Volume 2 No.10, October 2012 ISSN 2224-3577
International Journal of Science and Technology

http://www.ejournalofsciences.org

Evaluation of Different Feature Extraction Techniques for Continuous Speech Recognition

[1]Hamdy K. Elminir, [2]Mohamed Abu ElSoud, [3]L. M. Abou El-Maged
[1]Misr Academy for Engineering & Technology
[2,3]Computer Science Department , Faculty of Computer Science & Information System, Mansoura University

ABSTRACT

Extracting human's voice feature is the most important process in any speech recognition system. There are many feature extraction techniques which are already used such as MFCC, LPC and ZCPA; but still have some problems especially in the continuous speech.

It is important to evaluate different feature extraction techniques for continuous speech by making a comparison between these techniques as a trial to find the most suitable technique for speech recognition process, and trying to enhance the result by using PCA.

Using PCA gives great better results especially for ZCPA technique as a comparison to other techniques.

Keywords: *Mel-frequency cepstral coefficients (MFCC); Linear predictive coding (LPC); Zero Crossings with Peak Amplitudes (ZCPA); Hidden Markov Model(HMM);principal component analysis(PCA)*

1. INTRODUCTION

The feature extraction process is considered the most important phase in any speech recognition system, as its majority mission is to catch the features that may help the system to differentiate between utterances. There are some obstacles may be faced during the feature extraction process, these obstacles may be emerged from Variability from speakers: because of illness or emotion. Also, there is variability due to dialect foreign accent. Or; Variability from environments: This is because of background noise, reverberation, microphones, and transmission channels.

The process of sound producing is an acoustic filtering operation where Larynx and lungs provide input or source excitation and Vocal and nasal tracts act as filter, this leads us to identify two main features of the human's voice.

The main features of the human's voice are pitch and formants; a person's pitch originates in the vocal cords/folds, and the rate at which the vocal folds vibrate is the frequency of the pitch [1].

When air flows through the laryngeal tract, the air vibrates at the pitch frequency formed by the laryngeal tract as mentioned above. Then the air flows through the supralaryngeal tract, which begins to reverberate at particular frequencies determined by the diameter and length of the cavities in the supralaryngeal tract. These reverberations are called "resonances" or "formant frequencies" [1].

So the main task for the extracting features is to simulate the humanity audrity system [2,3].

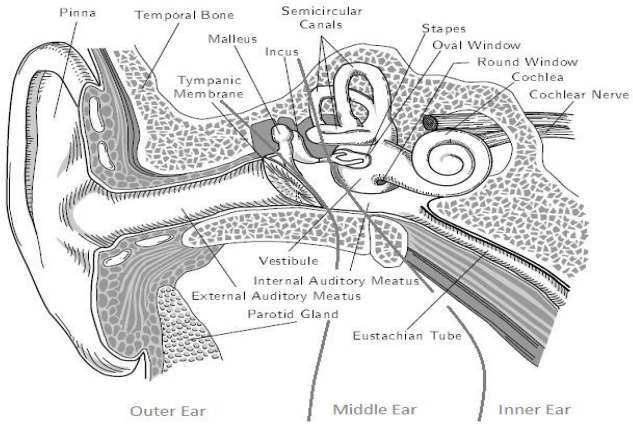


Figure (1) The human hearing system [3].

The ability of the human auditory system to perceive speech under adverse conditions has motivated researchers to include properties of human perception into the speech processing, which have contributed significantly to robustness of ASR under various types of noisy environments. Two important properties of human perception are the nonlinear frequency resolution of the basilar membrane (BM) and the saturating and compressive behavior of the inner hair cells to a wide range of speech stimulus. Computational auditory models replicate the psychoacoustic behaviors of the inner ear based on these perceptual features, transforming mechanical filtering and vibrations into neural representation. These models when used as a front-end ASR processor extract the essential speech features from the simulated probabilistic firing pattern of the auditory nerves. However, the ASR

performance of such models severely degrades under noisy conditions. It is well known that certain properties of human perception are invariant or less affected by additive and convolutive noise. On the other hand, some of these perceptual properties are related to loss of information such as in the case of masking and adaptation [4].

This paper will describe the extracting feature techniques which try to simulate the human's audrity system in their work, this will done after describing the segmentation of continuous speech in the next section. Section three will describe the experiments done to evaluate the feature extraction techniques then describe the usage of PCA and its effects on the results.

2. PROBLEM FORMULATION

A continuous speech system operates on speech in which words are connected together, i.e. not separated by pauses. Continuous speech is more difficult to handle because of a variety of effects. Such as the difficulty to find the start and end points of words. So; we can summarize the problem solution into the next steps:

2.1 Segment the Continues Speech

Segmentation and classification should account for differences in speaker variability, such as pronunciation duration and regional accent differences, in speaker independent automatic speech recognition (ASR) system. The segmentation divides the feature pattern into segments or pattern, each segment corresponds to a linguistic unit such as a phoneme or a word [5].When selecting the basic unit of acoustic information, we want it to be ***accurate***, ***trainable*** and ***generalizable***.

Basically there are three approaches to speech recognition with respect to the choice of sub-word units namely, word based, phone based and syllable based recognition [6].
Words are good units for small-vocabulary SR – but not a good choice for large-vocabulary continuous SR:

- Each word is treated individually – no data sharing, which implies large amount of training data and storage.
- The recognition vocabulary may consist of words which have never been given in the training data.
- Expensive to model inter-word coarticulation effects.

The alternative unit is a *Phoneme* .Phonemes are more *trainable* (there are only about 50 phonemes in English, for example) and *generalizable* (vocabulary independent) [6,7].

The type of sub-word units employed in a speech recognizer depends on the amount of available training data and the desired model complexity: while recognition systems designed for small vocabulary sizes (< 100 words) typically apply whole word models, systems developed for the recognition of large vocabularies (> 5000 words) often employ smaller sub-word units which may be composed of syllables, phonemes, or phonemes in context. Context-dependent phonemes are also referred to as n-phones. Commonly used sub-word units employed in large vocabulary speech recognition systems are n-phones in the context of one or two adjacent phonemes, so-called triphones or quinphones. Context-dependent phoneme models allow for capturing the varying articulation that a phoneme is subject to when it is realized in different surrounding phonetic contexts (coarticulation)[8].

2.1.1 Segmentation Techniques

The most often used current method is to use constant time segmentation, for example into 25 ms blocks. This methods benefit from simplicity of implementation and the ease of comparing blocks of the same length. Clearly, however, the boundaries of speech elements such as phonemes do not lie on fixed position boundaries; phonemes naturally vary in length both because of their structure and due to speaker variations. Constant segmentation therefore risks losing information about the phonemes. Different sounds may be merged into single blocks and individual phonemes lost completely. A number of approaches have previously been suggested for this task but these utilized features derived from acoustic knowledge of the phonemes. Such methods need to be optimized to particular phoneme data and the performance is often not as good on new speech data [9].

Algorithm 1 to calculate PSD	**Algorithm 2 to calculate ZCR**
-Divide speech signal into windows. - Calculate the square of amplitude of different samples in window. - Add all those squared values of amplitude to find PSD of window - Calculate normalized PSD of window - Repeat until all the windows are finished	- Divide the speech signal into windows - Compare successive samples in the window to find a transition from positive to negative - Mark a transition as zero crossing -Total number of zero crossings form a ZCR of the window - Calculate normalized ZCR - Repeat until all the windows are finished

Several techniques have been used for speech segmentation. *Manual segmentation* was the first: an expert linguist generates the segmentation based on spectrograms, energy curves, intonation and other features such as formant pattern, stress pattern, intonation pattern, rhythm, phoneme duration, rate of speech, zero crossing rate (ZCR), power spectral density (PSD) etc used in speech

analysis[8]. Table (1) illustrates the algorithms to calculate PSD and ZCR.

Formant pattern show a typical shape in the start and the end which provides us with information about specific consonants. PSD and ZCR play a major role in ***Table (1):*** *Algorithms to calculate PSD and ZCR [8]* segmenting phonemes. Spectrogram and spectrograph are very important in identifying PSD and ZCR respectively [8]. This technique possesses the advantage that the linguist experience assures a very good result in the segmentation. However, the costs in time and resources that this manual process carries are the highest and make it only applicable to very specialized studies. The second technique applicable to segmentation comes from automatic speech recognizers. In automatic speech recognition, the *hidden Markov models* (HMM) technique currently gives the best results [9,10]. Upon applying HMM to automatic speech recognition, there exists an implicit segmentation process (model alignment) and, with some modifications to reduce the computational cost of a complete recognition, it is possible to use them stand-alone for speech segmentation. However, the classical methods based on HMM alignment requires the full transcription of the speech input, in other words, a full speech-recognition process is needed [10].
In [9] use The Discrete Wavelet Transform (DWT) to segment the speech into phonemes, while [5] use the zero crossing rate and Frame Energy for the segmentation. In [11] used the pitch mark with DWT for the segmentation.

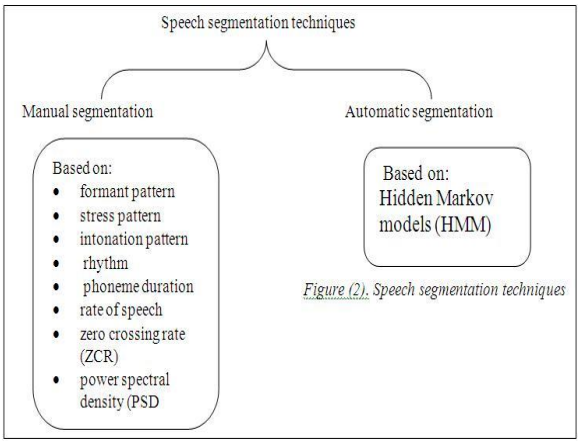


Figure (2). Speech segmentation techniques

2.2 Feature Extraction Techniques

For speech recognition purposes and research, MFCC is widely used for speech parameterization and is accepted as the baseline. It incorporates two perceptual features – the variable bandwidth Mel-spacing of triangular filter banks to simulate the frequency response of the BM, and compressive nonlinearity by taking the logarithm of filter bank amplitudes to mimic the effects of saturation of auditory nerve excitations[4].Figure(3) illustrates the stages of MFCC technique.
MFCC acts effectively in ideal operating conditions. However, it is well established that their performance degrades severely when there is a mismatch between the training and testing conditions, typically due to background noise [12,13].

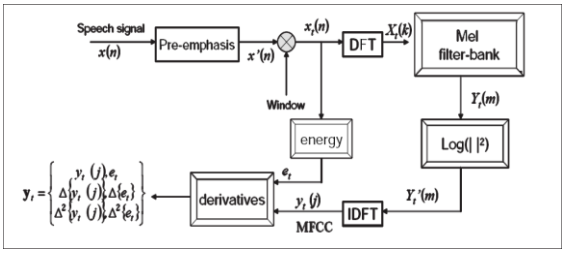


Figure (3) MFCC Stages [14]

Another popular technique is Linear predictive coding (LPC) which offers a powerful, yet simple method to provide exactly this type of information. Basically, the LPC algorithm produces a vector of coefficients that represent a smooth spectral envelope of the DFT magnitude of a temporal input signal. These coefficients are found by modeling each temporal sample as a linear combination of the previous p samples as in the equation

(equ.1)

$$x(n) = \sum_{k=1}^{p} a_k x(n-k) + e(n) = \hat{x}(n) + e(n)$$

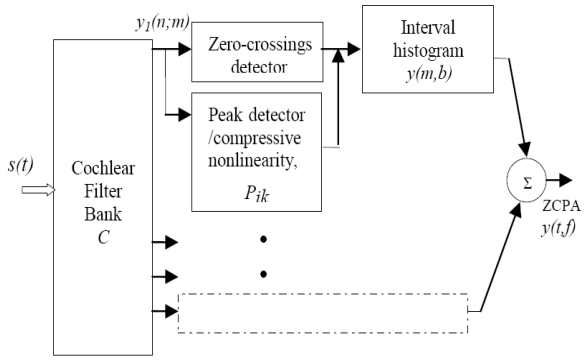
The Speech production process could be generally assumed as the convolution of the excitation E (ω) from the glottis and the all pole transfer function (vocal chords) H (ω) to result in speech, S (ω).

Using the Linear Prediction coefficients alone for the recognition process was not very successful because the all pole assumption of the vocal chord transfer function was not accurate and this method was not efficient enough to separate E (ω) from H (ω)[15].

ZCPA features were first proposed as adaptation of the Ensemble Interval Histogram model. The idea is to model the neural firing patterns of the human cochlea. In the proposed model, the speech signal is filtered with a set of auditory filters, and then the output of each filter is passed through a zero-crossing detector. The distance between adjacent upward going zero-crossings is used to give a frequency estimate. The resulting frequencies are collected in a histogram, with the weight of each histogram entry being given by a non-linear compression of the peak amplitude between the zero-crossings. The histograms across all filter channels are then summed to produce the feature vector [13]. A schematic for the algorithm is shown in Figure(4).

The auditory filter bank aims to simulate the frequency selectivity behavior of the human cochlea. It comprises of multiple channels with bandwidth and spacing determined by some non-linear scale.
For this purpose the Zero Crossings with Peak Amplitudes (ZCPA) method is based on the principle that any stimulus periodicity in the filter subband can be extracted from the

zero crossing intervals, which shows up as a dominant frequency corresponding to the formant peaks. This emphasis on dominant spectral peaks, and less emphasis on the valleys which are usually corrupted by noise, makes the model more robust in the presence of noise.



Figure(4): diagram of ZCPA [16]

The zero-crossing analysis (ZCA) of speech waveforms has advantages over autocorrelation, power spectrum and linear prediction methods. This is because in these methods data extraction by sampling a time waveform depends on the maximum frequency content in the time signal whereas ZCA requires a number of extracted samples determined by the average rate of zero-crossing intervals. ZCA is amenable to simple transformations instead of complex transformations between time and frequency domains. In the ZCPA the reciprocal of time intervals between two successive zero crossings are collected in frequency histograms from which frequency information are extracted (the speech spectral characteristics) . Moreover, the model uses the logarithm of the peak amplitude as a weighting factor to the frequency bins to extract the intensity information [4] .
This research tries to evaluate these techniques as a trial to reach the most suitable techniques for continues speech recognition.

3. EXPERIMENTS

The main target of this stage is to reach to the most suitable feature extraction technique; So we will follow the following steps in order to achieve our target.

3.1 Speech Database

The database was established by recording processing which is done by e-learning unit at Mansoura University, 5 males and 5 females recorded about 30 sentences contain about 90 words(i.e. the database contains 10X30X90 words) .the speech was recorded with sample rate 16khz,16bit per sample ,and with mono channel.

3.2 Segment the Continues Speech

1- Framing the input signal into hamming window frames.
2- Calculate the ZCR and energy for each frame.
3- Calculate ZCR's threshold and energy's threshold.
3- Extracting unvoiced frames and separate words.

3.3 Extract the feature for each word individually

1. MFCC
2. ZCPA
3. LPC

3.4 Feature Reduction using PCA

As a trial to improve performance we use PCA (principal component Analysis) to reduce the dimensionality of a feature vector while retaining as much information as is possible. It computes a compact and optimal description of the data set. Its job can be described shortly as elimination redundancy between dimensions based on correlation, collapsing of the correlated dimensions, and leaving uncorrelated ones intact.

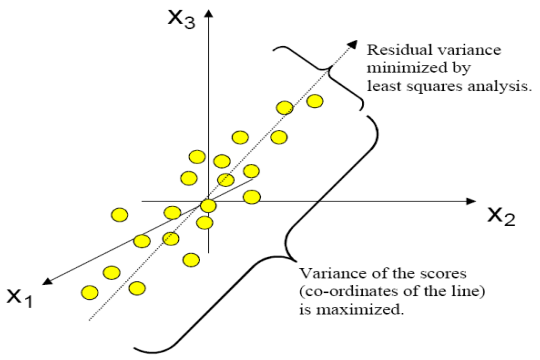


Figure (5) : PCA derives a model that fit data as well as possible[17]

4. RESULTS

4.1 Speech Recognition with Different Feature Extraction Techniques

In our experiments we use CHMM as classifier which give the results in the following table

Table (2): Recognition rates with different Extraction techniques

Feature Extraction technique	*Recogni_tion rate*
MFCC	*85.3*
ZCPA	*38.5*
LPC	*82.3*

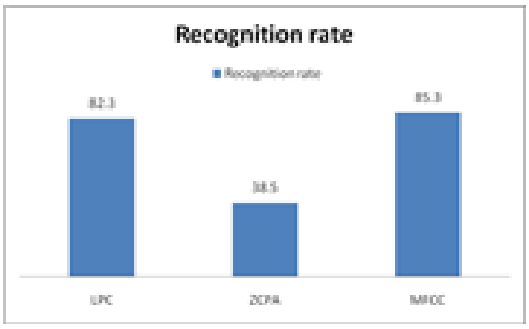


Figure 6: Recognition rates rates with different Extraction techniques

From table (2) and figure (6) we can notice that MFCC gives the best result.

4.2 Speech Recognition using PCA for Feature Vector Reduction with Different Feature Extraction Techniques

After using PCA with different parameters on the feature vector we notice that it affects on the recognition rate as in the following table.

Table (3): Recognition rates after using PCA with different parameters.

PCA parameters / Extraction techniques	6	8	10	12
MFCC	92.3	88.2	87.3	87.3
ZCPA	94	92.31	92.31	90.1
LPC	90	87.3	87.3	86.2

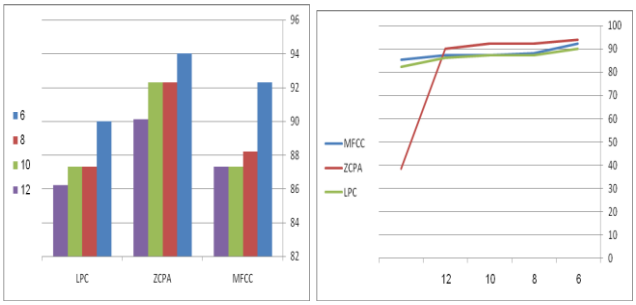


Figure 7: Recognition rates after using PCA Notice the greatest different in the result after using PCA especially with ZCPA feature extraction technique.

4.3 Calculation Feature Extraction Time with Different Feature Extraction Technique

Table (4) and figure (8) illustrate the results where the ZCPA takes the longest time than other techniques.

Feature Extraction technique	Feature Extraction time
MFCC	0.092
LPC	0.152
ZCPA	27.38

Table (4): feature extraction time

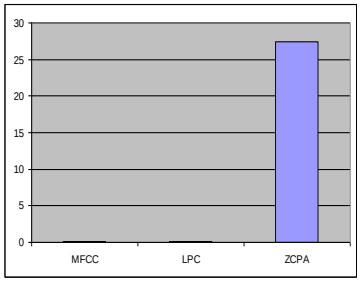


Figure 8: feature extraction time

4.4 Calculation the Training time with Different Feature Extraction Technique

Table (5) and figure (9) illustrate the results where the ZCPA still taking the longest time than other techniques.

Feature Extraction technique	Training Time
MFCC	0.345
LPC	0.213
ZCPA	1.811

Table (5): training time

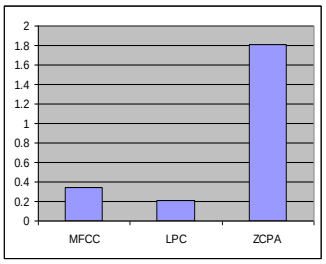


Figure 9: training time

4.5 Calculation the Time of PCA Conversion with Different Feature Extraction Technique

Table (6) and figure (10) illustrate the results where the ZCPA taking the longest time than other techniques that's because of its longest feature vector.

Feature Extraction technique	PCA convers_ ion time
MFCC	0.097
LPC	0.132
ZCPA	0.438

Table (6): PCA conversion time

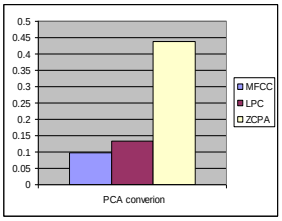


Figure 10: PCA conversion time

4.6 Calculation the Training Time after using PCA with Different Feature Extraction Technique

Table (7) and figure (11) illustrate the results where the ZCPA get back and MFCC is the longest

Table (7): Training time after PCA

Feature Extraction technique	Training time after PCA
MFCC	0.250
LPC	0.144
ZCPA	0.135

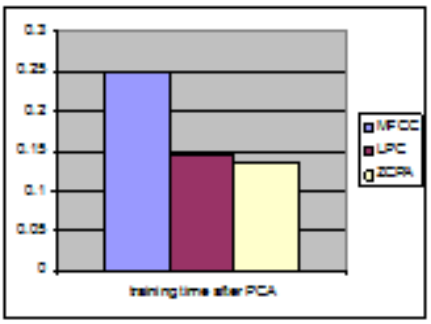


Figure 11: Training time after PCA

All previous results say that; when using CHMM as a classifier, MFCC gives a good result in the smallest time in comparison with LPC and ZCPA. When using PCA, the results are getting better than before and the time of recognition step is come down especially for ZCPA ,but the time of feature extraction still long when using ZCPA technique .

5. CONCLUSION

The speech recognition system tries to simulate the human's hearing system in order to get optimal result. The most important step in the recognition process is to extract the voice's feature, so; there are many feature extraction techniques are used. When using CHMM as a classifier we noticed that MFCC gives a good result in the smallest time in comparison with LPC and ZCPA. That's because MFCC is follows the human's hearing system in its work, but ZCPA takes along time than MFCC and gives bad results; this may be because of its long feature vector with redundant variables.

PCA is a famous technique used for data reduction, it is used in this work as trial to get better results, when it is used the results get better than before and the time of recognition step is come down especially for ZCPA ,but the time of feature extraction still long when using ZCPA technique .

The reason of the great transformation is the usage of PCA; which eliminate redundancy between dimensions based on correlation, collapse correlated dimensions, and leaving uncorrelated ones intact. This gives a feature vector with high variance variable.

The effect of using PCA gives different results according to the feature technique. In LPC; the effects shown to be quite better because the feature vector of LPC is consists of high correlated coefficients, these coefficients are composed of a linear combination of the previous p samples. In MFCC the results of the FFT will be information about the amount of energy at each frequency band and then complete the MFCC calculation step, and the energy calculation depends on the computation between features in window as shown in [18], so; the effects of using PCA on MFCC is also quite better, while in ZCPA; the using of PCA give a great better results this due to the calculation of ZCPA coefficients are uncorrelated.

6. FUTURE WORK

This work is applied on calm (not noisy) environment; the future work is to work in noisy environment.

REFERENCES

[1] K.R. Aida–Zade, C. Ardil and S.S. Rustamov , " Investigation of Combined use of MFCC and LPC Features in Speech Recognition Systems ", World Academy of Science, Engineering and Technology, 2006 .

[2] Om D. Deshmukh, " Synergy of acoustic-phonetics and auditory modeling towards robust speech recognition ", Doctor of Philosophy, Faculty of the Graduate School of the University of Maryland, College Park, 2006 .

[3] Cees-Jeroen Bes, "a front-end for sensing the stimulation and response of auditory nerve cells", master thesis Department of Microelectronics, Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, 2010.

[4] Serajul Haque, Roberto Togneri and, Anthony Zaknich," Zero Crossings with Peak Amplitudes and Perceptual Features for Robust Speech Recognition", http://www.ee.uwa.edu.au/~roberto/research/theses/tr06-01.pdf , March 2012.

[5] Ying Cui ," Recognition of Phonemes In a Continuous Speech Stream By Means of PARCOR Parameters In LPC Vocoder ",master thesis, College of Graduate Studies and Research, Department of Electrical & Computer Engineering University of Saskatchewan ,2007.

[6] R. THANGARAJAN , A.M. NATARAJAN and M. SELVAM ," Word and Triphone Based Approaches in Continuous Speech Recognition for Tamil Language",WSEAS TRANSACTIONS on SIGNAL PROCESSING , ISSN: 1790-5022, Issue 3, Volume 4, March 2008.

[7] http://www.liacs.nl/~erwin/SR2003/Students/10_SR_Triphones.ppt , Jan. 2012.

[8] Muhammad Jamil Anwar, M.M.Awais, Shahid Masud, and Shafay Shamail ," Automatic Arabic Speech Segmentation System ",International Journal of Information Technology Vol. 12 No.6 2006.

[9] Bartosz Ziołko, Suresh Manandhar and Richard C. Wilson ," Phoneme segmentation of speech ", http://www-users.cs.york.ac.uk/~suresh/papers/PSOS.pdf ,march ,2012.

[10] D. H. Milone , J. J. Merelo and H. L. Rufiner," Evolutionary Algorithm for Speech Segmentation ", IEEE WCCI,2002.

[11] I. Mporas, P. Zervas and N. Fakotakis ," Automatic Segmentation of Greek Speech Signals to Broad Phonemic Classes ", Wire Communications Laboratory, Electrical and Computer Engineering Department, University of Patras Rion Patras, 261 10 Greece, 2005.

[12] Finnian Kelly and Naomi Harte ," Auditory Features Revisited for Robust Speech Recognition ", 2010 International Conference on Pattern Recognition, 1051-4651/10 © 2010 IEEE .

[13] Finnian Kelly and Naomi Harte," A COMPARISON OF AUDITORY FEATURES FOR ROBUST SPEECH RECOGNITION ", 18th European Signal Processing Conference, August 23-27, 2010 © EURASIP, 2010 ISSN 2076-1465

[14] Daniel Jurafsky and James H. Martin, Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition (2ed.), Prentice Hall, 2008.

[15] W.B. Mikhael, and Pravinkumar Premakanthan, "Speaker Verification/Recognition and the Importance of Selective Feature Extraction: Review", 44th IEEE Proceedings on Midwest Symposium on Circuits and Systems, Ohio, 2001, Volume: 1, Page(s): 57 –61.

[16] Serajul Haque, Roberto Togneri and Anthony Zaknich,"Zero-Crossings with Adaptation for Automatic Speech Recognition", Proceedings of the 11th Australian International Conference on Speech Science & Technology, December 6-8, 2006. Copyright, Australian Speech Science & Technology Association Inc.

[17] http://www.umetrics.com/Content/Document%20Library/Files/Multimega_PartI-3.pdf , john, 2012.

[18] D. Torre Toledano, M. A. Rodríguez Crespo, J. G. Escalada Sardina," TRYING TO MIMIC HUMAN SEGMENTATION OF SPEECH USING HMM AND FUZZY LOGIC POST-CORRECTION RULES", the third ESCA/COCOSDN workshop (ETRW)on speech synthesis ,Australia ,November 20-29,1998