhance the real time performance of ASR. Now, a method of automatic segmentation is provided to verify the results by comparing it to TIMIT segmentation.

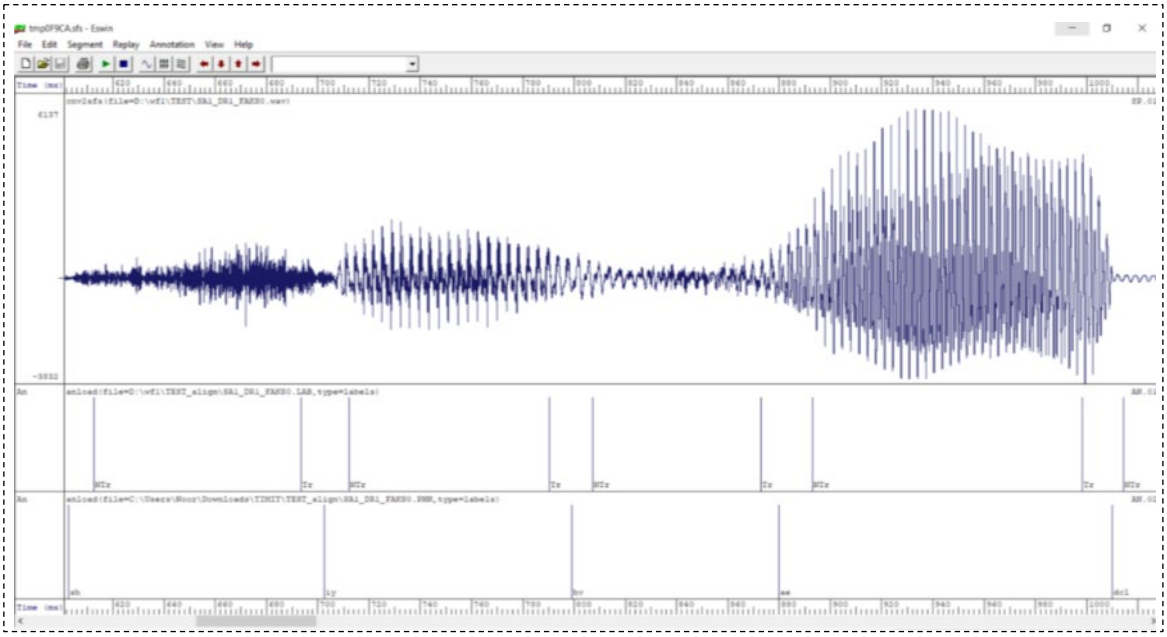

Figure 2: Tr, NTr labels in TIMIT

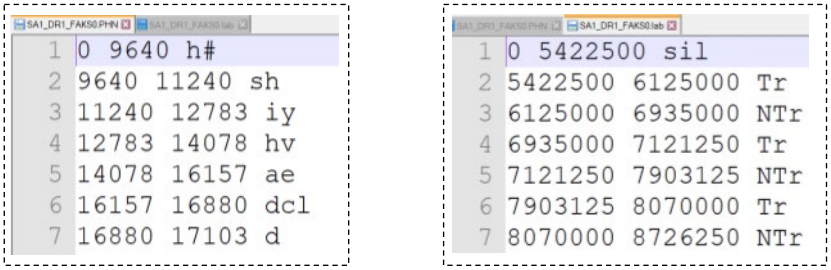

Figure 3: (a) .phn file from TIMIT before segmentation, should write references in order (b) .lab

B. Feature Extractions Methods

Feature extraction of speech signal is a crucial part since it allows the speech to be distinguished from others. There are different methods for feature extraction of speech signal such as statistical methods, spectral methods, model-based methods, transform based methods, pattern recognition methods, etc. [17]. Commonly, feature representations are extracted every 10ms over a window of 20 or 30ms for speech analysis. The aim is to find features that are stable for different examples of the same sounds in speech, despite differences in the speaker or the speaker's environment. The purpose of feature extraction is to transform input data into a set of properties for an utterance with acoustic correlation to the speech signal. That is, the intended properties which are termed as features can somehow be computed or estimated through the signal waveform processing. The features that had been suggested in the first model were derived from MFCC, while the second was derived from BTE, and the final were derived from Image INE which consists of a combination of two components: BTI and CNN ResNet-50. The output is a set of files that will be used in the recognition tasks for the next sections.

1) MFCC

There are a variety of feature representations in use, but the MFCC feature set is the most evident and popular feature extraction technique for speech recognition. MFCC approach can extract the efficient features of audio signals in time and frequency domains [18]. MFCC feature extraction procedure is broken down into several parts, which are detailed below and illustrated in Figure 4.

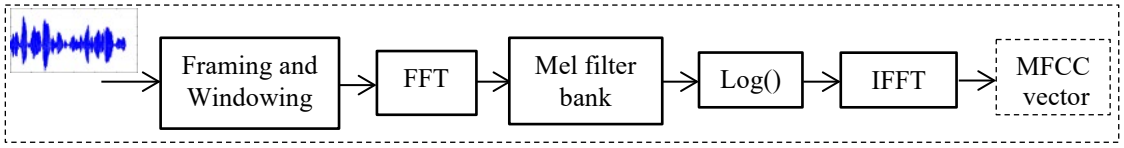

Figure 4: Block diagram of MFCC [18]

Speech signal is nonstationary; however, it is stationary at a certain time range (20-40 ms), which necessitates short window or frame. Consequently, the first process in MFCC is framing and windowing, In the framing process speech signal will be divided into several short frames (n) and then processed [19]. Thus, hamming windowing is responsible for creating a window shape by considering the next block of the feature extraction processing chain and integrating all the closest frequency lines. Hamming windows are computed based on equation (1) and equation (2) [20]:

$$Y[n] = X[n] * W[n] \quad \textbf{(1)}$$

$$W[n] = 0.54 - 0.46 * cos(\frac{2\pi n}{N-1}) \quad \textbf{(2)}$$

where,

- X: input signal.
- W: hamming window.
- Y: output signal.

FFT is a computational algorithm of Discrete Fourier Transform (DFT), where each frame is converted from N samples in the time domain into frequency domain [19]. This process preserves the convolution of glottal pulse and vocal tract impulse response h(t) in the time domain, where the computation is conducted based on equation (3) [20].

$$Y[w] = FFT\big(h(t) * x(t)\big) \quad \textbf{(3)}$$

The resulting spectrum of the DFT contains information in each frequency, where the spectrum is very wide and the voice signal does not follow the linear scale. Therefore, the Mel filter bank is used to ease the conversion and get a Mel frequency signal that is appropriate for human hearing and perception [19]. At this stage bank-filter analysis was used to perform linear predictions and the Mel-Scale computation is utilized by HTK according to equation (4) [20].

$$F_{(Mel)} = \left[2595 * log_{10}\left[\frac{1+f}{700}\right]\right] \quad \textbf{(4)}$$

where,

- $F_{(Mel)}$: frequency on Mel scale.
- f: frequency in Hertz.
- LOG ( ): is taking the logarithm of this provides Mel spectrum coefficients.

The first order of MFCC (Δ-MFCC) and the second-order derivatives of MFCC (ΔΔ-MFCC) are then added, and these are referred to as differential (13-coefficients) and acceleration, (13-coefficients) respectively. Equation (5) [21] is used to get the Δ-MFCC coefficients where (n=1), while its derivative yields the ΔΔ-MFCC coefficients, (n=2). Where, dt is the Δ coefficient at time t, from frame t computed in terms of the static coefficientsc_{t-n} , c_{t+n} and N is the window size of the delta.

$$d_t = \frac{\sum_{n=1}^{N} n(c_{t+n} - c_{t-n})}{2\sum_{n=1}^{N} n^2} \quad \textbf{(5)}$$

In this research all data had been pre-processed into frames using MFCC vectors plus delta and delta-delta coefficients (39 parameters) from a 20 ms window at a 20 ms frame rate. The obtained features of wave files in MFCC are saved in typical HTK format with file extension (.mfc), that had designated for storing the files. These files (ΔΔ-MFCC) are used to form the baseline results for comparisons and evaluations.

2) BTE

BTE was first introduced by Amr M. Gody in [16], and is a promising feature extraction technique used in ASR based on Wavelet Packet Decomposition (WPD). Consequently, the BTE can be considered as a new feature developed for ASR, the key of which is moving the ASR problem to new space where speech units can be effectively discriminated. Many phases of enhancement for BTE are done to enhance the efficiency. The procedure of extracting BTE features would be illustrated through the block diagram in Figure 5 [22].

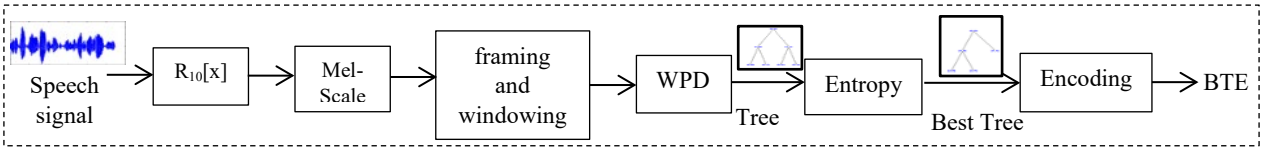

Figure 5: BTE block diagram

i. Resampling R10[x]:

The first step is to resample input signal to 10 kHz (5 kHz BW), to best distribute the wavelet tree structure through the significant bands as in Figure 6. Whereas human speech signal frequency band reach to 4 kHz bandwidth (BW) (8 kHz sampling by Nyquist rate). If the signal is resampled by 32 kHz (16 kHz BW), it will suffer from noise as in Figure 7 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. The bandwidth of human speech of interest will be contributed in the fourth quarter of the tree rectangle. The tree will be 75% of its area are not contributing to the information so that we are considering it as a noise. So, we resample the signal to 10 kHz (5 kHz BW).

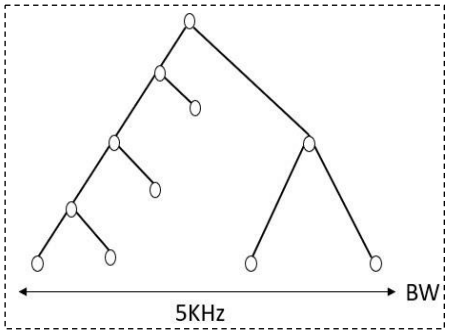

Figure 6: BTE frame after resampling at 10 KHz

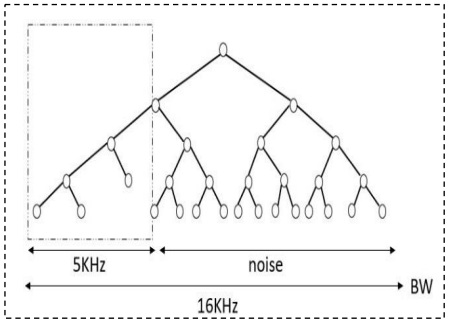

Figure 7: BTE frame resampled at 16 KHz

ii. Mel-scale:

Next step is to apply the resampled signal into the Mel-Scale for mapping the bandwidth of 5 kHz, where the formula for MS (f_{Mel}) is given as in equation(4). The weight of each node (in reference to Figure 6) is computed using this method based on its position on the MS curve in [23]. High weights will be allocated to nodes in the low frequency band, suggesting a high ability of human hearing, and vice versa.

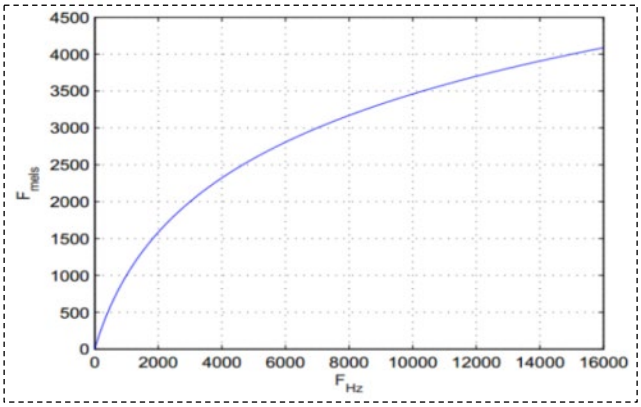

Figure 8: Mel-Scale curve [23]

iii. Framing and Windowing:

Being speech is non-stationary signal, it is converted into a series of small frames ranging from 20 to 40 ms. to ensure stationarity (signals whose frequency content does not change over time), the frame duration is set to 20ms (200 samples). Then, to make a seamless transition into continuous signal, a hamming window is utilized which is a

rectangular pulse with width equals to the frame length.

iv. Wavelet Packet Decomposition (WPD):

Wavelet is a short-duration waveform capable of expressing any function by scaling and shifting an original signal. After that, the signal is run through a high pass and low pass filters. Then, repeat the process for each part of the original signal, where the decomposition is considered the main term for this process. This process continues until the node in level 4, with a tree structure as the result (in reference to Figure 6). The following command was used to implement this in MATLAB;

$$\mathrm{t} = \mathrm{wpdec}(\mathrm{frame}, 4, 'db4', 'shannon')$$

Then, the frames of speech signal power are projected into defined filter banks using the entropy of the WPD coefficients.

v. Entropy:

Entropy is a tool for measuring the uncertainty of information content in given systems and it is widely used in signal processing, information theory, pattern recognition and other fields. The key to improving BTE is to increase entropy as it's used to measure the amount of information in each tree node. Thus, the best tree is decided by removing all low informative tree nodes according to Shannon entropy chosen in the original BTE.

vi. Best Tree:

The Best Tree selection model was described in detail [16], where it is started with the higher-level tree nodes and have one parent node for every two of them. If the parent node's entropy was greater than the total of both Childs' entropies, Childs would be deleted. This cycle would continue till the end with the best Tree. The point to be reminded, the key in BTE, is that each tree node represents a single frequency band while each node's component is the signal projection on the key frequency band. The information is expressed as a two dimensions image (2D) using best tree algorithm to keep the nodes of high informative contents.

vii. Encoding:

The last step is the encoding process, in which the Best Tree is obtained and encoded into 4 component features' vector. Each component represents quarter of the signal bandwidth and can be used to recall the best tree nodes that fall into the corresponding quarter represented by that component.
In this research, BTE features had been extracted and stored in typical User defined HTK format, where the file extension is designated as (.dat).

3) INE

In this research, we proposed the INE block as shown in Figure 9, which consists of hybrid features that combine the BTI and CNN ResNet-50. The BTI is applied as input into the CNN ResNet-50 block, where it is normalized to vectors for extracting the final features. This process is applied to each wave file and yields images which constitute the frames of that wave file and extracted from BTI block. Deep Residual Network (ResNet-50) is chosen to act as a type of CNN with 50 layers and this network can organize image into 1000 object classifications. It has a size for an input image of [224 224 3] (this size from the properties of ResNet-50 network) while the output feature vector is included in 1000 components. This network can be run on the Graphics Processing Unit (GPU) by using the tools in MATLAB. Finally, these features' vectors of wave file (.wav) are saved into HTK file format. The INE can be carried on using the following steps:

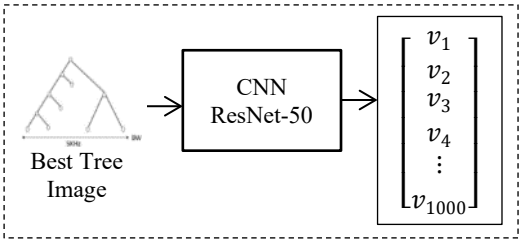

Figure 9: Image Normalized Encoder

i. Best Tree Image (BTI)

The output from BTI within the BTE block diagram (Figure 5) is the input to CNN ResNet-50, where all images are normalized to preserve both background color and image size. As a result, the tree image is drawn to fill the space of the image rectangle and ready for processing within the CNN.

ii. CNN

The availability of many speech feature extraction techniques makes the selection of proper feature extraction algorithm a challenging task. CNN is a form of Artificial Neural Networks (ANN) that is capable to detect patterns and make sense of them and this pattern detection makes the CNN useful for image analysis. CNN is a sequence of layers and every layer performs some unique functions on the data fed to it as illustrated in Figure 10. The five layers forming the CNN are considered input layer used to hold the raw data, while the convolution layer is used to perform dot product between image patch and all the filters in addition to the output-volume computation. CNN contains an activation function layer where the activation function is applied on every element of the output from the convolution layer. Next, the pooling layer makes the output of the previous layer memory efficient so that the computation costs are reduced. Finally, the fully-connected layer takes the input from the previous layer and yields the computed 1-D array class-scores. As the number of CNN layer increases, it is referred as Deep Neural Network (DNN) and improves the speech recognition success rate [24]. Among many of the CNN applications is the object detection and tracking algorithm where the features of image and video are extracted for security applications [25].

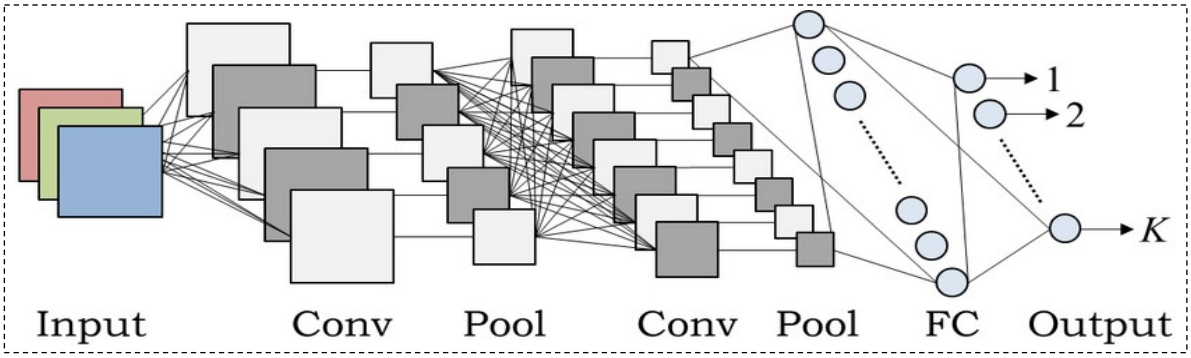

Figure 10: Architecture of CNN

C. Recognition Engine

ASR contains the most common generative learning approach which is based on GMM and HMMs. Conventional speech recognition systems utilize GMM based on HMMs to represent the sequential structure of speech signals. HMMs are used in speech recognition because speech signal can be viewed as piecewise stationary or short-time stationary signal. In short-time scale, speech can be approximated as a stationary process. Speech can be thought of as Markov Model for many stochastic purposes. Typically, each HMM state utilizes Gaussian mixture to model the spectral representation of sound waves. HMM-GMM is parameterized by $\lambda = (A, B, \pi)$, where π is the vector of state prior probabilities; $A=(a_{ij})$ is state transition probability matrix; $B=\{b_1,\ldots b_n\}$ is a set of b_j which represent the GMM of state j. The state is typically associated with a sub-segment of phone in speech [26], [27], [28], [29], [30].

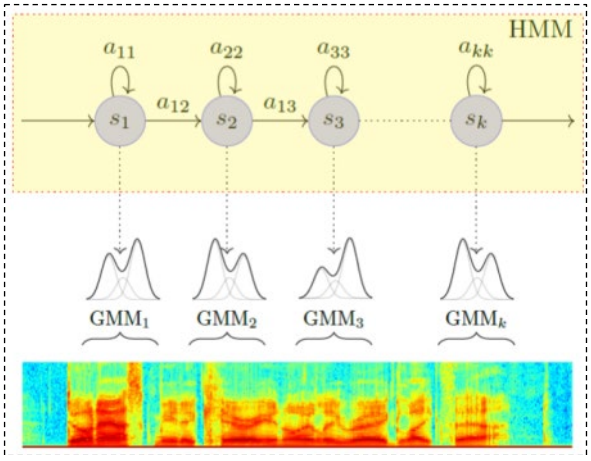

Figure 11: ASR Recognition System Using HMM\GMM

To achieve good levels of systems' performance, HMMs found its way to popularity within real applications as they handle the variable length data sequences resulting from variations in word sequence, speaker rate and accent. Even though the HMM-GMM approach had become the standard tool in ASR, it has its own advantages as well as disadvantages. HMMs-based speech recognition systems can be trained automatically with simple and computationally feasible usage. However, one of the main drawbacks of GMM is that they are statistically inefficient for modeling data

that lie on or near a non-linear manifold in the data space. The general structure of an HMM-GMM system for ASR is shown in Figure 11.

In this research the HTK toolkit outputs, obtained from different feature extractions, are used in building the HMM\GMM based acoustic model for ASR. Within the HMM model, all classifiers are trained using the same fixed state HMM as shown in Figure 12. This process contained three states, the first of which and the last are non-emitting states. The non-emitting states are needed to identify the entry and exit state in the HMM model. One emitting state is chosen because the best implementation for all phones is to have a short duration. Gaussian Mixtures with different counts are considered to construct the observation symbol probability function. To justify the performance of proposed algorithm, the system is tested against different Gaussian Mixture counts (2, 4, 8, 16 and 64).

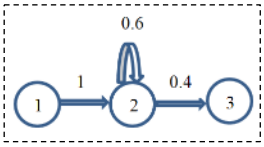

Figure 12: Static structure for all classes

D. Evaluation

The Evaluation is carried out using the success rate $\eta_{\lambda(m,n,p)}\%$, and the healthy factor $\sigma_{\lambda(m,n,p)}\%$, of the fixed state structure HMM design using the three techniques of feature extraction for each case as illustrated in TABLE 1, TABLE 2. The healthy factor (σ) is the relation between the success rate of the model (η) to the success rate of base model (MFCC_0_D_A) (η_0). In this work the reference model is considered $\lambda(1,64,2)$.

4 RESULTS AND DISCUSSIONS

A. Metrics and Measures

There are three techniques of features' extraction that are implemented in this research. The first is the MFCC (MFCC_0, MFCC_0_D, MFCC_0_D_A); where MFCC_0_D_A is the vector size which consists of 39 components. This number (39) is computed from the length of parameterized static vector (MFCC_0=13) plus the delta coefficient (D) (+13) plus the acceleration coefficient (A) (+13). The second is the BTE vector size that consists of 4 components, while the third is INE and this vector consists of 1000 components. Then all the models are trained using fixed HMM states and GMM. The comparative study is implemented to show the details and the key power in each specific feature set. All experiments are conducted; the results from which are verified against the most widely used feature technique (MFCC_0_D_A) to evaluate its performance. First, the model λ is related to the experiment variables as of equation (6), while the success rate is defined by equation (7).

$$\lambda(m, n, p) \tag{6}$$

where,

λ: is the model

$m: \text{features type} = \{\text{MFCC}, \text{BTE}, \text{INE}\}$

$n: \text{GMM count} = \{\text{without}, 2,4,8,16,64\}$

$p: \text{feature dervivatives} = \left\{\text{non: } (0), \left(\frac{\partial}{\partial x}\right) \text{Delta D: } (1), \text{Delta and } \left(\frac{\partial^2}{\partial x}\right) \text{Accelration A: } (2)\right\}$

$$\eta_{\lambda(m,n,p)} = SR = \frac{N - D - S}{N} \times 100\% \tag{7}$$

where, $\eta_{\lambda(m,n,p)}$ is the total success rate for each model, N indicates the total number of recognized classes in the expected transcription, D indicates deletions (class not found in the output), and S indicates substitutions (class replaced by other one). The confusion success rate for each class is defined by equation (8)

$$\Gamma_\lambda(u, v) = \frac{T_{uv}}{T_u}; (\text{confusion success rate for given model } \lambda) \tag{8}$$

where,

u: reference class {Tr, NTr}

v: test class {Tr, NTr}

T_{uv}: count of class u recognized as class v

T_u: count of class u

The healthy factor (σ_{uv}) is measured as in equation (9)

$$\sigma_{uv} = \frac{\Gamma_{\lambda(u,v)}}{\Gamma_{o(u,v)}}\% \quad (9)$$

where, $\Gamma_o(u,v)$ is the confusion success rate for the given reference model. The healthy factor $\sigma_{\lambda(m,n,p)}$ is calculated from the relation between total success rate for any model $\eta_{\lambda(m,n,p)}$ and the reference model $\eta_{\lambda(MFCC,64,2)}$.

$$\sigma_{\lambda(m,n,p)} = \frac{\eta_{\lambda(m,n,p)}}{\eta_{\lambda(MFCC,64,2)}}\% \quad (10)$$

TABLE 1 represents the obtained values of the total success rate $\eta_{\lambda(m,n,p)}$%, and the healthy factor $\sigma_{\lambda(m,n,p)}$ % for the fixed state structure HMM with different counts of Gaussian Mixture (2,4,8,16 and 64). The obtained results from $\lambda(MFCC, 64,2)$ yields the maximum success rate and consequently it is selected as reference result for each class to evaluate its performance. Results in TABLE 1 were calculated from equation (7), and it show that using $\lambda(MFCC, 64,0)$ achieved η 26.48% and σ is 32.05%, while using $\lambda(MFCC, 64,1)$ yields η of 76.38%, and σ is 92.46%. on the other hand using $\lambda(BTE, 64,0)$ achieved η of 92.94% and σ is 112.51%, and finally using $\lambda(INE, 64,0)$ achieved η of 89.22% and σ is 108.01%. The results clarify that using $\lambda(BTE, 64,0)$ yields the highest success rate (η = 92.94% , σ =112.51%), where all results are compared to that obtained with the reference $\lambda(MFCC, 64,2)$ that achieved (η = 82.6%, σ = 100%) as shown in Figure 13.

TABLE 1

SUCCESS RATE AND HEALTH FACTOR FOR DIFFERENT FEATURE EXTRACTIONS, VARIOUS GM

Feature Type (m)	GMM Count (n)	Feature derivatives(p)	$\eta_{\lambda(m,n,p)}$%	$\sigma_{\lambda(m,n,p)}$ %
MFCC_0	Without GM	0	25.24	30.55
	2		25.5	30.87
	4		25.68	31.08
	8		25.99	31.46
	16		25.29	30.61
	64		26.48	32.05
MFCC_0_D	Without GM	1	23.98	29.03
	2		24.66	29.85
	4		27.35	33.11
	8		30.24	36.61
	16		73.26	88.69
	64		76.38	92.46
MFCC_0_D_A	Without GM	2	36.92	44.69
	2		39.8	48.18
	4		41.74	50.53
	8		44.06	53.34
	16		45.94	55.61
	64		82.6	100
BTE	Without GM	0	83.08	100.58
	2		35.41	42.86
	4		53.27	64.49
	8		78.04	94.47
	16		91.21	110.42
	64		92.94	112.51
INE	Without GM	0	88.54	107.19
	2		88.54	107.19
	4		88.54	107.19
	8		88.54	107.19
	16		89.2	107.99
	64		89.22	108.01

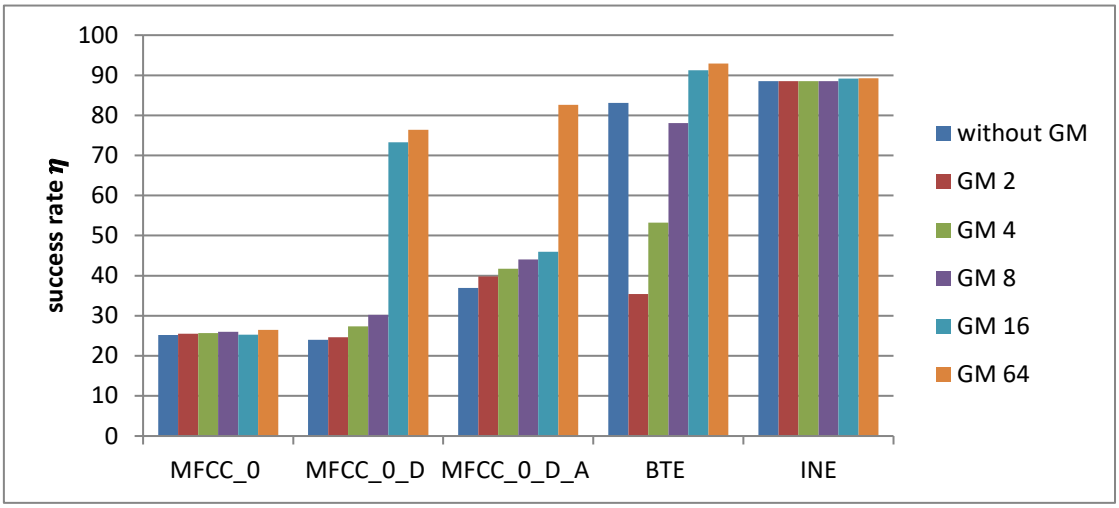

Figure 13: SR% for different feature extractions, various GM

In reference to Figure 14, this research concludes that using the high number of GM with $\lambda(\text{MFCC}, 64,0)$, $\lambda(\text{MFCC}, 64,1)$, $\lambda(\text{MFCC}, 64,2)$, and $\lambda(\text{INE}, 64,0)$ gives high prediction for the classes confusion success rate. However, using high GM in BTE gives high prediction for $\Gamma_\lambda(\text{Tr}, \text{Tr})$ but low numbers of GM are better for $\Gamma_\lambda(\text{NTr}, \text{NTr})$. The best confusion success rate for both classes Tr and NTr are obtained from model $\lambda(\text{INE}, 64,0)$ which has got $\Gamma_\lambda(\text{Tr}, \text{Tr})$ is 97.1%, and the $\Gamma_\lambda(\text{NTr}, \text{NTr})$ is 99.1%; compared with the result of the reference model $\lambda(\text{MFCC}, 64,2)$ $\{\Gamma_\lambda(\text{Tr}, \text{Tr}) = 95.6\%$, and $\Gamma_\lambda(\text{NTr}, \text{NTr}) = 96.2\%\}$. The proposed model gives better results and clarifies that the BTE and INE feature extractions with automatic segmentation outperformed MFCC.

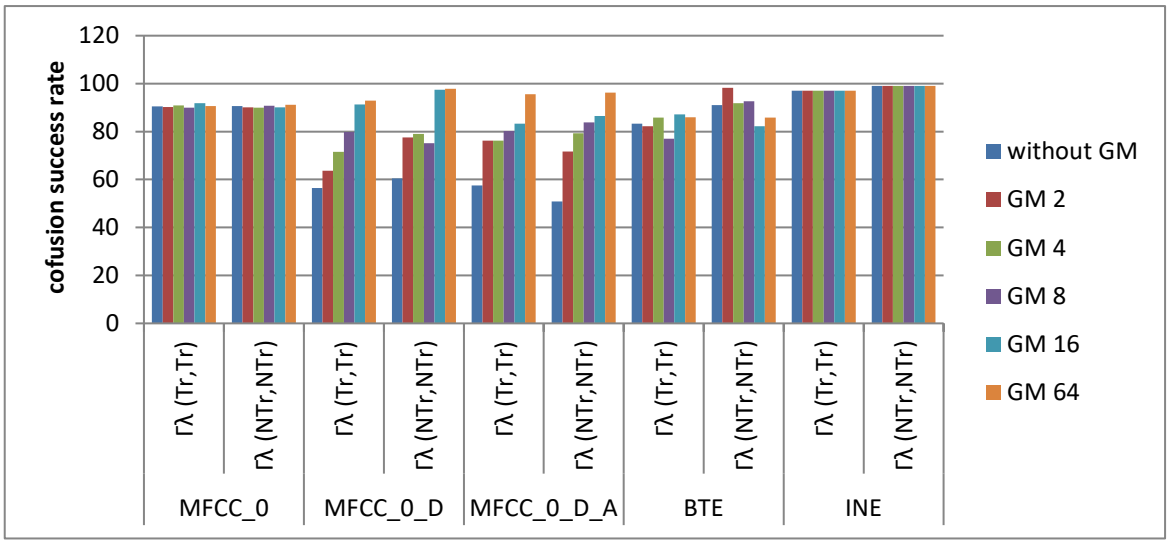

Figure 14: $\Gamma_\lambda(u, v)$ for each class for different feature extractions, various GM

TABLE 2 illustrates the relationship $\Gamma_o(u, v) = \frac{\Gamma_\lambda(u,v)}{\sigma_{uv}}$ for each class, when using different extracted features. For the first model $\lambda(\text{MFCC}, 64,0)$ the result is $\Gamma_\lambda(\text{Tr}, \text{Tr}) = 91.8\%$, and $\Gamma_\lambda(\text{NTr}, \text{NTr}) = 91.2\%$, while in the second model $\lambda(\text{MFCC}, 64,1)$ it was found that $\Gamma_\lambda(\text{Tr}, \text{Tr}) = 92.9\%$, and $\Gamma_\lambda(\text{NTr}, \text{NTr}) = 97.8\%$. For the third model $\lambda(\text{BTE}, 64,0)$ the result is $\Gamma_\lambda(\text{Tr}, \text{Tr}) = 86\%$ and in $\lambda(\text{BTE}, 2,0)$ the result is $\Gamma_\lambda(\text{NTr}, \text{NTr}) = 85.8\%$. The last model $\lambda(\text{INE}, 64,0)$ yields the results $\Gamma_\lambda(\text{Tr}, \text{Tr}) = 97.1\%$, and $\Gamma_\lambda(\text{NTr}, \text{NTr}) = 99.1\%$ which are the best results for both classes Tr and NTr; compared with the result of the reference model $\lambda(\text{MFCC}, 64,2)$ $\{\Gamma_\lambda(\text{Tr}, \text{Tr}) = 95.6\ \%$, and $\Gamma_\lambda(\text{NTr}, \text{NTr}) = 96.2\%\}$.

Furthermore, the proposed approach ($\lambda(INE, 64,0)$) achieved the best success rate (89.22%) compared to literature [15] in which it equals 74.11% using the same DB (TIMIT) and the same feature extraction. In addition, the proposed approach ($\lambda(BTE, 64,0)$) achieved the highest success rate (92.64%) compared to similar work in literature [14] which gave 81.01% using same DB and same feature extraction.

TABLE 2
$\frac{\Gamma_{\lambda}(u,v)}{\sigma_{uv}}$ FOR EACH CLASS FOR DIFFERENT FEATURE EXTRACTIONS, VARIOUS GM

Feature Type (m)	**GMM Count (n)**	**Feature Derivatives (p)**	$\Gamma_{\lambda}(\mathbf{Tr},\mathbf{Tr})$	$\sigma_{\mathbf{Tr},\mathbf{Tr}}$	$\Gamma_{\lambda}(\mathbf{NTr},\mathbf{NTr})$	$\sigma_{\mathbf{NTr},\mathbf{NTr}}$
MFCC_0	Without GM	0	90.5	0.95	90.7	0.93
	2		90.2	0.94	90.1	0.92
	4		90.9	0.95	90	0.92
	8		90	0.94	90.8	0.93
	16		91.8	0.96	90.1	0.92
	64		90.6	0.95	91.2	0.93
MFCC_0_D	Without GM	1	54.4	0.59	60.5	0.62
	2		63.6	0.66	77.5	0.79
	4		71.6	0.75	79	0.81
	8		79.8	0.84	75.1	0.77
	16		91.3	0.96	97.4	0.99
	64		92.9	0.97	97.8	1.0
MFCC_0_D_A	Without GM	2	57.5	0.60	50.9	0.52
	2		76.2	0.80	71.7	0.73
	4		76.2	0.80	79.3	0.81
	8		80.2	0.84	83.8	0.85
	16		82.3	0.87	86.5	0.88
	64		95.6	1.0	96.2	0.98
BTE	Without GM	0	83.3	0.87	91	0.93
	2		82.2	0.86	98.3	1.0
	4		85.8	0.90	91.8	0.94
	8		77	0.81	92.6	0.94
	16		87.2	0.91	82.2	0.84
	64		86	0.90	85.8	0.88
INE	Without GM	0	97.1	1.02	99.1	1.01
	2		97.1	1.02	99.1	1.01
	4		97.1	1.02	99.1	1.01
	8		97.1	1.02	99.1	1.01
	16		97.1	1.02	99.1	1.01
	64		97.1	1.02	99.1	1.01

5 CONCLUSIONS

This paper presents a novel automatic segmentation method with different types of feature extraction to enhance total success rate of ASR; the first is MFCC [MFCC_0, MFCC_0_D, MFCC_0_D_A], while the second is BTE and the last one is INE. The proposed model is trained by using hybrid model of HMM and GMM. The proposed model experiments are conducted with GM counts of 64 mixtures, from which the results showed that the highest total success rate (92.64%) is achieved by BTE compared to other features' extraction (MFCC, INE) which gave 82.6%, 89.22%, respectively.

Confusion success rate for class Tr is $\Gamma_{\lambda}(\mathrm{Tr},\mathrm{Tr})$, and for NTr is $\Gamma_{\lambda}(\mathrm{NTr},\mathrm{NTr})$ gave the best result in INE with 97.1% and 99.1% respectively. Furthermore, the proposed approaches ($\lambda(\mathrm{INE},64{,}0)$) and ($\lambda(\mathrm{BTE},64{,}0)$) achieved the best total success rate (89.22%, 92.64%), compared to literature [15], [14] which achieved (74.11%, 81.01%) respectively using the same DB (TIMIT) and the same feature extraction. In the future, some portions can be added to obtain better results. Examples of these portions are to choose various entropy functions and to prepare it in BTE, INE, using smaller parts of syllables and choosing the best HMM to represent it. In addition, increasing the number of GM to improve the recognition success rate necessitate more justification. Finally Recurrent Neural Network (RNN) can be in model training and evaluated instead of HMM.

REFRENCES

[1] K. F. Lee, H. W. Hon, "Speaker-independent phone recognition using hidden Markov models," *IEEE Transactions on Acoustics, Speech, and Signal Processing,* vol. 37, no. 11, pp. 1641-1648, Nov. 1989.

[2] F. Fallside, H. Lucke, T. P. Marsland, P. J. O'Shea, M. S. J. Owen, R. W. Prager,… N. H. Russell, "Continuous speech recognition for the TIMIT database using neural networks," in *International Conference on Acoustics, Speech, and Signal Processing*, Albuquerque, NM, USA, 1990.

[3] H. A. Bourlard and N. Morgan, Connectionist speech recognition: a hybrid approach, Kluwer Academic, 2012.

[4] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Vancouver, BC, Canada, 2013.

[5] G. Hinton, L. Deng, D. Yu,G. Dahl, A. Mohamed, N. Jaitly, B. Kingsbury, "Deep Neural Networks for Acoustic Modeling in Speech Recognition: The shared views of four research groups," *IEEE Signal Processing Magazine,* vol. 29, no. 6, pp. 82-97, 2012.

[6] A. Mohamed, G. E. Dahl, & G. Hinton, "Acoustic Modeling Using Deep Belief Networks," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 20, no. 1, pp. 14-22, 2011.

[7] A. Stolcke, N. Ryant, V. Mitra, J. Yuan, W. Wang, & M. Liberman, "Highly accurate phonetic segmentation using boundary correction models and system fusion," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, 2014.

[8] K. A. Qazi, T. Nawaz, Z. Mehmoud, M. Rashid, & H. A. Habib, "A hybrid technique for speech segregation and classification using a sophisticated deep neural network," *PLOS ONE,* vol. 13, no. 3, 2018.

[9] P. Bansal, A. Pradhan, A. Goyal, A. Sharma, and M. Arora, "Speech Synthesis-Automatic Segmentation," *International Journal of Computer Applications,* vol. 98, no. 4, pp. 29-31, july 2014.

[10] A. Chittora and H. A. Patil, "Classification of phonemes using modulation spectrogram based features for Gujarati language," in *International Conference on Asian Language Processing (IALP)*, Kuching, Malaysia, 2014.

[11] M. Javed, M.M.A. Baig & S.A Qazi, "Unsupervised Phonetic Segmentation of Classical Arabic Speech Using Forward and Inverse Characteristics of the Vocal Tract," *Arabian Journal for Science and Engineering,* vol. 45, no. 3, p. 1581–1597, July 2012.

[12] N. Zerari, S. Abdelhamid, H. Bouzgou and C. Raymond, "Bi-directional recurrent end-to-end neural network classifier for spoken Arab digit recognition," in *2nd International Conference on Natural Language and Speech Processing (ICNLSP)*, Algiers, Algeria, 2018.

[13] C. S. Anoop and A. G. Ramakrishnan, "Automatic Speech Recognition for Sanskrit," in *2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT)*, Kannur, India, 2019.

[14] D. N. Senousy, A. M. Gody, and S. F. Saad,, "Syllables Classification for ASR using Variable State Hidden Markov Model," in *18th Conference on Language Engineering*, Ain Shams University, pp. 1-11, 2018.

[15] D.A. Lehabik, M. H. Merzban, S. F. Saad, A. M. Gody, "Broad Phonetic Classification of ASR using Visual Based Features," *The Egyptian Journal of Language Engineering,* vol. 7, no. 1, pp. 14-26, 2020.

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

[17] B. Jolad and R. Khanai, "An Art of Speech Recognition : A Review," in *2nd International Conference on Signal Processing and Communication (ICSPC)*, Coimbatore, INDIA, 2019.

[18] M.R Hasan, M. Jamil, S. Rahman, "Speaker identification using Mel Frequency Cepstral Coefficients," in *3rd*

International Conference on Electrical and Computer Engineering, Dhaka, Bangladesh, 2004.

[19] E. S. Wahyuni,, "Arabic speech recognition using MFCC feature extraction and ANN classification," in *2nd International conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, Yogyakarta, Indonesia, 2017.

[20] P. Kurzekar, R. Deshmukh, V. Waghmare, and P. Shrishrimal, "A Comparative Study of Feature Extraction Techniques for Speech Recognition System," *International Journal of Innovative Research in Science, Engineering and Technology,* vol. 3, no. 12, pp. 18006-180016, 2014.

[21] Akpudo, U. E., & Hur, J. W., "A Cost-Efficient MFCC-Based Fault Detection and Isolation Technology for Electromagnetic Pumps," *Electronics,* vol. 10, no. 4, pp. 1-20, 2021.

[22] A. M. Gody, R. A. AbulSeoud, M. M. Ibraheem, "Hybrid Model Design for Baseline-Context-Independent-Mono-Phone Automatic Speech Recognition," *International Journal of Engineering Trends and Technology (IJETT),* vol. 27, no. 6, pp. 304-313, September 2015.

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

[24] J. Rownicka, S. Renals and P. Bell, "Simplifying very deep convolutional neural network architectures for robust speech recognition," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Okinawa, Japan, 2017.

[25] A. Raghunandan, Mohana, P. Raghav and H. V. R. Aradhya, "Object Detection Algorithms for Video Surveillance Applications," in *International Conference on Communication and Signal Processing (ICCSP)*, Chennai, India, 2018.

[26] L. Rabiner and J. Bing-Hwang, Fundamentals of Speech Recognition, Prentice-Hall, 1993, pp. Rabiner, L. and Bing-Hwang, J. 1993. Fundamentals of Speech Recognition. Prentice-Hall.

[27] M. A. Anusuya and S. K. Katti, "Speech Recognition by Machine:A Review," *International Journal of Computer Science and Information Security,* vol. 6, no. 3, pp. 181-205, 2009.

[28] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing,* vol. 26, no. 1, pp. 43-49, February 1978.

[29] J. Baker, "The DRAGON system--An overview," *IEEE Transactions on Acoustics, Speech, and Signal Processing,* vol. 23, no. 1, pp. 24-29, 1975.

[30] J. Bilmes, "HMMs can do," *IEICE Trans. Inf. Syst,* Vols. E89-D, no. 3, p. 869–891, 2006.

BIOGRAPHY

M. O. Gbaily received a B.Sc. degree in Communications and Electronics Department with very good and honor degree from Pyramids Higher Institute for Engineering and Technology in 2014.
She joined the teaching staff of the Communications and Electronics Department, PHI, Egypt, in 2015. She joined the M.Sc. program at Fayoum University - Communications and Electronics Department in 2015. Her areas of interest include automatic speech recognition.

A. M. Gody received the B.Sc. M.Sc., and PhD. from the Faculty of Engineering, Cairo University, Egypt, in 1991, 1995, and 1999. He joined the teaching staff of the Electrical Engineering Department, Faculty of Engineering, Fayoum University, Egypt, in 1994. He is the Acting Chief of the 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).

G. A. El-Sheikh received B.Sc. and M.Sc. in EE (guidance, control and navigation) from MTC, Cairo, in 1980, 1990. Instructor, MTC, Cairo, Egypt, 1985-1987. PhD degree in EE (robust self-tuning control with aerospace applications) from the Industrial Control Centre, Strathclyde University, UK, 1994. Lecturer and Chief for the GC, MTC, Cairo, 1994-2000. Ass Prof. in GC, MTC, Cairo, 2000-2004. Full time professor in EC, MSA University, 2004-2006. Full time professor, SVA, 2006-2008. Visiting Professor in Karary University, Republic of Sudan, 2008-2009. Head of EC, PHI, 6th of October, Giza, Egypt. Cooperative postgraduate-research supervision with MTC, Ain Shams University, Cairo University aerospace Engineering, Alexandria University, Helwan University, Banha University. Supervisor of (38) PhD and MSc/PhD Thesis and (3) under-supervision; Examiner: Internal (38) and External (14 MSc + 4 PhD); and author of more than (79) papers published in local and international journals and conferences. Research interests include control theories and design (Classical, Modern, Robust –LQG-GLQG-H_∞- GH_∞, Adaptive, Self-Tuning); Systems identification; CC; Sys Sim and Data Acquisition; Autopilot design and analysis; Embedded systems, automation of industrial applications, PLC, microcontrollers, Arduino, VHDL; Inertial Sensors: conventional, laser, fiber-optic, solid state and MEMS, MEMS actuators, GPS; Guidance, Navigation and Control, Autopilot design, embedded Flight Control.

Ahmed A. Nashat received his B.Sc. in communication and electronics engineering from Cairo University – Egypt in 1982. In 1985, he received his M.Sc. degree in communication and electronics engineering from King Fahd University of Petroleum and Minerals – Saudi Arabia. In 1990, he received his Ph.D. in communication Engineering from New Mexico State University – USA. From 1990 till 1995, he worked for Telstra Research Lab. – Australia. Currently, he is working as an associate professor at Fayoum University – Egypt. His research interest includes application of digital signal processing techniques to spectral analysis, smart antennas, signal detection and estimation theory, radar and sonar systems, pattern recognition analysis, image processing, digital filters, and adaptive noise cancellation.

التقسيم الآلي لقاعدة البيانات باستخدام الطيف الهجين والنهج البصري

أحمد علي نشأت [*4] , جمال أحمد الشيخ[**3] , عمرو محمد جودي [*2] , منار عثمان جبيلي[*1]

[*] قسم الهندسة الكهربائية ، كلية الهندسة، جامعة الفيوم ،مصر

1 noragbaily@yahoo.com

2 amg00@fayoum.edu.eg

4 aan01@fayoum.edu.eg

**قسم الالكترونيات والاتصالات،معهد الاهرامات للهندسة، السادس من اكتوبر، الجيزة، مصر

3 gaelsheikh@gmail.com

الملخص:

في الوقت الحاضر، اجتذب التقسيم الآلي لإشارات الكلام العديد من الباحثين في جميع أنحاء العالم ، تتطلب العديد من أنظمة معالجة الكلام تجزئة شكل الموجة الكلامية إلى وحدات صوتية رئيسية. في هذا البحث ، يتم استخدام TIMIT DataBase (DB) لمواصلة هذه العملية وتبرير عملها أو نتائجها. وبالتالي، تقدم هذه الورقة طريقة جديدة لتجزئة الصوتيات الكلامية ، حيث تساعد الإستراتيجية المقترحة في اختيار تقنية استخراج الميزة المناسبة لتجزئة الكلام. هناك ثلاث تقنيات رئيسية لاستخراج الميزات المستخدمة في بحثنا ؛ التقنية الأولى هي معامل Cepstral للتردد الميل (MFCC)، والتقنية الثانية معروفة من خلال أفضل ترميز شجرة (BTE)، بينما التقنية الثالثة هي تشفير الصورة الطبيعية (INE)، وهي تقنية هجينة بين أفضل صورة شجرة (BTI)، والشبكة العصبية الالتفافية CNN(ResNet-50). بعد ذلك ، يتم تدريب البيانات باستخدام نموذج هجين يتكون من Hidden Markov Model (HMM)، و (Gaussian Mixture Model (GMM لتحسين أداء التعرف التلقائي على الكلام. تم اختبار النموذج المقترح والتحقق منه مقابل الميزة الأكثر استخدامًا MFCC بالإضافة إلى معاملات دلتا ودلتا دلتا (39 معلمة) لتقييم أدائها. يمكن استخدام هذا النهج في تطبيقات مثل التعرف التلقائي على الكلام والتعرف التلقائي على اللغة. أظهرت النتائج التجريبية أن تقنية BTE حققت أعلى نسبة نجاح (92.64) (η٪) من استخدام تقنية INE. ومع ذلك ، فإن تقنية INE تعطي معدل نجاح ارتباك للانتقال (Tr) و Non-Transition (NTr) للقيم 97.1٪ و 99.1٪ ، على التوالي.

الكلمات الدالة:

ASR، تقنيات التجزئة ، HTK، WPD، BTE، MFCC، CNN، INE،HMM.