Automatic Speech Segmentation Using Hybrid Wavelet Features and HMM

Amr M. Gody*[1], Manal Shaaban*[2], Amr Saleh*[3]

[*] *Electrical Engineering Department, Faculty of Engineering, Fayoum University*
El-Fayoum, Egypt

[1]amg00@fayoum.edu.eg

[2] manal.mohammed66@yahoo.com

[3]aae00@fayoum.edu.eg

Abstract: ***In this research, a novel feature set is used to automatically segment speech signal. Automatic segmentation is very useful especially for large database. A hybrid features model is created from wavelet packet analysis and mel-scale is used to train Hidden Markov Model (HMM) for phone boundary detection. HMM is implemented using the Hidden Markov Model Toolkit (HTK).The database (Ked-TIMIT) is used for result verifications and Mel Frequency Cepstral Coefficients (MFCC) is used as reference for evaluating the results of the proposed Hybrid model. The results are categorized for vowels, consonants and short phones. Phone duration and start location are used as metrics to evaluate the system success rate. Success rate of 74% is achieved for consonant detection, 72% for vowel detection and 58% for short phone detection. Using the simple metric that relies only on boundary locations but ignoring duration, the achieved results are 92.5% for consonant detection, 90% for vowel detection and 77.5% for short phoneme detection. In addition to boundary detection the proposed hybrid model is utilized to compare newly developed features called Mel scale Best Tree Encoding (Mel-BTE) to the mostly used popular features MFCC along with all experiments using the same database. The relative results for Mel-BTE with respect to MFCC are 94.77% for consonant detection, 87.5% for vowel detection and 93.33% for short phoneme detection.***

Key words*:Mel scale, BTE, MFCC, HTK, Gaussian Mixture, Speech Segmentation.*

1 INTRODUCTION

Phone segmentation is a process of finding the boundaries of a sequence of known phones in a spoken utterance. Phone segmentation process is still an active topic, as is shown by the range of research directions suggested in this section.

In [2] the authors proposed a hybrid optimization scheme for clear phonetic segmentation applied on the TIMIT database, they used both context independent (CI) and context dependent (CD) models. Table I shows an improved Mean absolute error (MAE) and root mean square error (RMSE)of the segmentation of Support Vector Machine (SVM) to get the phone boundary using corrective statistical model on different frame duration (Time resolution). Below is part of the table from [2] that includes the best results and the best technique according to their experimental results. The grayed cell is that one that is compared to the work in this research. In this research the time resolution is 20 (ms). The obtained results using the proposed hybrid model is about 91%, which is competitive to the achieved results in [2] but in this research paper the model is much simpler than the proposed model in [2].

TABLE I
REFINEMENTS WITH PREDICTIVE MODEL [2]

Time Resolution of acoustic model	<5 (ms) %	<10 (ms) %	<20 (ms) %	<30 (ms) %	<40(ms) %	<50 (ms) %	MAE	RMSE
SVM	58.02	81.31	94.19	97.56	98.7	99.36	6.54	11.92

In [3] the authors proposed phone boundary models significantly improved forced alignment accuracy applied on the Mandarin Chinese database. The system achieved 93.1% agreement (of phone boundaries) within 20 (ms) compared to manual segmentation on the test set without boundary correction.

In [4] the authors proposed a combination of special one-state phone boundary models and mono phone HMMs .The proposed system achieves 93.92% agreement (of phone boundaries) within 20 (ms) compared to manual segmentation on the TIMIT database.

In [5] the authors proposed a neural network method which is used to learn the mapping between phoneme boundaries and auditory attention features. The proposed method achieves 86.8% phoneme boundary detection accuracy at frame-level when tested on TIMIT database.

In [6] the authors proposed a series of objective functions for optimal segmentation, Log Determinant (LD), Rate Distortion (RD), Bayesian Log Determinant (BLD), Mahalanobis distance (MD), and Euclidean distance (ED). The

proposed five measures are compared through experiments on TIMIT database. RD achieves the highest recall rate among the five objectives. They found that their method achieved a correlation of 92% with manual scores on utterance-level evaluation and a correlation of 79%with TOEIC scores.
In [7] the authors proposed a text-independent method for boundary correction applied on the TIMIT database. Boundary correction improves the segmentation by 2.3% relative for manual and 3.0% relative for automatic transcription for a 20(ms) maximum deviation.
In [8] the authors proposed a hybrid architecture for automatic alignment of speech waveforms and their corresponding phone sequence. They achieved an accuracy of 83.56%, which corresponds to improving the baseline system's accuracy by 6.09 %.
In [9] the authors proposed a phoneme segmentation system which had two models. In the first model 94.4% phoneme recognition accuracy with 95.2% of phoneme boundaries is reached within 70 (ms). For the second model, phoneme recognition accuracy increased to reach 96.8% with 96.1% of phoneme boundaries within 70 (ms).
In[10] the authors proposed two novel approaches(DTW and HMMs) to phonetic speech segmentation. In the DTW approach 90% phoneme recognition accuracy is reached within <15 (ms). For the HMMs approach, 93%phoneme recognition accuracy is reached within <15(ms). Results reached by HMM-based system are close to that given by the manual segmentation.
In [11] the authors proposed studying the performance of automatic phone segmentation from two viewpoints. The first point of view is temporal precision and the second point of view is effect on the naturalness of synthetic speech. The absolute error of the phone onset time for the best 90% and worst 10% were 4.6 (ms) and 25.9 (ms) respectively.

2 PROCEDURE

The experiments are implemented according to the following procedure:
Ked-TIMIT database is processed for preparing reference database. Data includes a list of information records {KEY, SYM, Start (ms), End (ms), Duration (ms), SEQ}, where:

KEY : is a unique identification of the sample.

SYM : is a text representation of the associated phone.

Start(ms) : is a boundary start in ms.

End(ms) : is a boundary end in ms.

Duration(ms) : is the duration of the associated Symbol in ms.

SEQ : is the order of the symbol in the associated sample.

1- BTE features extractor function is applied on all Ked-TIMIT databases to construct {SET A}.
2- MFCC features extractor function is applied on all Ked-TIMIT databases to construct {SET B}.
3- HMM models are constructed for phone recognition task using HTK for both sets {SET A} and {SET B}. The models are created for different Gaussian mixture counts. The models are identified by the models-group identifier {A1, A2, A3…etc.} and {B1, B2, B3…etc.}, where for example A1 means that models for {SET A} and single Gaussian and B2 means that models for {SET B} and 2 Gaussians are used in each state of the associated HMMs.
4- Training and testing tasks of the HMM models created in step 3 are done using HTK. This step creates recognition results for each models-group created in step 3. The results are identified as REC_A1, REC_A2…etc. and REC_B1, REC_B2…etc.,where REC_A1 is the results of Group A1 models. This can be read as such REC_A1 is the recognition results of BTE based phone recognizer with single Gaussian into each emitting state.
5- Database is prepared for each Recognition-Group in step 4. The database schema is similar to the reference database in step1. {KEY, SYM, Start (ms), End(ms), Duration(ms), SEQ}.
6- Boundary adjustment algorithm is applied on the databases generated in step 5.
7- Database consolidation is implemented in order to group the results for each symbol. This is to construct a result database for each models-group. The schema of the results is {SYM, GM, DD_A%, DD_B%, PD_A%, PD_B% }, where SYM is the phone symbol and GM is the Gaussian Mixture Count.
   DD_A% : is the average duration deviation percentage off the reference database for {SET A}.
   DD_B% : is the average duration deviation percentage off the reference database for {SET B}.
   PD_A% : is the average boundary start deviation percentage off the reference database for {SET A}.
   PD_B% : is the average boundary start deviation percentage off the reference database for {SET B}.
8- Results from step7 are segregated into {Voiced, Unvoiced, short and long phones}. Charts are constructed to view the comparison between Set A and Set B.

In the next sections the procedure steps will be illustrated in details. By the end of this paper, conclusions will be provided to explain the obtained results.

3 DATABASE PROCESSING

In this research KED TIMIT database is used[12]. The database contains 453 utterances spoken by a US male speaker. This database was collected at University of Edinburgh's Centre for Speech Technology Research. The database is hand labeled and carefully corrected.
The database includes both the digital samples files and the associated label files. The processing of the database is illustrated through the below steps:

1- Converting all label files into HTK standard format label files. This conversion is illustrated in Table II.

TABLE III
KED TIMIT LABEL FORMAT VERSUS HTK STANDARD FORMAT

Initial label format	HTK format		
0.399028 121 sil ; ref 0;	0	3990280	sil
0.481458 121 sh ; ref 1 ;	3990280	4814580	sh
0.556038 121 iy ; ref 2 ;	4814580	5560380	iy
0.604122 121 h ; ref 3 ;	5560380	6041220	h
0.666925 121 ae ; ref 4 ;	6041220	6669250	ae
0.679682 121 dcl ; ref 5 ;	6669250	6796820	dcl
0.687533 121 d ; ref 6 ;	6796820	6875330	d

The units in HTK standard format is in 100(nano seconds) but in the original Ked TIMIT the units are in seconds. The row in HTK format includes both boundaries of the associated phone but in the original format it is just the end boundary of the associated phone.

2- Importing the database into schema table to make it easy for querying information. The original database is delivered as twin files for each utterance. Each utterance is a one set of two files like this sample {kdt_001.lab, kdt_001.wav}. The name of the file is the identifier of the associated sample. As explained in step 1, all label files are converted into standard label format. In this step all label files are consolidated into schema data table. The record in the table is formatted like this part shown in TableIV.

TABLEV
DATA TABLE SCHEMA FOR THE CONSOLIDATED LABEL FILES

KEY	SYM	Start(ms)	End(ms)	Duration(ms)	SEQ
kdt_001	Sil	0	0.399028	0.399028	1
kdt_001	Sh	0.399028	0.481458	0.08243	2
kdt_001	Iy	0.481458	0.556038	0.07458	3
kdt_001	H	0.556038	0.604122	0.048084	4

where:

KEY : is unique identification of the sample.
SYM : is text representation of the associated phone.
Start(ms) : is boundary start in ms.
End(ms) : is boundary end in ms.
Duration(ms) : is the duration of the associated Symbol in milli-seconds.
SEQ : is the order of the symbol in the associated sample.

4 FEATURES EXTRACTION

In this step, the features are extracted from the database. The output of this step is the files that will be used into the subsequent recognition tasks. Two features models are used in this work, BTE and MFCC.
The output is grouped into two sets. Set A is for BTE files and Set B is for MFCC files. BTE with mel-entropy is considered into this research.

A. BTE with mel-entropy

First, an idea of Best Tree Encoding should be introduced. The process of creating BTE file [13] starts by producing frames of the speech signal. The second step is the preprocessing phase. Wavelet packet decomposition (WPD) is used in the preprocessing phase. Next step is to select the proper entropy type. Then get the best tree that contains the significant

signal power using the entropy obtained from the previous step. The last step is to encode the tree structure into 4-Dimensional vector of integer values [1]. The figure which describes this model is shown in [13].

New direction is considered in the version of BTE(BTE with mel-entropy). This version is proposed in [14]. The algorithm of estimating the best tree is targeted in this version of BTE. Mel scale is included to estimate the best tree nodes. Moreover Mel-Scale; resembling the original waveform is considered to map the bandwidth to 5(KHz).
The formula which is used for MS (f_{Mel}) is given as follows:
$f_{Mel} = 2595 * log_{10}(1 + \frac{f_{Hz}}{700})$, where f_{Hz} is the frequency in Hertz.
In this approach, the weight is calculated for each node based on the position of this node on the MS curve [14]. Nodes on low frequency band will be given high weights which indicate high ability of human hearing and vice versa.
An improvement for Best Tree Encoding is considered in this version of Mel scale BTE. This improvement is detailed in [15].The hybrid model reordering technique is implemented over BTE5 and BTE 7 to enhance the encoding.

B. MFCC features

In MFCC approach [16], Mel Frequency Cepstral Coefficients (MFCCs) is a feature widely used in automatic speech and speaker recognition. Mel-Frequency analysis of speech is based on human perception experiments. It is observed that human ear acts as filter.
Otherwise, MFCC is based on known variation of the human ear's critical bandwidth with frequency. A particular pitch is present on Mel Frequency Scale to pick up important characteristics of phonetics in speech. Figure of the process of the MFCC shown in [16].

5 HMM MODELS DESIGN

In this step, the design approach of the HMM models will be illustrated. Each phone into the database is altered individually. The following assumptions are included in this work:

1- No consideration of phone context is considered in this design. Only baseline phones are included. This is to avoid sophistications needed for state tying due to database limitations.
2- No consideration of the long or short durations of the phone is considered. This is to simplify the HMM design of this preliminary research.
3- No discrimination between silence and speech pause is considered.
4- All symbols are assumed as 3 states.
5- Gaussian mixtures with different counts are considered to construct the observation symbol probability function

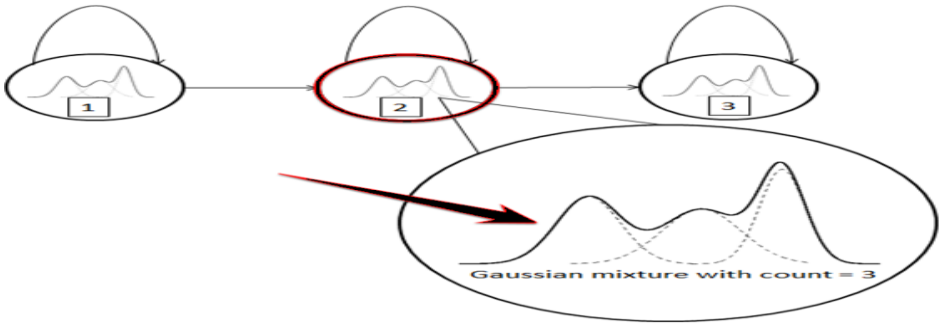


Figure 1: State model design of phone recognition with Gaussian mixture for symbol probability is considered

Figure 1 illustrates the model design. It consists of 3 states. Each state is modeled using Gaussian mixture. The count of mixtures is variable. It is one of the experiment parameters. HTK is used for training and testing the designed models.
This design approach is considering that the phone is 3 states, two transient states and one intermediate stationary state. Each state can accommodate as much as 3 or 4 sequence of frames. This makes it last for about 20 x 3 = 60(ms). This comes up with model that can cover at most 60 x 3 = 180 (ms). State transition penalty will not allow for more than 3 consecutive frames. This duration is suitable for almost all symbols excluding silence, speech pauses and short duration phones like the plosive phones. This may be considered as a weakness point in this research that will be concluded later on in the conclusions section.

6 TRAINING AND TESTING THE HMM MODELS FOR PHONE RECOGNITION

In this step, the models created in section 8 will be trained for phone recognition. The models will be trained using Set A and Set B databases that are constructed in section 7. HTK is used for the training and testing tasks for its popularity in the area of Automatic Speech Recognition [17, 18].The following assumptions are considered for the training:

1- Phones appear in balanced frequency. No phones are ignored accordingly.
2- Silence and Speech pauses are the same.
3- No state tying for bi-phone and tri-phone similar groups.

The output of this step is the recognition files. They are in the standard master label format from HTK. This format is illustrated in Figure 2.

The Recognition files are consolidated into schema data table like that one provided by Table VI.

```
#!MLF!#
"*/kdt_001.rec"
0 200000 sil -47.628166
200000 400000 ng -32.497784
400000 600000 ng -32.497784
600000 1200000 ai -103.804939
1200000 1600000 ix -66.654396
```

Figure 2 MLF file format for the HTK recognition results

7 BOUNDARY ADJUSTMENT ALGORITHM

In this step, the algorithm for boundary adjustment will be illustrated. The database in section 8 for the test samples will be processed against some rules to adjust the boundaries and to remove the errors. Two methods are included, the single duration ranking and the Tri-duration ranking.

The algorithm is illustrated in Figure3. The recognition file for certain utterance is identified by "Recognition File". The recognition file is the group of records from the data table for a certain set and a certain key as previously explained in the earlier sections. The reference transcription is the sequence of the symbols that represents the transcription of the key. It is not including the segmentation information.

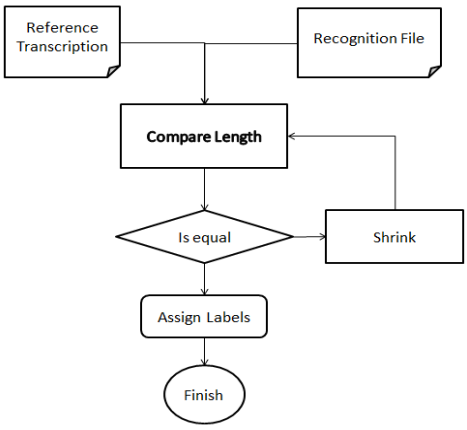


Figure 3: Boundary Adjustment Algorithm

Comparison is done between the count of phones from the reference transcription {CRef} and the count of segments from the recognition file {CRec}. If CRec > CRef, then the shrink algorithm will be implemented to remove a single boundary. The loop continues till the counts are equal. Then the reference transcriptions are assigned to the boundaries.

There are 2 shrinking methods considered in this work. The first one is the single duration shrinking and the other one is the Tri-duration shrinking.

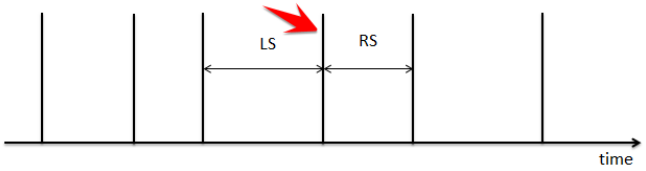


Figure 4 Single duration rank. The arrow is pointing to the boundary for which the rank is being evaluated

The single duration shrinking is illustrated in Figure 4. The method iterates all boundaries. For each boundary the Left space {LS} and the Right Space {RS} are estimated. The rank is evaluated by equation 1.

$$Rank(i) = \begin{cases} \infty & i = 1 \\ LS & LS < RS \; i = 1 \dots N \\ RS & RS < LS \end{cases} \quad (1)$$

where i is the index of the boundary. The boundary with lowest rank is removed according to equation 2, where j is the index of the boundary with lowest rank value.

$$j = MIN_i\{Rank(i)\} i \forall \{1 \rightarrow N\} \quad (2)$$

The Tri-duration method is slightly different of the single duration method. The rank of the boundary to be removed is evaluated using equation 3.

$$Rank(i) = \begin{cases} \infty & i = 1 \\ BP + CP + NP & i \forall \{2 \rightarrow N-1\} \\ BP + CP & i = N \end{cases} \quad (3)$$

where BP = previous period in (ms), CP = Current duration in (ms) and NP = Next duration (ms)
The symbols and the keywords are illustrated in Figure 5.

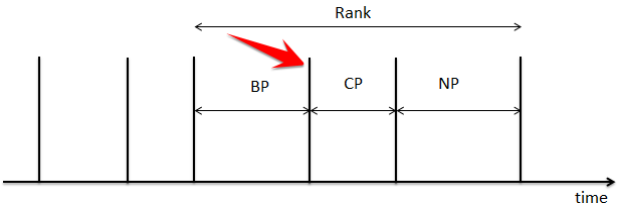


Figure 5 Tri-duration rank. The arrow is pointing to the boundary for which the rank is being evaluated

By the end of this step, the recognition results are aligned into the result table to the reference table. The same key in the test will be of the same length of boundary ticks as the reference.

8 RESULTS

In this step, the results will be illustrated. All samples are aligned as of the illustration in section 9. The key measurement equations are the duration deviation percentage {DD} and the boundary position deviation percentage {PD}.

$$\delta_d = \frac{\tau_r - \tau_0}{\tau_0} \times 100\% \quad (4)$$

$$\delta_p = \frac{l_r - l_0}{l_0} \times 100\% \quad (5)$$

where
τ_r: Duration in (ms) of the test unit. (Set A, BTE)
τ_0: Duration in (ms) of the reference unit. (Set B, MFCC)
l_r: The position of the phone start in (ms) measured from the sample 0 position. This is for (Set A, BTE).
l_0: The position of the phone start in (ms) measured from the sample 0 position. This is for (Set B, MFCC).
To illustrate the obtained results let us recall both Table VII and Figure2. Chart into Figure2 is representing the data in Table VIII. Each row in Table IXV is a summary of one phone results for Set A and Set B. The results are all in percentage as explained in equations 4 and 5.

TABLE XV
SAMPLE RESULTS COMPILED INTO TABLE.

Symbol Index	***Symbol***	$\delta_d(Set\ A)$	$\delta_d(Set\ B)$	$\delta_p\ (Set\ A)$	$\delta_p\ (Set\ B)$
1	Sil	-60	6.25	0	0
2	D	640.7407407	362.962963	-60	6.25
3	Ow	138.8059701	123.880597	-15.690867	28.8056206
4	N	244.8275862	417.2413793	5.26315789	41.7004049

The negative value indicates less than the reference value. The positive value indicates more than the reference value. For duration this is indicating the % deviation from the reference duration. The reference duration is the exact duration according to the database human verified segmentation. To explain this deviation from the reference, the ray chart is including circle at value 0.

This Zero-Circle is highlighted in dark black as shown in Figure 6. Set A and Set B are discriminated in the gray level such that Set A is darker than Set B as shown in Figure 6. This convention will be implemented in all the result graphs.
The obtained results are segregated into Gaussian mixture, single duration and Tri duration techniques. They are also grouped by vowels, consonants, long and short duration.

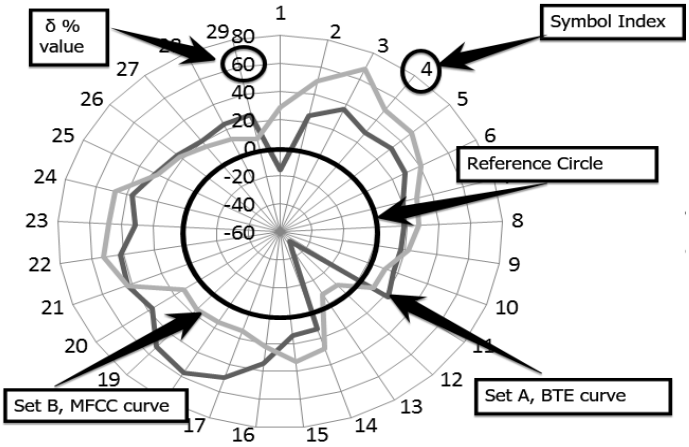


Figure 6 sample chart with info graphic

Case 1: Vowels comparison for different Gaussian mixture count

Figure 7 and Figure 8 illustrate the comparison results of vowels phones. Figure 7 indicates the effect of GM count as well as the segmentation method {Single duration or Tri duration} on the **phone duration** deviation δ_d.

As shown in Figure 7, the best curve is that one for GM = 2 and Tri duration method. The curve is selected by visual inspection. It is clear that the curves in this selection are the closest to the Zero-Circle than the other option of GM and segmentation method.

Using visual inspection in Figure 7, it is also clear that Set B curve is much closer to the Zero-Circle than of Set A curve. Both curves for Set A and Set B are almost identical but in some symbols there are sudden drop in Set A. Set A is almostfluctuating between 20% and 30% (The darker curve as of the convention in Figure 6), while Set B is almost fluctuating between 0% to 20%.

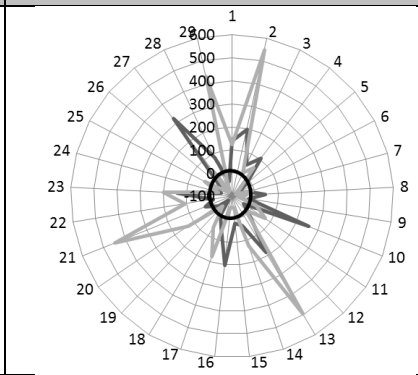
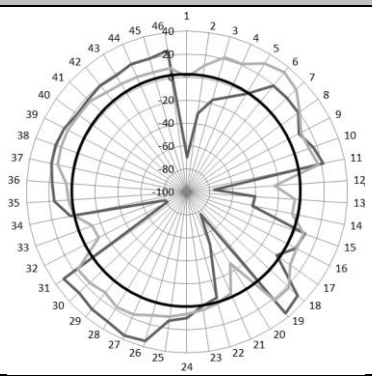
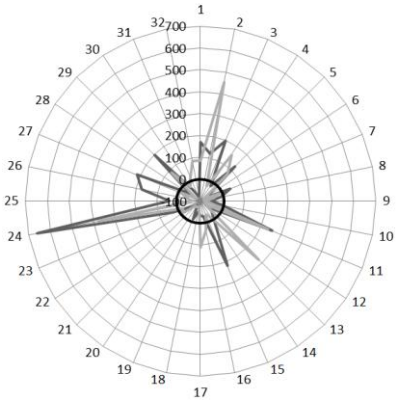
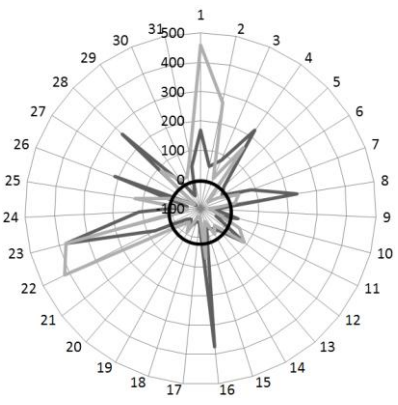
GM	Single Duration	Tri-Duration
2		
4		

Figure 7 Vowels comparison chart for duration deviation

Figure 8 illustrates the position deviation for the vowels phones. The curves are much better than the duration deviation. They don't include sudden deviations from the Zero-Circle. But by visual inspection it is clear that the best curves are in GM = 2 and Tri duration. Excluding the odd cases, Set A is almost fluctuating between 7% and 25% and Set B is almost fluctuating about 10% to 30%.

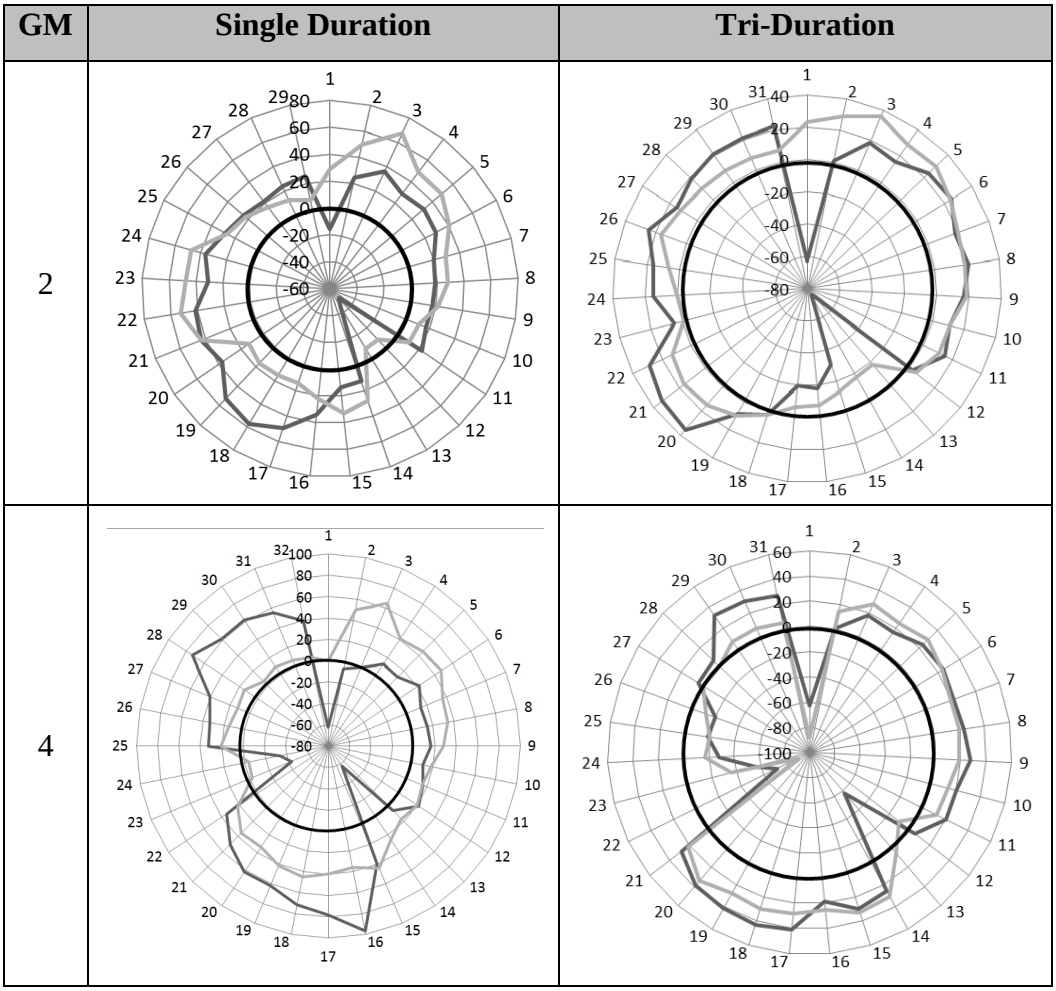


Figure 8 Vowels comparison chart for position deviation

Case 2: Consonant comparison for different Gaussian mixture count

Figure 9 and Figure10 illustrate the comparison results of Consonant phones. Figure 9 indicates the effect of GM count as well as the segmentation method {Single duration or Tri duration} on the **phone duration** deviation δ_d.

As shown in Figure 9, the best curve is that one for GM = 2 and Tri duration method. The curve is selected by visual inspection. It is clear that the curves in this selection are the closest to the Zero-Circle than the other option of GM and segmentation method.

Using visual inspection in Figure 9, it is also clear that Set B curve is much closer to the Zero-Circle than of Set A curve. Both curves for Set A and Set B are almost identical but in some symbols there are sudden drop in Set A. Set A is almost fluctuating between 10% to 20% while Set B is almost fluctuating between 0% to 15%.

Figure 10 illustrates the position deviation for the vowels phones. The curves are much better than the duration deviation. They don't include sudden deviations from the Zero-Circle. But by visual inspection it is clear that the best curves are in GM = 2 and Tri duration. Excluding the odd cases, Set A is almost fluctuating between 10% and 25% and Set B is almost fluctuating about 10% to 30%.

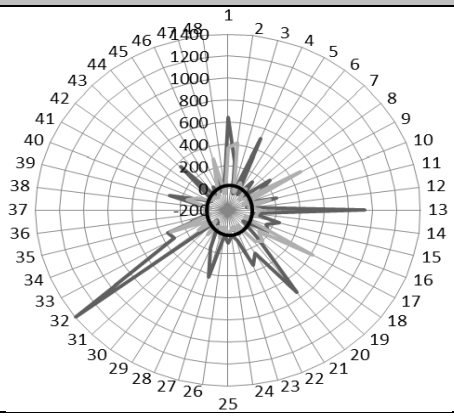
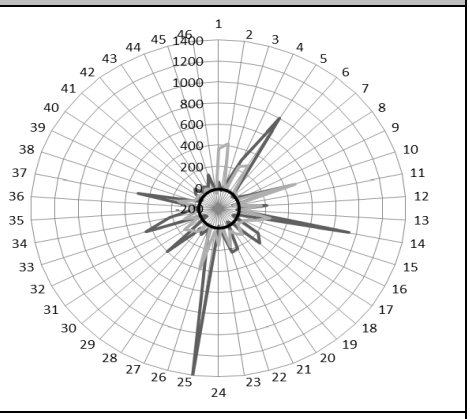
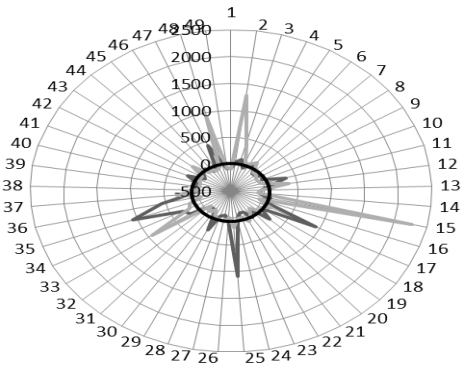
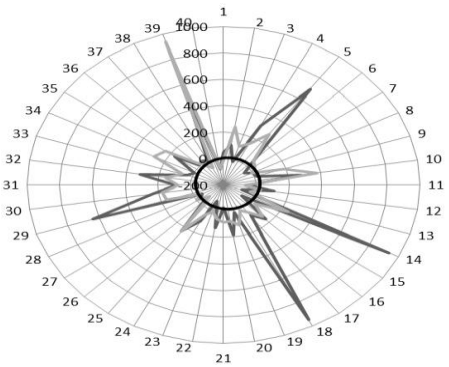
GM	Single Duration	Tri-Duration
2		
4		

Figure 9 Consonant comparison chart for duration deviation

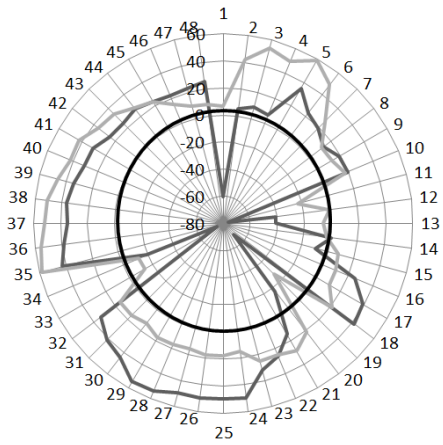
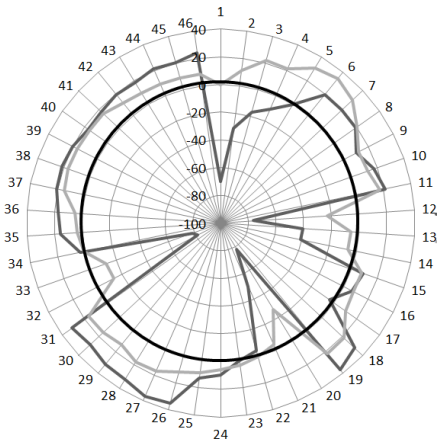
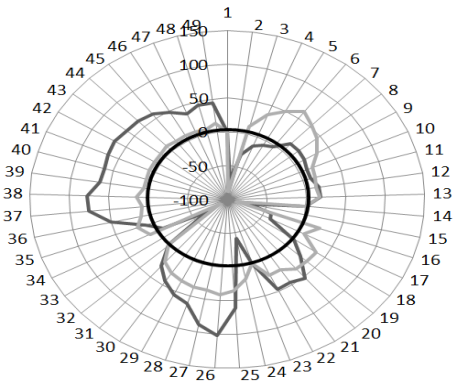
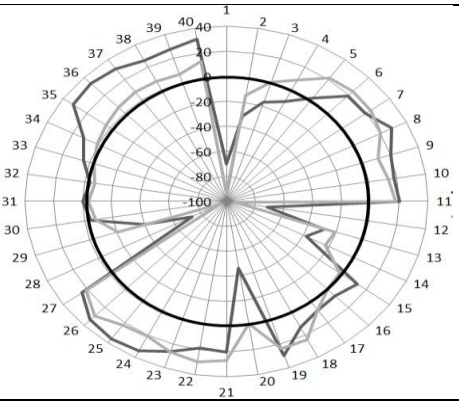
GM	Single Duration	Tri-Duration
2		
4		

Figure 10 Consonant comparison chart for position deviation

Case 3: Short phones comparison for different Gaussian mixture count

Figure 11 and Figure 12 illustrate the comparison results of short phones. Figure 11 indicates the effect of GM count as well as the segmentation method {Single duration or Tri duration} on the phone duration deviation δ_d.
As shown in Figure 11, the best curve is that one for GM = 2 and tri duration method. The curve is selected by visual inspection. It is clear that the curves in this selection are the closest to the Zero-Circle than the other option of GM and segmentation method.

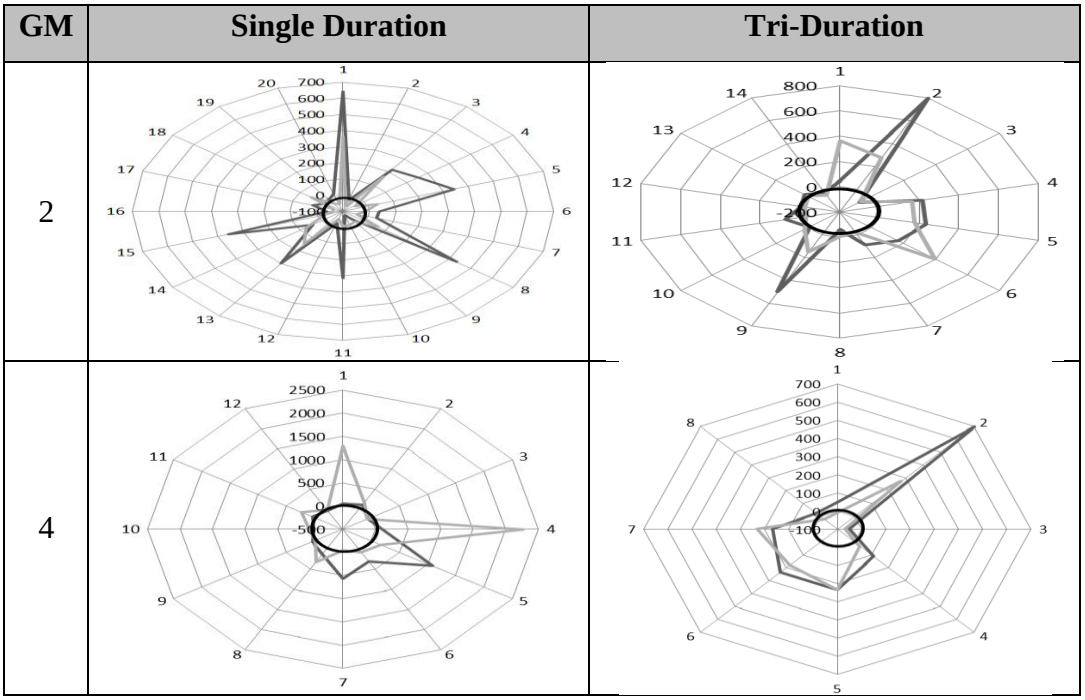


Figure 11 Short phones comparison chart for duration deviation

Using visual inspection in Figure 11, it is also clear that Set B curve is much closer to the Zero-Circle than of Set A curve. Both curves for Set A and Set B are almost identical but in some symbols there are sudden drop in Set A. Set A is almost fluctuating between 20% to 40% while Set B is almost fluctuating between 10% to 35%.

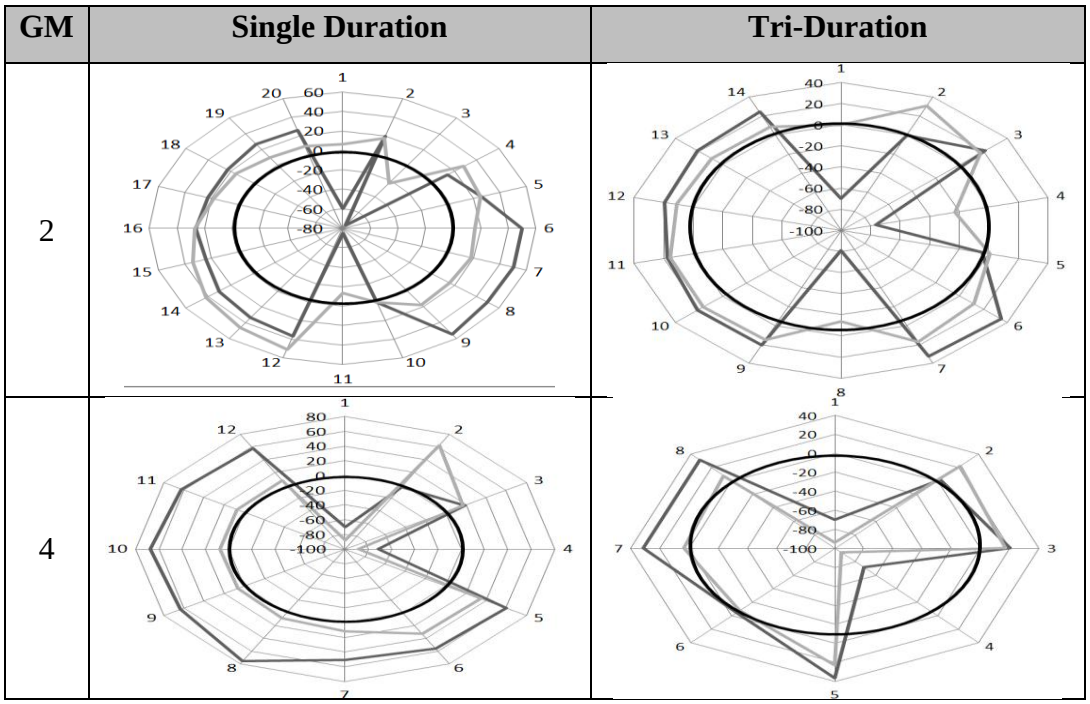


Figure 12 Short phones comparison chart for position deviation

Figure 12 illustrates the position deviation for the vowels phones. The curves are much better than the duration deviation. They don't include sudden deviations off the Zero-Circle. But by visual inspection it is clear that the best curves are in GM = 2 and Tri duration. Excluding the odd cases, Set A is almost fluctuating between 10% and 35% and Set B is almost fluctuating about 10% to 40%. Short phones have more sudden deviations off the Zero-Circle than any phones.

TABLE XI
CONSOLIDATED RESULTS

		Position		**Duration**		**Consolidated Results**		
Category	**Features set**	**Low** $\delta_{p_{low}}$	**High** $\delta_{P_{High}}$	**Low** $\delta_{d_{low}}$	**High** $\delta_{d_{High}}$	**Average duration Error** e_d %	**Average Position Error** e_p %	**SR %**
Vowels	Set A	7	25	20	30	16	25	63
	Set B	10	30	0	20	20	10	72
Consonant	Set A	10	25	10	20	17.5	15	70.13
	Set B	10	30	0	15	20	7.5	74
Short phones	Set A	10	35	20	40	22.5	30	54.25
	Set B	10	40	10	35	25	22.5	58.13

Table V shows the consolidated results of all experiments. It includes the best results according to Figures 7 to 12. Table V indicates that Set A features are very competitive with the popular MFCC features for all detection categories. To explain the results and the success rate in the table the following equations are considered.

$$e_d = \frac{\delta_{d_{High}} - \delta_{d_{Low}}}{2} \quad (6)$$

$$e_p = \frac{\delta_{p_{High}} - \delta_{p_{Low}}}{2} \quad (7)$$

$$SR\% = SR_d \times SR_p = (100 - e_d) \times (100 - e_p) \quad (8)$$

The way of measure in this research is very tight. Not only the boundary location is considered but also the duration of the detect segment. The success rate is considered as the success rate of both position and duration. Most of the researches paperare considering only the boundary location of each phone which in turn may get little bit higher results. For example if the boundary only is considered the results will be related to only position deviation error rate which in this case can be obtained from the equation $SR_p = (100 - e_p)$. As shown in Table VI this will give results with higher level of success ranges from 70% in the short duration phones up to 92.5% in case of consonants.

TABLE XII
CONSOLIDATED RESULTS ON BOUNDARY DEVIATION

Category	**Features set**	SR_p **%**
Vowels	Set A	75
	Set B	90
Consonant	Set A	85
	Set B	92.5
Short phones	Set A	70
	Set B	77.5

Analyzing the obtained results, it is clear that the consonant boundary detection gives the best results. The short phonemes give the worst results. This is expected for the short phonemes as the duration is too small. Due to the short duration nature of short phonemes, any boundary deviation will be evaluated as big error. The consonants are very different from the short phonemes with respect to the duration point of view. They have the biggest duration, so that the boundary deviation will be the least in all categories.

9 CONCLUSION

It is indicated that through this research paper the automatic segmentation problem can be altered using some hybrid techniques that are related to spectrum analysis and statistical model.

Spectrum analysis is inherited from the wavelet packet features BTE or MFCC and the statistical technique is based on the HMM. In this work comparison between MFCC and BTE in solving the segmentation problem is implemented. Very well manually segmented database called Ked-TIMIT is used to get accurate results. The results in this work indicated that the consonants boundary detection is the best over short phonemes detection or vowels detection. Considering very harsh metrics that held both boundary positions and phoneme duration to calculate the success rate, the consonants average success rate of 74% is achieved. This is achieved using MFCC and HMM with Tri duration method as explained in this paper. This result will be 92.5% if only boundary deviation error is considered as metric. BTE gives a success rate for consonants boundary detection as such 70.13%. This is almost 94.77% as relative to MFCC success rate of the same class.

The results can be enhanced by updating the key parameters of the hybrid model. The key parameters are the features and the HMM model. All odd cases in Figures 7 to 12 that give odd results can be considered for special update to their HMM model. By analysing the database for Tri-Phone statistics, the HMM models can be updated for best frequent trip models by tying the states of rarely appearing phones. The silent and speech pauses can be also modified for better HMM models. This model modification can enhance both hybrid models that depend on BTE or MFCC. It will be also considered to modify BTE itself in order to add more information for discriminating the different categories of phones that are indicated into this research.

10 REFERENCES

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

[2] S. Zhao, Y. Soon, S. N. Koh, and K. K. Luke, "A Hybrid Refinement Scheme for Intra-and Cross-Corpora Phonetic Segmentation,"ComputerSpeech and Language, Elsevier, vol. 29, no. 1, pp. 81–97, 2015.

[3] J. Yuan, N. Ryant, and M. Liberman, "Automatic Phonetic Segmentation in Mandarin Chinese: Boundary Models, Glottal Features and Tone", Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on, pp. 2539–2543, 2014.

[4] J. Yuan, N. Ryant, M. Liberman, A. Stolcke, V. Mitra, and W. Wang, "Automatic Phonetic Segmentation using Boundary Models", Interspeech, pp. 2306–2310, 2013.

[5] O. Kalinli, "Automatic Phoneme Segmentation Using Auditory Attention Features", Interspeech conference, 2012.

[6] Y. Qiao, D. Luo, and N. Minematsu, "A study on unsupervised phoneme segmentation and its application to automatic evaluation of shadowed utterances", Technical report, 2012.

[7] S. Hoffmann and B. Pfister, "Fully Automatic Segmentation for Prosodic Speech Corpora", Interspeech, pp. 1389–1392, 2010.

[8] I. Mporas, T. Ganchev, and N. Fakotakis, "A Hybrid Architecture for Automatic Segmentation of Speech Waveforms", in Acoustics, Speech and Signal Processing, ICASSP 2008. IEEE International Conference, pp. 4457–4460, 2008.

[9] P. Lanchantin, A. C. Morris, X. Rodet, and C. Veaux, "Automatic Phoneme Segmentation with Relaxed Textual Constraints", in LREC, 2008.

[10] J. Adell, A. Bonafonte, J. A. Gómez, and M. J. Castro, "Comparative Study of Automatic Phone Segmentation Methods for TTS", in IEEE International Conference onAcoustics, Speech, and Signal Processing(ICASSP'05), pp. 309–312, 2005.

[11] H. Kawai and T. Toda, "An Evaluation of Automatic Phone Segmentation for Concatenative Speech Synthesis", in IEEE International Conference on Acoustics, Speech, and Signal Processing(ICASSP'04), vol. 1, pp. I–677, 2004.

[12] A. Lazaridis, I. Mporas, T. Ganchev, G. Kokkinakis, and N. Fakotakis, "Improving Phone Duration Modelling using Support Vector Regression Fusion", Speech Communication, vol. 53, no. 1, pp. 85–97, 2011.

[13] Amr M. Gody, Rania Ahmed AbulSeoud, Mohamed Hassan, "Automatic Speech Annotation Using HMM Based on Best Tree Encoding (BTE) Feature", The Eleventh Conference on Language Engineering, Ain-Shams University, pp. 153-159, December 2011.

[14] Amr M. Gody, Rania Ahmed AboulSoud, Mai Ezz El-Din, "Using Mel-Mapped Best Tree Encoding for Baseline-Context-Independent-Mono-Phone Automatic Speech Recognition", The Egyptian Journal of Language Engineering, Vol. 2, No. 1, pp. 10-24, April 2015.

[15] Amr M. Gody, Rania Ahmed AbulSeoud, Marian M. Ibraheem , "Hybrid Model design for Baseline-Context-Independent-Mono-Phone Automatic Speech Recognition, "International Journal of Engineering Trends and Technology (IJETT) – Volume 27 Issue :2231-5381- September 2015.
[16] Bala, Anjali, Abhijeet Kumar, and Nidhika Birla., "Voice Command Recognition System Based on MFCC and DTW", International Journal of Engineering Science and Technology 2.12 (2010): 7335-7342.
[17] Al-Qatab, Bassam AQ, and Raja N. Ainon. "Arabic Speech Recognition using Hidden Markov Model Toolkit (HTK)", Information Technology (ITSim), 2010 International Symposium in. Vol. 2. IEEE, 2010.
[18] Resch, Barbara, "Automatic Speech Recognition with HTK.", Signal Processing and Speech Communication Laboratory. Inffeldgase. Austria. Available on Internet: http://www. igi. tugraz. at/lehre/CI (2003).

BIOGRAPHY

Amr M. Godyreceived the B.Sc. M.Sc., and PhD. from the Faculty of Engineering, Cairo University, Egypt, in 1991, 1995 and 1999 respectively. He joined the teaching staff of the Electrical Engineering Department, Faculty of Engineering, Fayoum University, Egypt in 1994. He is author and co-author of about 40 papers in national and international conference proceedings and journals including Springer {INTERNATIONAL JOURNAL OF SPEECH TECHNOLOGY}. He is the Acting chief of Electrical Engineering department, Fayoum University in 2010, 2012, 2013, 2014 and 2016. His current research areas of interest include speech processing, speech recognition and speech compression.

ManalShabaan received the B.Sc. degree in Electrical Engineering – Communications and Electronics Department with very good honor degree, from the Faculty of Engineering - Fayoum University in 2012. She joined theM.Sc program in Fayoum University - Communications and Electronics Department in 2013 .She received the Pre-Master degree fromFayoum University with very good degree, in 2014. Her areas of interest include Automatic Speech Segmentation.

AmrSalehreceived the B.Sc., M.Sc., and Ph.D. from the Faculty of Engineering, Cairo University, Egypt, in 1996, 2002 and 2007 respectively. He had joined the University of Siegen, Germany through a DAAD channel scholarship (2004-2006) where he had performed the laboratory experimental work of his Ph.D. He joined the teaching staff of the Electrical Engineering Department, Faculty of Engineering, Fayoum University, Egypt in 1998. His current research areas of interest include electrical machines and drives, automatic control, embedded systems and renewable energy.

التقطيع الاوتوماتيكي للصوت باستخدام نموذج مهجن

عمرو جودي، منال شعبان ،عمرو عبد الله
قسم الهندسه الكهربية، كليةالهندسة، جامعةالفيوم، مصر

الملخص
تجزئة الصوت الهستمروفقا للتحويل الصوتي هى مهمة أساسية في أي نظام صوت.التجزئة اليدوية شاقة تستغرق وقتا طويلا وعرضة للخطأ . وعلاوة على ذلك، فإنه يكاد يكون من المستحيل إعادة إنتاج التجزئة اليدويه بسبب كبر حجم قاعده البيانات. هذا البحث يتم استخدام خصائص جديده لتقطيع الكلام اوتوماتيكيا وهو النموذج المهجن ويتكون من تحليل حزمة المويجات وميل النطاق . ويتم تنفيذه باستخدام أدوات نموذج ماركوف المخفي .وتستخدم قاعده البيانات (Ked-TIMIT) للتحقق من النتائج المحققة عن طريق النموذج المقترح . تستخدم MFCC كمرجع لتقييم نتائج النموذ ج المقترح وتصنف النتائج الى الحروف المتحركه، الساكنة والحروف القصيرة. وتستخدم مدة الحرف وموقع بدايته مقاييس لتقييم نسبة نجاح النظام . ويتم تحقيق نسبة نجاح ٧٤٪ للكشف عن ساكن، ٧٢٪ للكشف عن متحرك و ٥٨٪ للكشف عن حرف قصير. وعن طريق استخدام مقياس بسيط هو أن تعتمد فقط على المواقع الحدودية ولكن رتجاهل المدة، تحققت النتائج ا لآتيه هي ٩٢.٥٪ للكشف عن ساكن، و ٩٠٪ للكشف عن متحرك و٧٧.٥٪ للكشف عن صوت قصير . بالإضافة الكشف عن طريق الحدود يستخدم النموذج المهجن المقترح لمقارنة الميزات المطورة حديثا يسمى (Mel-BTE) ليحقق نتائج مقارنه بال MFCC وهي ٩٤.٧٧٪ للكشف عن ساكن، ٨٧.٥٪ للكشف عن علة و٩٣.٣٣٪ للكشف عن صوت قصير.