Optimal Entropy to Enhance the Structure of the Wavelet-Packets-Best-Tree for Automatic Speech Recognition

Fatma M. Abd El_latif [*1], Amr M. Gody [**2], Waleed A. Maguid Ahmed [*3]

[*]*Engineering Math & Physics Department, Faculty of Engineering, Fayoum University, Fayoum, EGYPT.*

[1]fma06@fayoum.edu.eg

[3]waa01@fayoum.edu.eg

[**]*Electrical Engineering Department, Faculty of Engineering, Fayoum University, Fayoum, EGYPT.*

[2]amg00@fayoum.edu.eg

Abstract: ***Best Tree Encoding (BTE) is a promising feature extraction technique based on wavelet packet decomposition that is utilized in Automatic Speech Recognition (ASR). This research introduces an enhancement of Wavelet Packet Best Tree (WPBT) Calculations. The standard features BTE encodes the tree structure using a mathematical model into a features vector of 4 components. The best tree structure has been calculated using the entropy function. In the standard version of BTE, Shannon entropy has been chosen as the entropy function. In this research, Shannon Entropy (SE), Renyi Entropy (RE), and Tsallis Entropy (TE) are used to construct the Best Tree. The encoding of the Best Tree has been done using the same mathematical model approach in the standard 4-Point BTE. The proposed model is tested and Verified against the most widely used feature Mel Frequency Cepstral Coefficient (MFCC) plus delta and delta-delta coefficients (39 parameters) to evaluate its performance. The TIMIT database is used in this research. All phones are divided into five classes: Vowels, Fricatives, Silences, Nasals, and Plosives. The acoustical model has been implemented using Hidden Markov Model (HMM). No language model has been applied. The HMM Tool Kit (HTK) software is used for model implementation. The experiments show that BTE using Tsallis entropy yields the highest overall success rate of 75.85% which is better than MFCC's overall success rate of 71.76%. Comparing the vector of 4 components of BTE to the 39 components vector of MFCC makes it a very promising feature vector to be considered for research and development.***

Keywords: ***ASR, BTE, WPD, HMM, Shannon entropy, Renyi Entropy, Tsallis entropy, MFCC.***

1 INTRODUCTION

Speech recognition is the method of automatically extracting and evaluating linguistic information transmitted by a speech signal using computers or electronic circuits. Automatic speech recognition techniques, which have been studied for several years, have mostly been aimed at realizing transcription and human-computer interaction systems. The speech signal is usually represented in terms of phones, and words are simply seen as concatenations of phone sequences. Phone classification is the process of assigning speech categories to a small segment of speech signals. This research provides an enhancement for a new feature extraction approach; Best Tree Encoding (BTE) [1], which is based on wavelet packet decomposition and best tree to increase the success rate of ASR.

The vital element that affects the success rate of the Best Tree Encoding (BTE) is the type of entropy. Shannon entropy (SE), Renyi entropy (RE), and Tsallis entropy (TE) are used in this research to extract features from the wavelet packet tree (WPT) to obtain the best tree. Fixed and variable state structure Hidden Markov Model (HMM) is used to train the data. BTE is a vector of 4 components only compared to a vector of 39 components of Mel Frequency Cepstral Coefficient (MFCC) plus delta and delta-delta coefficients.

The paper is organized as follows: Section 2 includes a literature review on relevant topics. Section 3 provides an overview of BTE. Section 4 represents an Experiment environment, which includes a database and HMM model's design. Section 5 represents and discusses the results. Finally, Section 6 contains the conclusion.

2 LITERATURE REVIEW

The definition of different types of speech classes, feature extraction methods, speech classifiers, and performance evaluation issues to consider when developing a speech recognition system. Feature extraction is the most essential aspect of speech recognition since it separates one speech from another. The utterance can be derived from a wide variety of feature extraction techniques proposed and successfully used for speech recognition tasks [2]. There are various techniques for feature extraction: Linear prediction coding (LPC) [3], Mel frequency Cepstral Coefficient (MFCC) [4], Linear prediction cepstral coefficient (LPCC) [5], Wavelet Packet Features (WPF) [6]. Best tree encoding (BTE) is a new

feature extraction technique first introduced by Amr M. Gody in [1]. From the literature review, there are some feature extraction studies to improve the phone classification recognition rate.

Anil Kumar et al. in [7] presented a study on detecting the points of the beginning of vowels in encrypted speech. In this study's experiments, TIMIT, and broadcast news corpus were used. The data set consisted of 95 consonant-vowel (CV) classes. Several methods for extracting features were used, including excitation source (EXC), spectral peaks (SP), modulation spectrum (MOD), and mixed methods (COMB). A comparative study was given among these various feature methods using the missing rate and average deviation. As compared to other models, the presented approach had a higher success rate and a lower average deviation. The data was trained using a hybrid of the Hidden Markov Model (HMM) and Support Vector Machine (SVM). The proposed method achieved the highest success rate 66.14% in clean speech.

J. Ye et. al. in [8] introduced a novel approach for classifying speech phonemes based on histograms of reconstructed phase spaces This method is a new methodology that is significantly different from conventional techniques. Preliminary results show that the approach is a promising way to create a phoneme recognizer. The proposed method classifies phonemes to vowel, fricative, and nasal in the TIMIT database. The results indicated that a reconstructed phase space approach is a specific classification method, with overall recognition rates of 61.59%, 34.49%, and 30.21% for fricative, vowel, and nasal phonemes, respectively.

Nasereddin et al. in [9] presented a study on the classification of speech signals into four classes. MFCC was used as a feature extraction technique. Hidden Markov Model (HMM), Dynamic Time Warping (DTW), and Dynamic Bayesian Network (DBN) were used in the classification stage. The results showed that DBN outperformed in recognizing one class while HMM achieved the highest success rate for the others.

Keshet et al. in [10] proposed a method for Precise plosive detection based on pattern matching. The problem of false silence detection has been solved using the hierarchical treatment and multi-class decisions. The presented method has been tested using the TIMIT corpus, which yielded a very high detection success rate. The results showed that the presented method was very effective at detecting plosives. The distribution of insertions for the different classes is 23% for vowels, 34% for silences, 6.5% for nasals, 28% for fricatives, 1.5% for affricates, and 7% for glides.

G. Tryfou et al. in [11] introduced research into the classification of speech signals into five classes These classes are vowels, stops, nasals, fricatives, and liquids. In this study, a subset of the TIMIT database was used. They translate 61 phonemes into 48 phonemes. This research used two feature extraction techniques, MFCC and time-frequency reassigned cepstral coefficients (TFRCC). HTK, and GMM were used to train the data. TFRCC had the highest success rate in classifying stops with 53.74 %, while MFCC had the highest success rates in the other classes. TFRCC increased the overall success rates when transitioning to vowels, fricatives, and liquids by 36.5%, 26.6%, and 70.84%, respectively.

G. Deekshitha et al. in [12] proposed a novel method for detecting fricative and plosive regions in continuous speech. A two-stage recognition system is designed for detecting and verifying the fricative and plosive regions. In the first stage, a Deep Neural Network (DNN) based broad classifier is used to convert the input speech signal to corresponding broad phoneme classes. Silence, Vowel, Nasal, Fricative, and Plosive are the broad classes. Thus, fricative and plosive regions are prioritized. In the second step, the fricative regions are verified using a spectral centroid, and the plosive regions are verified using a difference in spectral distribution. Automatic detection of fricative and plosive regions can be used to improve speech recognition performance. Using the TIMIT database, the verification rate for Fricative is 78.37% and for Plosive it is 68.75%.

T. Jeff Reynolds et al. in [13] introduced research into classifying speech signals into seven classes. Fricatives, semi-vowels, diphthongs, plosives, nasals, closures, and vowels are the classes. A set of 39 TIMIT phones was used. Four feature extraction techniques were gathered to perform this work: MFCC, perceptual linear prediction (PLP), LPC, and posterior. In this study, HMM and Multi-Layer Perceptron (MLP) were used. Gathering MFCC, PLP, and LPC resulted in the highest levels of recognition. The phone classification rate achieved was 84.1%.

P. Scanlon et al. in [14] presented a novel approach using a neural network multilayer perceptron classifier with a modular order of experts. Phonemes are classified into seven classes vowel, semi-vowels, diphthongs, stops, fricatives, nasals, and silence. PLP was used as feature extraction. The experiment has been running on a TIMIT database. The highest success rate was 74.2%.

A. Rizal et al. in [15] presented a new feature extraction of lung sound, which is multilevel wavelet packet entropy (MWPE) calculations using Shannon entropy, Renyi entropy, and Tsallis entropy. The test was performed on five classes of lung sound databases. In this research MLP was used as the classifier. The results showed that MWPE using Shannon calculation could yield the highest success rate of 97.98% for N = 4 decomposition level. On the other hand, MWPE

using Renyi entropy yielded the highest success rate of 93.94% and the one using Tsallis entropy yielded 57.58% success rate.

Doaa N. Senousy et al. in [16] 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.

Doaa A. Lehabik et al. in [17] presented 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. They are (FS-HMM-GM-MBTI-CNN-VQ), (VS-HMM-GM-MBTI-CNN-VQ), (FS-HMM-GM-MBTI-CNN), and (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 BEST TREE ENCODING

Best Tree Encoding (BTE) was first introduced by Amr M. Gody in [1]. The procedure of extracting BTE will be illustrated through the block diagram in Figure 1. The process of creating BTE starts with converting the input speech signal into short time duration frames. The preprocessing phase is the second step. In the preprocessing phase, wavelet packet decomposition (WPD) is used. The proper entropy type is applied in the next step to obtain what is called the best tree. The last step is to encode this best tree into a feature vector of four components to obtain the BTE feature.

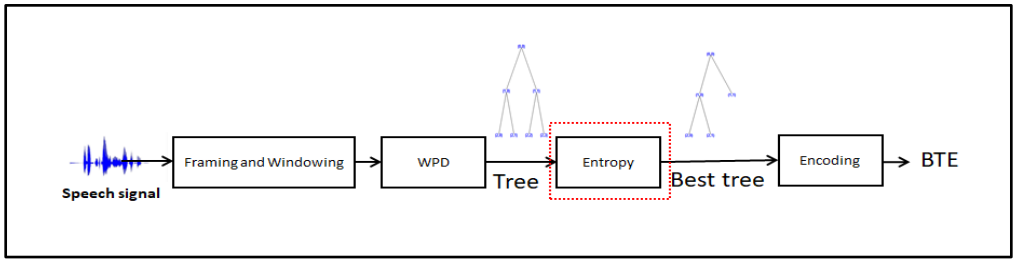


Figure 1: BTE block diagram

A. Framing and Windowing

The input speech signal is divided into small duration frames to deal with it as a stationary signal. The frame length is 20 ms. Then the Hamming window that is a rectangular pulse whose width is equal to the frame length is applied to make a smooth transition to the signal to be continuous.

B. Wavelet Packet Decomposition

The wavelet packet approach is a generalization of wavelet decomposition that provides a wider variety of signal analysis options and allows for the most accurate signal analysis [18]. Wavelet packet elements are waveforms that are indexed by three naturally interpreted parameters: position, scale, and frequency. The wavelet transform is defined below as the inner product of a signal $x(t)$ with the mother wavelet $\psi(t)$ [19]:

$$\psi_{a,b}(t) = \psi\left(\frac{t-b}{a}\right) \tag{1}$$

$$W_{\psi}x(a,b) = \frac{1}{\sqrt{a}}\int_{-\infty}^{+\infty} x(t)\ \psi^{*}\left(\frac{t-b}{a}\right)dt \tag{2}$$

where ψ^{*} is a conjugate of mother wavelet ψ, a and b denote the scale and shift parameters. By modulating a and b, the mother wavelet can be dilated or translated. The wavelet packets transform recursively decomposes the speech signal generated by the recursive binary tree as shown in Figure 2. The wavelet packet transform (WPT) is like the Discrete wavelet transform (DWT), but the WPT decomposes all details and approximations rather than only approximations. The wavelet packet (WP) principle states that for a given signal, a pair of low-pass and high-pass filters are used to generate two sequences that capture different frequency sub-band features of the original signal [19].

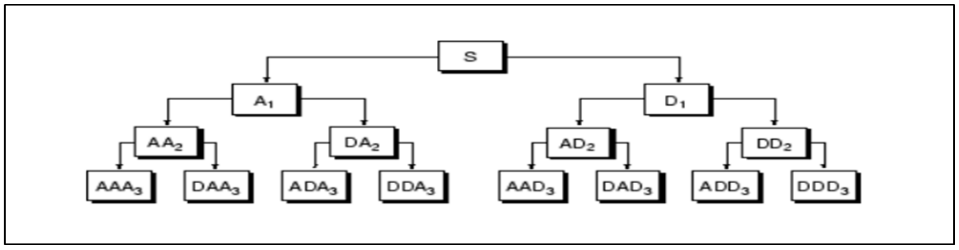


Figure 2: 3 Levels Wavelet packets decomposition [1]

C. Wavelet Packet Entropy

Each node in the wavelet packet tree gives the contribution of the signal power into a certain frequency band. Not all nodes have useful information about the speech signal so, the proper entropy function is applied to obtain the optimal nodes that contribute the most to signal information. Beginning with the higher-level tree nodes, every 2 nodes have one parent node. If the parent node's entropy is greater than the sum of the entropies of both Childs, Childs will be removed. This process will be repeated until the end. The obtained nodes collection is called the best tree.

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. Some common types of entropy are Shannon entropy (SE), Renyi entropy (RE), and Tsallis entropy (TE). Entropy can be calculated using energy [20].

The information of the kth coefficient of the jth node at ith level can be calculated by wavelet energy which is defined as Equation (3) [20]:

$$E_{i,j,k} = \|d_{i,j,k}\|^2 \tag{3}$$

Then, the total energy for the jth node at ith level can be calculated by Equation (4) [20]:

$$E_{i,j} = \sum_{k=1}^{N} E_{i,j,k} \tag{4}$$

where N is the number of the corresponding coefficients in the node. The probability of the kth coefficient at its corresponding node can be calculated by Equation (5) [20]:

$$p_{i,j,k} = E_{i,j,k}/E_{i,j} \tag{5}$$

where the sum of $p_{i,j,k}$ equals 1.

SE is a measure of uncertainty associated with random variables in information theory, and it can be calculated by Equation (6) [20]:

$$SE_{i,j} = -\sum_{k=1}^{N} p_{i,j,k} * log(p_{i,j,k}) \tag{6}$$

RE entropy can be defined as Equation (7) [20]:

$$RE_{i,j,q} = \frac{1}{1-q}\ log\ (\sum_{i=1}^{N} {p^q}_{i,j,k}) \tag{7}$$

where q is the order of entropy (q $\geq$ 0 and q $\neq$ 1).

TE is another type of entropy that is defined at various q values as Equation (8) [20]:

$$TE_{i,j,q} = \frac{1}{q-1}\ (1 - \sum_{i=1}^{N} {p^q}_{i,j,k}) \tag{8}$$

Both RE and TE are extensions of SE. The parameter q in RE and TE needs to be optimized in practical applications. RE and TE tend to SE for q $\rightarrow$ 1.

BTE uses Shannon entropy (SE) to obtain the best tree. In this research, Renyi Entropy (RE) and Tsallis entropy (TE) are used to obtain the best tree.

D. Encoding

The final step is the encoding process. The best tree obtained in step 3 is encoded into four component feature vector. Figure 4 and Figure 3 show the four points encoding algorithm for BTE-4. For detailed discussion the reader may refer to paper [1]. In summary the nodes are rearranged in order to minimize the distance between the adjacent in frequency features vectors. As a quick example, MatLab wavelet packet indexing system in Figure 4 shows that node 2 and node 3 are subsequent but they are not in the frequency band (V1 at low band while V2 at High band). Figure 3 shows the proposed coding. Note that node 2 and 3 fall into the same band as well as they are consecutive numbers. Each component represents a quarter of the signal's bandwidth. Each component can be used to recall the best tree leaf nodes in the relevant quarter that fall into the corresponding quarter represented by the component.

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

Figure 4: Matlab function "besttree" indexing scheme [1]

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

Figure 3: Wavelet packet tree of 4 points encoding [1]

Figure 5 shows an example of Tree Structure Encoding into 4 components. Each component is 7 bits. Each bit maps to a tree node. Circles point to leaf nodes in BTE that contain information (Those leaves of high Entropy than the Children). To obtain feature vectors, it is represented as a decimal number as listed in Table 1.

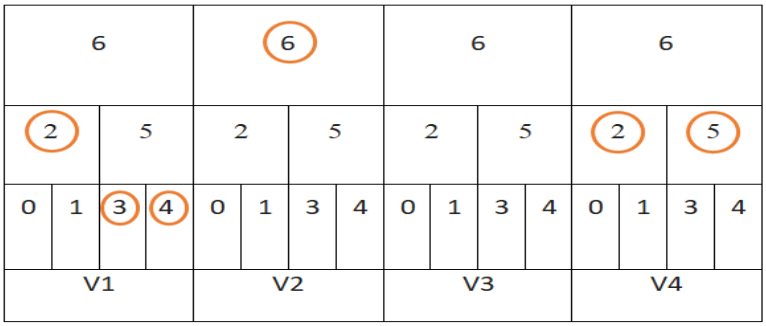


Figure 5: Best tree 4-point encoding example

TABLE 1

BEST TREE 4 POINT ENCODING EVALUATION [1]

Element	Binary Value	Decimal Value	Frequency Band
V1	0011100	28	0 - 25 %
V2	1000000	64	25% - 50%
V3	0000000	0	50%-75%
V4	0100100	36	75%- 100%

The features vector for this example will be,

$$F = \begin{pmatrix} 28 \\ 64 \\ 0 \\ 36 \end{pmatrix}$$

Mel scaled (MS) Best Tree Encoding that is used in this research is an enhanced version of BTE that is introduced in [21]. This version of BTE is focused on the algorithm for evaluating the best tree. The Mel scale is used to evaluate the best tree nodes. In addition to Mel-Scale; the input speech signal is resampled at 10 kHz to map the bandwidth of 5(kHz).

The formula for MS (f_{Mel}) is given as follows:

$$f_{Mel} = 2595 * \log_{10}(1 + \frac{f_{node}}{700}) \tag{9}$$

In this approach, each node's weight is calculated depending on its position on the MS curve of Figure 6 [21]. Nodes in the low-frequency band will be assigned high weights, indicating a high ability of human hearing and vice versa.

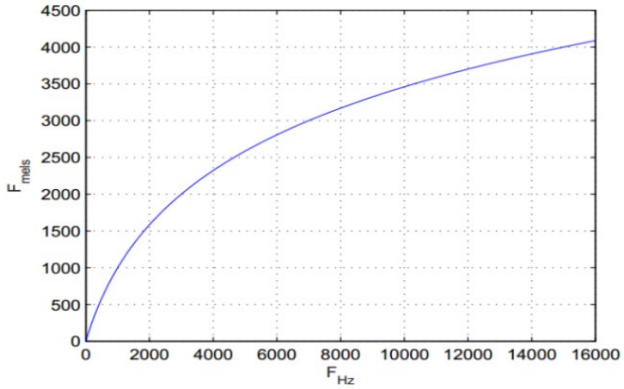


Figure 6: Mel scale curve [21]

4 EXPERIMENT ENVIRONMENT

A. *Database*

The TIMIT database is used in this research. The TIMIT corpus was designed to supply speech knowledge for the acquisition of acoustic-phonetic information and for developing automatic speech recognition systems [22]. This database contains 6300 sentences, ten sentences spoken by each of 630 speakers from eight major dialect regions of the United States (US). The original version of the TIMIT database includes 61 phones. The database is processed to modify transcription files for the character recognition to be suitable for the objective of this research. Vowels (V), Fricatives (F), Nasals (N), Plosives (P), and Silences (Si) as in [12]. Each classifier with phones assigned to it is listed in Table 2.

TABLE 2

PHONE CLASSIFIERS [12]

Classifiers	Number of labels	TIMIT Labels
Vowels (V)	25	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)	16	p, t, k, b, d, g, jh, ch, bcl, dcl, gcl, pcl, tcl, kcl, q, dx
Fricatives (F)	10	s, sh, z, zh, f, th, v, dh, hh, hv
Nasals (N)	7	m, em, n, nx, ng, eng, en
Silences (Si)	3	h#, epi, pau

B. HMM Models Design

HMMs have the advantage of being able to model variable-length sequences, whereas other models usually need a fixed feature set. There are two different models of HMM. The first model is the fixed state structure HMM model shown in Figure 7; in which all classifiers are trained using the same HMM fixed number of states which has three emitting states and two non-emitting states. The Non-emitting states are needed in the HMM model to define the entry and exit states. The second model is the variable state structure HMM model; in which Vowel, Fricatives, Nasals, and Silences are modeled by three emitting states as shown in Figure 8. Plosives are modeled using two emitting states due to their short time duration as shown in Figure 9.

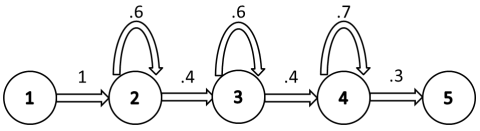


Figure 7: Fixed states HMM for all classifiers in the first model

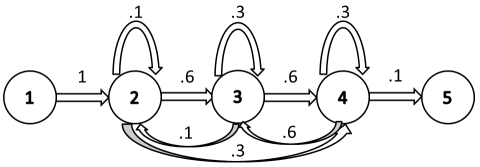


Figure 8: Vowels, Fricatives, Nasals, and Silences design in the second model

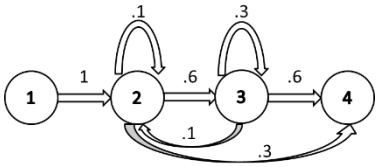


Figure 9: Plosives design in the second model

5 RESULTS AND DISCUSSIONS

Two sets of feature extraction techniques are used. First, the Mel frequency cepstral coefficient (MFCC) technique (MFCC_0_D_A) is used which has 39 coefficients [4]. The zeroth coefficient represents the average log-energy of the input signal it is often ignored because it contains little speaker-specific information, D represents delta coefficients is the first-order derivative; it informs us about the speech rate, and A represents the delta-delta coefficients is the second-order derivative; it provides information like speech acceleration. Second, Best Tree Encoding (BTE). The standard features BTE encodes the tree structure using a mathematical model into a features vector of 4 components. In the standard version of BTE, Shannon entropy has been chosen as the entropy function. In this research, Shannon entropy (SE), Renyi entropy (RE), and Tsallis entropy (TE) are used to construct the Best Tree. The proposed model is tested and Verified against the most widely used feature technique (MFCC_0_D_A) to evaluate its performance.

The success rate can be defined by Equation (10). D represents Deletions (class not found in the output). S represents Substitution (class replaced by other phones). N represents the total number of classes in the expected transcription.

$$SR = \frac{N - D - S}{N} \tag{10}$$

Success Rate (SR) has been expressed as a function of (α, β, q), as of Equation (11).

$$\phi_n(\alpha, \beta, q) = SR \tag{11}$$

where,
n : HMM structure $\{f$: fixed structure, v: variable structure$\}$.
α : Feature Type $\{MFCC, BTE\}$.
β : Entropy Function $\{SE, TE, RE\}$.
q : Entropy Order.

The success rate of the Fixed State Structure HMM Design using the two techniques of feature extraction is illustrated in Table 3. MFCC_0_D_A is taken as a reference. It shows that the best results are $\phi_f(BTE, RE, .2) = 64.62\%$, $\phi_f(BTE, SE, 1) = 65.14\%$, and $\phi_f(BTE, TE, 1.2) = 67.38\%$ which is compared to MFCC success rate of 70.37%.

The model health indicates the relative success of the model to the base model (MFCC_0_D_A). To quantize the model health, a new term has been introduced. The Health Factor (η_r) is the relative success rate of BTE that can be calculated by Equation (12).

$$\text{Health Factor } (\eta_r) = \frac{SR\ of\ BTE}{SR\ of\ MFCC} \tag{12}$$

According to this relative equation, BTE using Tsallis entropy for q= 1.2 can achieve $\eta_r = 95.75\%$ in the Fixed State Structure HMM Design as listed in Table 3.

TABLE 3

$(\phi_f(\alpha,\beta,q)\%)$ FOR THE FIXED STATE STRUCTURE HMM DESIGN

Feature Type (α)	**Entropy Function (β)**	**Entropy Order (q)**	$\boldsymbol{\phi_f(\alpha,\beta,q)\%}$	$\boldsymbol{\eta_r}$ **%**
MFCC			70.37	100
BTE	**SE**	q=1.0	65.14	92.56
BTE	**TE**	q=0.2	52.24	74.23
		q=0.3	54.20	77.02
		q=0.4	56.83	80.75
		q=0.6	60.03	85.30
		q=0.8	63.27	89.91
		q=1.2	67.38	95.75
		q=1.3	64.88	92.19
		q=1.4	62.32	88.56
		q=1.5	58.38	82.96
BTE	**RE**	q=0.2	64.62	91.82
		q=0.3	40.96	58.20
		q=0.4	62.99	89.51
		q=0.5	42.71	60.69

The relationship between ϕ_n, η_r can be expressed as follows:

$$\Gamma_n(\alpha,\beta,q) = \left.\frac{\phi_n(\alpha,\beta,q)\%}{\eta_r}\right|_C \tag{13}$$

where

C: $phone\ classes\ \{V,P,F,N,Si\}$.

Table 4 illustrates this relationship ($\Gamma_f(\alpha,\beta,q)$) for each class. BTE outperformed MFCC in the Fixed State Structure HMM Design in recognizing the Vowels class using Renyi entropy and in recognizing the Silences class using Renyi and Tsallis entropy. Figure 10 shows the success rate of BTE for each class in the Fixed State Structure HMM ($\phi_f(BTE,\beta,q)|_C\%$) using Shannon entropy, Tsallis entropy and Renyi entropy at different entropy order (q). It shows that $\phi_f(BTE,TE,.3)|_F = 78.8\%$, this is the best for Fricatives class (F). $\phi_f(BTE,SE,1)|_P = 79\%$, this is the best for Plosives class (P). $\phi_f(BTE,TE,.4)|_N = 88.6\%$, this is the best for Nasals class (N). $\phi_f(BTE,RE,.2)|_V = 97.2\%$, this is the best for Vowels class (V). $\phi_f(BTE,RE,.3)|_{Si} = 99.9\%$, this is the best for Silences class (Si).

TABLE 4

$(\Gamma_f(\alpha,\beta,q)\big|_C)$ FOR THE FIXED STATE STRUCTURE HMM DESIGN

Feature Type (α)	**Entropy Function (β)**	**Entropy Order (q)**	$\Gamma_f\big\|_V$	$\Gamma_f\big\|_P$	$\Gamma_f\big\|_F$	$\Gamma_f\big\|_N$	$\Gamma_f\big\|_{Si}$
MFCC			84.2 / 1.0	97.0 / 1.0	84.8 / 1.0	96.9 / 1.0	89.1 / 1.0
BTE	**SE**	q=1.0	69.2 / 0.82	79.0 / 0.81	47.4 / 0.55	76.5 / 0.78	86.0 / 0.96
BTE	**TE**	q=0.2	28.3 / 0.33	49.8 / 0.51	77.6 / 0.91	85.5 / 0.88	91.1 / 1.02
		q=0.3	35.7 / .42	39.6 / 0.40	78.8 / 0.92	88.3 / 0.91	89.4 / 1.0
		q=0.4	43.5 / 0.51	44.6 / 0.54	76.6 / 0.90	88.6 / 0.91	88.2 / 0.98
		q=0.6	52.3 / 0.62	49.5 / 0.51	73.7 / 0.86	88.4 / 0.91	87.2 / 0.97
		q=0.8	62.7 / 0.74	49.5 / 0.51	66.1 / 0.77	86.1 / 0.88	88.0 / 0.98
		q=1.2	78.5 / 0.93	50.2 / 0.51	58.3 / 0.68	64.8 / 0.66	92.2 / 1.03
		q=1.3	78.8 / 0.93	62.4 / 0.64	67.4 / 0.79	56.2 / 0.57	92.2 / 1.03
		q=1.4	79.1 / 0.93	72.6 / 0.74	61.0 / 0.71	47.3 / 0.48	94.3 / 1.05
		q=1.5	80.4 / 0.95	76.8 / 0.79	52.2 / 0.61	38.6 / 0.39	96.7 / 1.08
BTE	**RE**	q=0.2	97.2 / 1.15	26.8 / 0.27	34.8 / 0.41	30.4 / 0.31	86.8 / 0.97
		q=0.3	33.0 / 0.39	11.7 / 0.12	18.5 / 0.21	23.8 / 0.24	99.9 / 1.12
		q=0.4	94.7 / 1.12	39.9 / 0.41	15.6 / 0.18	39.9 / 0.41	88.2 / 0.98
		q=0.5	78.9 / 0.93	66.9 / 0.68	22.7 / 0.26	80.0 / 0.82	97.7 / 1.09

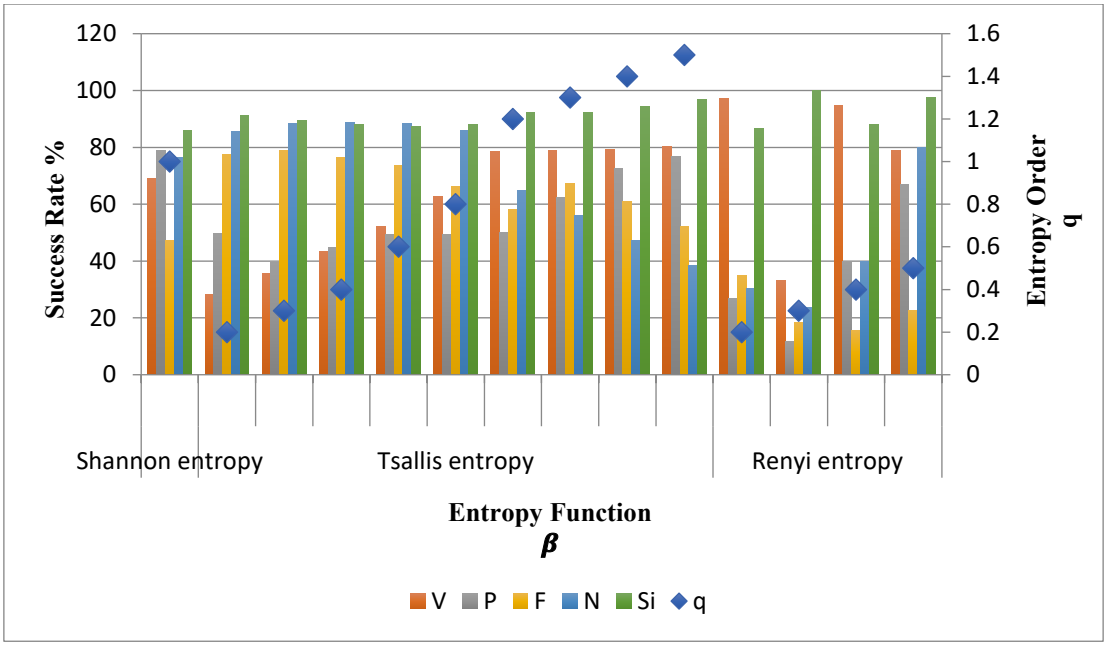


Figure 10: $\phi_f(BTE,\beta,q)|_C\%$ For the Fixed State Structure HMM Design

Table 5 illustrates the success rate of the Variable State Structure HMM Design using the two techniques of feature extraction. It shows that the best results are $\phi_v(BTE,RE,.2)=51.40\%$, $\phi_v(BTE,SE,1)=74.02\%$, and $\phi_v(BTE,TE,1.2)=75.85\%$, which is compared to MFCC success rate of 71.76%. BTE using Tsallis entropy for q= 1.2 can achieve $\eta_r = 105\%$ in the Variable State Structure HMM Design.

TABLE 5

$(\phi_v(\alpha,\beta,q)\%)$ FOR THE VARIABLE STATE STRUCTURE HMM DESIGN

Feature Type (α)	Entropy Function (β)	Entropy Order (q)	$\phi_v(\alpha,\beta,q)\%$	η_r %
MFCC			71.76	100
BTE	**SE**	q=1.0	74.02	103
BTE	**TE**	q=0.2	59.32	82.66
		q=0.3	61.83	86.16
		q=0.4	64.02	89.21
		q=0.6	66.85	93.15
		q=0.8	69.57	96.94
		q=1.2	75.85	105
		q=1.3	73.86	102
		q=1.4	69.68	97.10
		q=1.5	64.28	89.57
BTE	**RE**	q=0.2	51.40	71.62
		q=0.3	41.35	57.62
		q=0.4	50.56	70.45
		q=0.5	40.27	56.11

Table 6 illustrates the relationship $(\Gamma_v(\alpha,\beta,q))$ for each class. BTE outperformed MFCC in the Variable State Structure HMM Design in recognizing Vowels, Plosives, Fricatives, and Silences classes. Figure 11 shows the success rate of BTE for each class in the Variable State Structure HMM $(\phi_v(BTE,\beta,q)|_C\%)$ using Shannon entropy, Tsallis entropy and Renyi entropy at different entropy order (q). It shows that $\phi_v(BTE,TE,.2)|_F = 82.6\%$, this is the best for Fricatives class (F). $\phi_v(BTE,SE,1)|_V = 83\%$, this is the best for Vowels class (V). $\phi_v(BTE,TE,1.5)|_P = 94.4\%$, this is the best

for Plosives class (P). $\phi_v(BTE, RE, .5)|_N = 94.4\%$, this is the best for Nasals class (N). $\phi_v(BTE, RE, .4)|_{Si} = 98.4\%$, this is the best for Silences class (Si).

TABLE 6

$(\Gamma_v(\alpha, \beta, q)|_C)$ FOR THE VARIABLE STATE STRUCTURE HMM DESIGN

Feature Type (α)	**Entropy Function (β)**	**Entropy Order (q)**	$\Gamma_v\|_V$	$\Gamma_v\|_P$	$\Gamma_v\|_F$	$\Gamma_v\|_N$	$\Gamma_v\|_{Si}$
MFCC			79.1 / 1.0	94.4 / 1.0	82.4 / 1.0	95.1 / 1.0	76.6 / 1.0
BTE	**SE**	q=1.0	83.0 / 1.05	67.4 / 0.71	57.5 / 0.70	71.7 / 0.75	91.2 / 1.19
BTE	**TE**	q=0.2	36.0 / 0.45	68.1 / 0.72	82.6 / 1.0	87.5 / 0.92	92.8 / 1.21
		q=0.3	46.0 / 0.58	57.6 / 0.61	81.8 / 0.99	86.6 / 0.91	90.7 / 1.18
		q=0.4	54.9 / 0.69	58.5 / 0.62	80.5 / 0.98	84.8 / 0.89	88.8 / 1.16
		q=0.6	63.9 / 0.81	65.0 / 0.69	80.2 / 0.97	79.3 / 0.83	85.8 / 1.12
		q=0.8	70.9 / 0.90	67.9 / 0.72	76.1 / 0.92	74.5 / 0.78	86.7 / 1.13
		q=1.2	82.5 / 1.04	79.0 / 0.84	60.2 / 0.73	68.0 / 0.71	93.6 / 1.22
		q=1.3	76.8 / 0.97	89.8 / 0.95	58.2 / 0.71	58.7 / 0.62	94.4 / 1.23
BTE	**TE**	q=1.4	74.3 / 0.94	93.4 / 0.99	41.3 / 0.50	38.6 / 0.40	95.1 / 1.24
		q=1.5	73.8 / 0.93	94.4 / 1.0	26.3 / 0.32	27.2 / 0.29	96.0 / 1.25
BTE	**RE**	q=0.2	81.1 / 1.02	90.7 / 0.96	29.1 / 0.35	60.9 / 0.64	96.0 / 1.25
		q=0.3	66.3 / 0.84	94.4 / 1.0	34.3 / 0.42	58.3 / 0.61	94.5 / 1.23
		q=0.4	79.5 / 1.01	83.8 / 0.89	35.2 / 0.43	45.2 / 0.47	98.4 / 1.28
		q=0.5	40.8 / 0.51	77.4 / 0.82	38.0 / 0.46	94.4 / 0.99	94.0 / 1.23

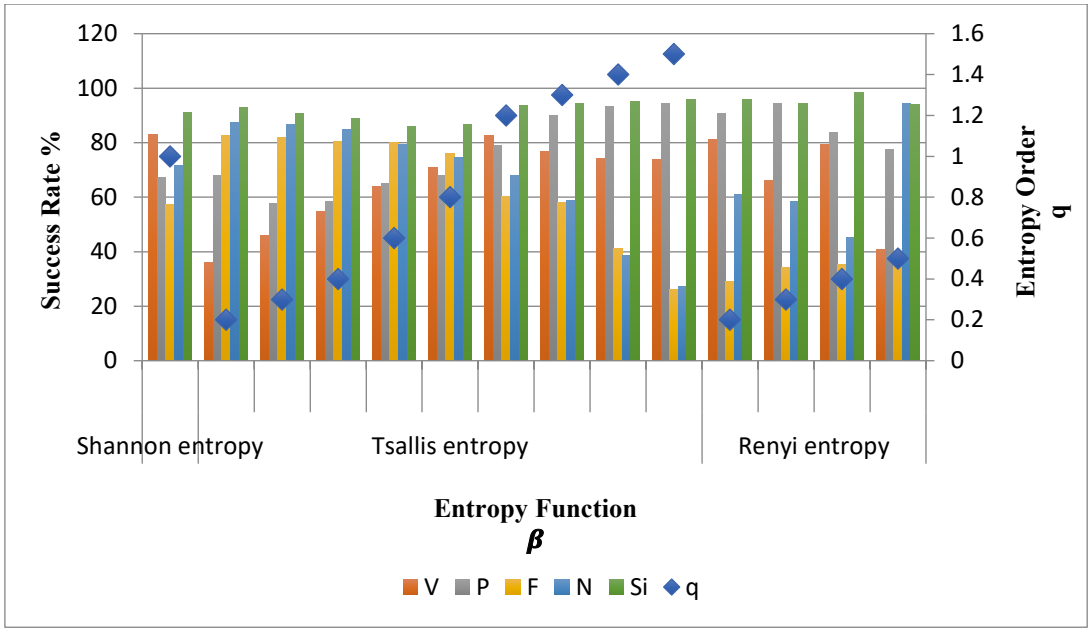


Figure 11: $\phi_v(BTE,\beta,q)|_C$% For the Variable State Structure HMM Design

From Table 4 and Table 6 the best success rate for each class and its indicating Health Factor (η_r) are plotted in Figure 12. It shows that the best ϕ for the Vowels and Silences classes was achieved using the fixed structure HMM. The best ϕ for the Plosives, Fricatives, and Nasals classes was achieved using variable structure HMM. The Health Factor $\eta_r > 1$ for Vowels and Silences classes in the fixed and variable structure HMM is an enhancement of the baseline model.

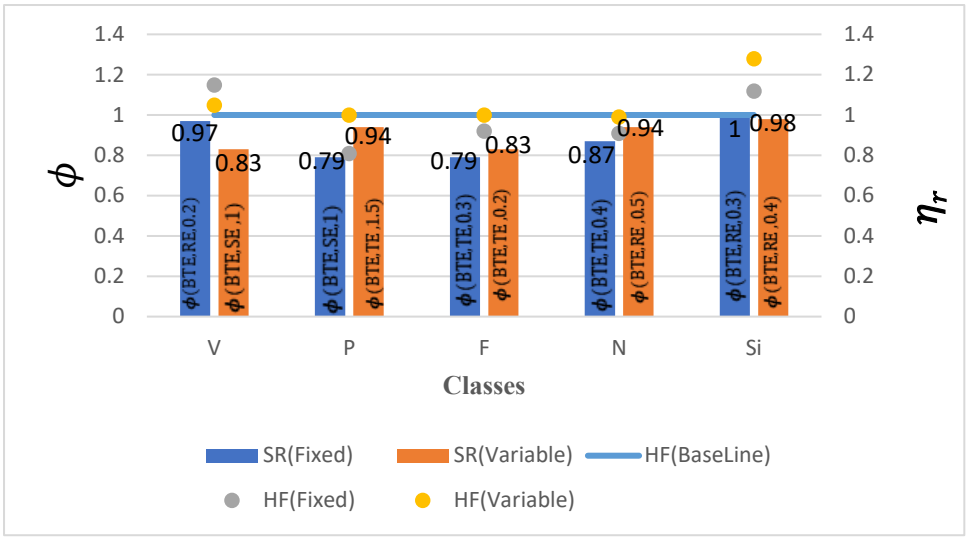


Figure 12: Best Results for Classes

By comparing our results with [17], which used the same database (TIMIT) and the same classes, the best result was achieved by the proposed approach BTE using Tsallis entropy, which equals 75.85% compared with [17], which equal 74.11 %. The vowels class was achieved by 97.2%; by comparing with [17], which is 91.7 %. Plosives class was achieved by 94.4% by comparing with [17], which is 92.5 %. The Silences class was achieved by is 99.9% by comparing with [17], which is 99 %. These results have been improved using various entropy functions with BTE feature extraction, which are Tsallis and Renyi entropy instead of Shannon entropy, which was used in [17].

6 CONCLUSIONS

This paper focused on enhancing the BTE feature extraction technique. The first issue that affects the success rate of BTE in classification is the type of entropy. Shannon entropy (SE), Renyi entropy (RE), and Tsallis entropy (TE) are used. The highest overall success rate of 75.85% was achieved using Tsallis entropy. The best success rate for the

Vowels class is 97.2%. This is achieved using fixed structure HMM and entropy RE with entropy order 0.2. The health factor indicates a 1.15 enhancement relative to the baseline model. The best success rate for the Plosives class is 94.4%. This is achieved using variable structure HMM and entropy TE with entropy order 1.5. The health factor indicates 1.0. The best success rate for the Fricatives class is 82.6%. This is achieved using variable structure HMM and entropy TE with entropy order 0.2. The health factor indicates 1.0. The best success rate for the Nasals class is 94.4%. This is achieved using variable structure HMM and entropy RE with entropy order 0.5. The health factor indicates 0.99. The best success rate for Silence's class is 99.9%. This is achieved using fixed structure HMM and entropy RE with entropy order 0.3. The health factor indicates 1.12 enhancement for the baseline model. The highest overall success rate 75.85% is achieved by BTE using Tsallis entropy compared to similar work that was using BTE with Shannon entropy in [17], which gave 74.11%. This indicates that Tsallis entropy is more efficient than the Shannon entropy. In the future, other classification techniques instead of HMM such as Recurrent Neural Network (RNN) can be used, also Convolutional Neural Networks (CNN) to improve the results. In addition, using the results obtained in this paper to enhance Mel scaled Best Tree Image (MBTI) used in [17] will achieve better results.

REFERENCES

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

[2] N. Desai, K. Dhameliya, and V. Desai, "Feature Extraction and Classification Techniques for Speech Recognition: A Review," *International Journal of Emerging Technology and Advanced Engineering*, vol. 3, no. 12, pp. 367-371, December 2013.

[3] B. S. Atal, Suzanne L. Hanauer, "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave," *The Journal of the Acoustical Society of America*, vol. 50, no. 2, pp. 637-655, 1971.

[4] Hasan, M.R., Jamil, M., Saifur Rahman, "Speaker identification using Mel Frequency Cepstral Coefficients," in *3rd International Conference on Electrical and Computer Engineering*, Dhaka, Bangladesh, 2004, pp. 256-258.

[5] R. L. K. Venkateswarlu, R. Vasantha Kumari, and A. K. V. Nagavya, "Efficient Speech Recognition by Using Modular Neural Network," *Int. J. Comp. Tech. Appl.*, vol. 2, no. 3, pp. 463-470, 2003.

[6] O. Farooq, S. Datta, "Mel filter-like admissible wavelet packet structure for speech recognition," *IEEE Signal Processing Letters*, vol. 8, no. 7, pp. 196-198, 2001.

[7] A. Vuppala, J. Yadav, S. Chakrabarti and K. Rao, "Vowel Onset Point Detection for Low Bit Rate Coded Speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 6, pp. 1894-1903, 2012.

[8] 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.*, Pine Mountain, GA, USA, 2002, pp. 37-40.

[9] A. A. R. O. Hebah H. O. Nasereddin, "Classification techniques for automatic speech recognition (ASR) algorithms used with real time speech translation," in *Computing Conference 2017*, London, UK , 2017, pp. 200-207.

[10] J. Keshet, D. Chazan, and B.-Z. Bobrovsky, "Plosive spotting with margin classifiers," in *7th European Conference on Speech Communication and Technology*, Aalborg, Denmark, 2001, pp. 1637-1640.

[11] G. Tryfou, M. Pellin, and M. Omologo, "Time-frequency reassigned cepstral coefficients for phone-level speech segmentation," in *22nd European Signal Processing Conference (EUSIPCO)*, Lisbon, Portugal, 2014, pp. 2060-2064.

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

[13] 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.

[14] 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.

[15] A. Rizal, R. Hidayat, and H. A. Nugroho, "Comparison of Multilevel Wavelet Packet Entropy using Various Entropy Measurement for Lung Sound Classification," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 10, no. 2, pp. 77-82, 2019.

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

[17] Doaa A. Lehabik, Mohamed H. Merzban, Sameh F. Saad, Amr 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.

[18] M. Y. Gokhale, Daljeet Kaur Khanduja, "Time Domain Signal Analysis Using Wavelet Packet Decomposition Approach," *Int. J. Communications, Network and System Sciences*, vol. 3, no. 3, pp. 321-329, March 2010.

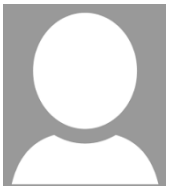
[19] J.-D. Wu, B.-F. Lin, "Speaker identification using discrete wavelet packet transform technique with irregular decomposition," *Expert Systems with Applications*, vol. 36, no. 2, pp. 3136-3143, 2009.

[20] Taiyong Li and Min Zhou, "ECG Classification Using Wavelet Packet Entropy and Random Forests," *Entropy*, vol. 18, no. 8, p. 285, August 2016.

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

[22] John S. Garofolo, Lori F. Lamel, William M. Fisher, Jonathan G. Fiscus, and David 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, p. 27403, 1993.

BIOGRAPHY



Fatma M. Abd El_latif received her B.Sc. degree in Electrical Engineering – Communications and Electronics Department with an excellent and honor degree from the Faculty of Engineering - Fayoum University, 2014. She joined the teaching staff of the Mathematics and Physics Department, Faculty of Engineering, Fayoum University, Egypt, in 2015. She joined the M.Sc. program at Fayoum University - Mathematics and Physics Department in 2016. Her areas of interest include entropy and its application in automatic speech recognition.

Amr 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 the 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).

Waleed A. Maguid Ahmed was born in Cairo, Egypt, 1974. He received the B.Sc. degree (Distinction with honors) in Electronics and Communication Engineering from Cairo University, Cairo, Egypt in 1996, and the M.Sc. and Ph.D. degree in Engineering Mathematics and Physics from Cairo University, Cairo, Egypt in 2001, and 2005 respectively, He was promoted to associate professor in 2014. Since 1997 he has been with the department of Engineering Mathematics and Physics, Fayoum university, Fayoum, Egypt. where he is now an Associate Professor. Since 2017 he has been with the applied mathematics program in Zewail City of Science and Technology as a Full-time associate professor. His main research interests are in Fractional Fourier Analysis and its Engineering Applications, Fractional Calculus and its Engineering Applications, Analysis of Neural Networks, and Fluid Mechanics.

البحث عن الإنتروبي الأمثل لتحسين بنية الشجرة المثُلى لحزمة المويجات لإنتاج ميزات أكثر ملاءمة لمهمة التعرف الآلى على الكلام

فاطمه محمد عبد اللطيف[*1] , عمرو محمد جودي [**2], وليد عبد المجيد أحمد [*3]
[*] قسم الرياضيات والفيزيقا الهندسية, كلية الهندسة, جامعة الفيوم, الفيوم, مصر.
[1]fma06@fayoum.edu.eg
[3]waa01@fayoum.edu.eg
[**] قسم الهندسة الكهربية, كلية الهندسة, جامعة الفيوم ,مصر.
[2]amg00@fayoum.edu.eg

ملخص

نقدم تقنية مطورة حديثًا لخصائص الكلام والمصممة لمهمة التعرف الآلى على الكلام. هذه التقنية تسمى تشفير الشجرة المثُلى (BTE). يقدم هذا البحث تحسينًا لحسابات (WPBT) الشجرة المثُلى لحزمة المويجات. تم ترميز أفضل شجرة (BTE) باستخدام نموذج رياضي بمتجه مكون من 4 مكونات. أفضل هيكل شجرة تم حسابه باستخدام دالة الانتروبي. في الإصدار القياسي من تشفير الشجرة المثُلى ، تم اختيار إنتروبيا شانون كوظيفة إنتروبي. أما في هذا البحث ، فقد تم استخدام إنتروبيا شانون (SE) ، إنتروبيا ريني (RE) وانتروبيا تساليس (TE) لبناء أفضل شجرة. تم إجراء تشفير الشجرة المثلى باستخدام نفس نهج النموذج الرياضي المستخدم في معيار BTE المكون من 4 مكونات. تم اختبار النموذج المقترح والتحقق منه مقارنة بتقنية استخراج الخصائص الأكثر استخدامًا على نطاق واسع و هى معامل ميل التردد (MFCC) بالإضافة إلى معاملات دلتا ودلتا-دلتا (39 معامل) لتقييم أدائها. تم استخدام قاعدة بيانات TIMIT في هذا البحث. جميع المقاطع الصوتية مقسمه إلى 5 مجموعات وهي حروف متحركة (V) وحروف احتكاكية (F) وصامت الذي لا يحتوي على اي كلام (Si) وحروف انفية (N) وحروف انفجارية (P). تم تنفيذ النموذج الصوتي باستخدام نموذج ماركوف المخفي (HMM). لم يتم تطبيق أي نموذج لغوي . يتم استخدام برنامج HMM Tool Kit (HTK) لتنفيذ النموذج. أظهرت التجارب أن BTE باستخدام الانتروبيا Tsallis ينتج أعلى معدل نجاح بنسبة 75.85٪ مقارنة بمعدل نجاح 71.76٪ لـ MFCC. إن مقارنة المتجه المكون من 4 مكونات من BTE بمتجهًا مكونًا من 39 مكون لـ MFCC يجعله ناقل للخصائص واعدًا للغاية للبحث والتطوير.

الكلمات المفتاحية

التعرف الآلي على الكلام، الشجرة المثلى للتعرف الآلي على الكلام، تحليل حزمة المويجات، نموذج ماركوف المخفي، Shannon إنتروبي، Renyi إنتروبي، Tsallis إنتروبي، معامل cepstral Mel للتردد (MFCC).