مجلـة الـتربـوي
Journal of Educational
ISSN: 2011- 421X
Arcif Q3
معامل التأثير العربي 1.63
العدد 23

مجلة التربوي

مجلة علمية محكمة تصدر عن كلية التربية

جامعة المرقب

العدد الثالث والعشرون
يوليو 2023م

- المجلة ترحب بما يرد عليها من أبحاث وعلى استعداد لنشرها بعد التحكيم .
- المجلة تحترم كل الاحترام آراء المحكمين وتعمل بمقتضاها .
- كافة الآراء والأفكار المنشورة تعبر عن آراء أصحابها ولا تتحمل المجلة تبعاتها .
- يتحمل الباحث مسؤولية الأمانة العلمية وهو المسؤول عما ينشر له .
- البحوث المقدمة للنشر لا ترد لأصحابها نشرت أو لم تنشر .

مجلة التربوي
Journal of Educational
ISSN: 2011- 421X
Arcif Q3

معامل التأثير العربي 1.63
العدد 23

ضوابط النشر:

يشترط في البحوث العلمية المقدمة للنشر أن يراعى فيها ما يأتي :

- أصول البحث العلمي وقواعده .
- ألا تكون المادة العلمية قد سبق نشرها أو كانت جزءا من رسالة علمية .
- يرفق بالبحث تزكية لغوية وفق أنموذج معد .
- تعدل البحوث المقبولة وتصحح وفق ما يراه المحكمون .
- التزام الباحث بالضوابط التي وضعتها المجلة من عدد الصفحات ، ونوع الخط ورقمه ، والفترات الزمنية الممنوحة للتعديل ، وما يستجد من ضوابط تضعها المجلة مستقبلا .

تنبيهات :

- للمجلة الحق في تعديل البحث أو طلب تعديله أو رفضه .
- يخضع البحث في النشر لأولويات المجلة وسياستها .
- البحوث المنشورة تعبر عن وجهة نظر أصحابها ، ولا تعبر عن وجهة نظر المجلة .

Information for authors

1- Authors of the articles being accepted are required to respect the regulations and the rules of the scientific research.
2- The research articles or manuscripts should be original and have not been published previously. Materials that are currently being considered by another journal or is a part of scientific dissertation are requested not to be submitted.
3- The research articles should be approved by a linguistic reviewer.
4- All research articles in the journal undergo rigorous peer review based on initial editor screening.
5- All authors are requested to follow the regulations of publication in the template paper prepared by the editorial board of the journal.

Attention

1- The editor reserves the right to make any necessary changes in the papers, or request the author to do so, or reject the paper submitted.
2- The research articles undergo to the policy of the editorial board regarding the priority of publication.
3- The published articles represent only the authors' viewpoints.

مجـلة الـتربـوي
Journal of Educational
ISSN: 2011- 421X
Arcif Q3

معامل التأثير العربي 1.63
العدد 23

Speaker recognition from speech using Gaussian mixture model (GMM) and (MFCC)

Hanan A. Algrbaa,
Computer. Dept. Academe Tertiary for Sciences and Technique ,Kasr Akiear

Abstract :This research presents a comprehensive presentation of speaker recognition technology, beginning with the basics of self-identification, extracting some features from the voice, models used, updates and current developments, and identifying methods for the Speaker Recognition System 'SRS'. first we extracted features from the speech signal and then we give them to the statistical model . This study We use GMM as statistical model to create a unique voice print for each identity

Keywords (Speaker recognition , Feature extraction , Gaussian mixture model [GMM], Mel Frequency Cepstral Coefficients[MFCC].

Introduction:

Identification of the speaker: It is the process of identifying a person through his voice, as there are no two people with the same voice, and this is due to the difference in the vocal apparatus responsible for producing or issuing the sound from one person to another, and the size of the larynx, in addition to that each voice has what distinguishes it by it or it, with a specific dialect, and a special rhythm or intonation.

In the recognition system, a number of these features of each person are extracted and an attempt is made to reach these differences as a way to identify the person more accurately. SRS is to convert the acoustic voice signal into a computer-readable format and to identify the speakers depending upon their vocal characteristics [1].

What is a speech recognition system?

Speaker recognition is the process of automatically recognizing who is speaking on the basis of individual Information included in speech waves.

Problem Definition and Applications Speaker recognition involves two stages: identification and verification,

In identification, the goal is to determine which voice in a known group of voices best matches the speaker. In verification, the goal is to determine if the speaker is who he or she. Recent studies were focused on the automatic speech recognition (ASR) because of its importance in many fields as banking, security, forensics , remote access to computers etc. Speaker recognition is the process of automatically recognizing who is speaking on the basis of individual Information included in speech waves. It is involves two types: Speaker identification and Speaker verification. Speaker identification is the process of determining which registered speaker provides a given utterance, It is determine which voice in a known group of voices best matches the speaker. Speaker verification is the process of accepting or rejecting the identity claim of a speaker, It is determine whether the person speaking is the same person he/she.[18].This study search in Speaker recognition (identification).The process recognition can be divided into several stages: features extracting, Features Selection, and Classification.

The first stage Features Extracting: this stage is important to recognize, it is the computation of a sequence of feature vectors which provides a compact representation of the given speech signal. This stage will be read in the following section.

the MFCC is the most common important features fore speaker recognition and it gives of accuracy recognition, it was used in many previous studies it was perfect, there for it was selected and it has been improved by apply Ifft instead of DCT .The MFCC were extracted from database, for 10 speakers whit 196 sentences, using matlab program, The number of features was 13 features.

Features Selection stage: The number the extracted features should be taken into account after extracting the speech features usually there are huge number of features , are repeated and less effective for recognition speaker, it should not be a big number, statistical models for example as Gaussian mixture model cannot cope with the higher dimensions data. therefore we need to selection or reduction dimension this features.

Dimension Reduction These techniques are typically used while solving machine learning problems to obtain better features for a classification or regression task.

Normalization is the process of scaling individual samples to have unit norm. This process can be useful to use a quadratic form such as the dot-product or any other kernel to quantify the similarity of any pair of samples.
in this stage used different methods to reduce dimension of the features to obtain good classification results. Before apply reduction dimension apply normalization on each data (features). In this study was used three reduction methods: Low Variance Filter, Principal Component Analysis (PCA) and Backward Feature Elimination. whit two normalization methods .So this created six methods for classification.

Classification stage: classification is the process of grouping the patterns,

are sharing the same set of properties. It is Based on the feature extraction a model of the voice is generated and stored in the speaker recognition system. After features selection, the output are input to Training and testing the model . here get model classification (recognition process). [25] machine learning tool used to evaluate the accuracy of the model. there are many models that used in classification as Gaussian Mixture Model (GMM), Support Vector Machines (SVM), Hidden Markov Model (HMM) and K-mean. In this study was used (GMM ,K-mean). Recent researches show that MFCCs are successful in processing the voice signal with high accuracies. MFCCs represents a sequence of voice signal-specific features[30].MFCC, PLP and LPC are the most widely used features in area of speaker processing (Namrata Dave 2013).

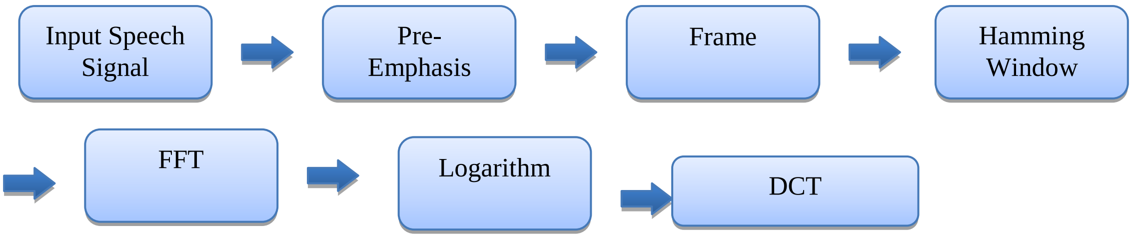
MFCC are chosen for the following reasons:-

1. MFCC are the most important features, which are required among various kinds of speech applications.
2. It gives high accuracy results for clean speech.

مجلة التربوي
Journal of Educational
ISSN: 2011- 421X
Arcif Q3
معامل التأثير العربي 1.63
العدد 23

3. MFCC can be regarded as the "standard" features in speaker as well as speech recognition.

1.The Mel-scale frequency cepstral coefficients (MFCC) extraction is used in front-end processing



MFCCs being considered as frequency domain features are much more accurate than time domain features [9, 10, 11]. MFCC features are based on the short-term analysis and thus from each frame 13 MFCC features are computed (Desai and etal 2013). MFCC method for feature extraction analyses the acoustic features in a speech to determine the Mel coefficients for processing a speech in ASR [20]. MFCCs are coefficients, which represent speaker, based on perception of human auditory systems (Taabish Gulzar and etal2014). MFCC has two types of filter, which are spaced linearly at low frequency below 1000 Hz and logarithmic spacing above 1000Hz [11].The MFCC algorithm is used to extract the features.

Remaining calculation for features extraction is same as for image [13].

1.1 MFCC Method : In the flowing section we will present how calculate the MFCC in six steps as browser.

1- Pre–emphasis

This step processes the passing of signal through a filter which emphasizes higher frequencies. This process will increase the energy of signal at higher frequency is given by eq (1)

$$y(n) = x(n) - ax(n-1) \qquad (1)$$

Where x is the original signal and n is the index of the sample. The range of n is from 1 to the length of signal, y is the pre-emphasis signal obtained by eq (1) as shown in fig2.1, and a is a constant, which has a typical value of 0.95.

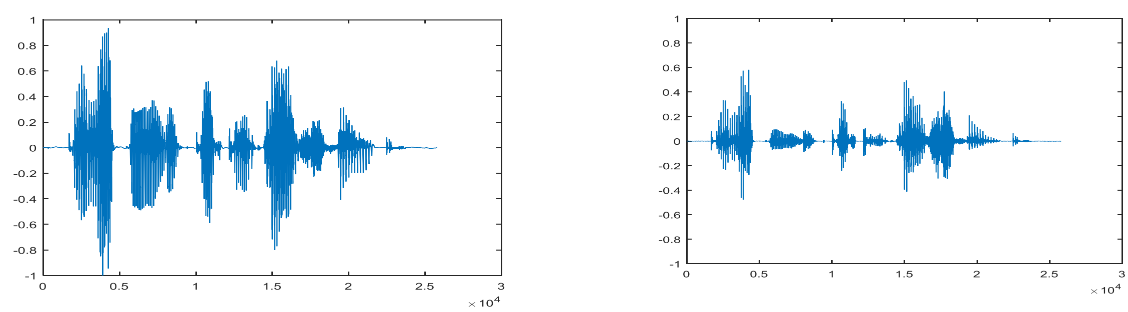


fig.1.1 (a.b) (a) Original signal (b) Pre–emphasis

2- Framing and blocking

In this process, the Pre–emphasis signal is divided into overlapped frames. This process segmenting the wave into small frames. Each frame has a duration from 20 to30 mel second (ms). which contains N samples, the overlap length (M) usually less than N. for example the overlap length (M)=100 when the frame length N =256. If the frame is much shorter, we don't have enough samples to get a reliable spectral estimate, if it is longer the signal changes too much throughout the frame. It is assumed that although the speech signal is non-stationary, but is stationary for a short duration of time.

3- Windowing

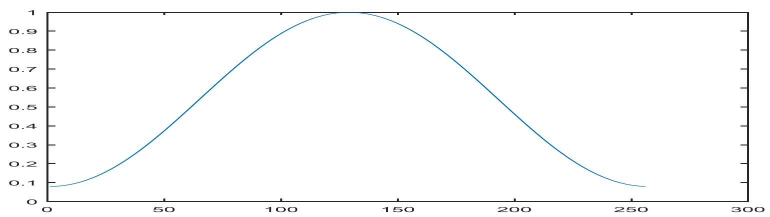
In this step Hamming window is multiplied with each of the above frames, this is done for minimizing the disruptions at the starting and at the end of the frame, the output after windowing the signal will be presented as Y (n) = y(n) × W (n) (2)

Where Y(n) is the frame after windowing process, y(n) is the Pre–emphasis frame, and W(n) represents the Hamming window. Basically many window functions exist such as rectangular window, flat top window and hamming window, However mainly hamming window is applied for carrying out windowing. Hamming usually represented by the following equation:

$$W(n) = 0.54 - 0.45\cos(\frac{2\pi n}{N-1}) \qquad (3)$$

Where 0 ≤ n≤N -1 when n is the sample number, N is the length.

This equation can be represented as shown in fig2.3.a where N=256 sample



(a) Hamming window

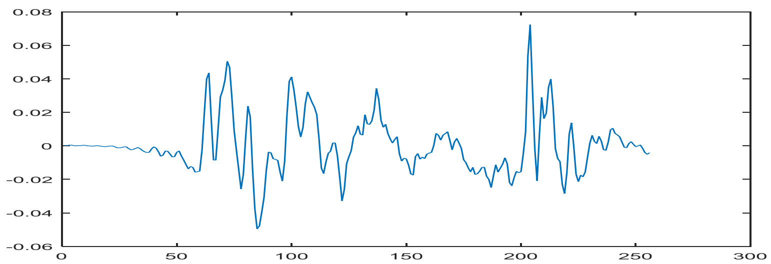


Fig.1.3(a,b) (b)windowing

مجـلة الـتربـوي
Journal of Educational
ISSN: 2011- 421X
Arcif Q3
معامل التأثير العربي 1.63
العدد 23

From the final process as shown in Fig .2.3(b) the starting and ending are smooth.

4- **Fast Fourier Transform (FFT)** : FFT is used to convert each frame of N samples from time domain into Frequency domain. Fourier transformation is a fast algorithm to apply Discrete Fourier Transform (DFT), on the given set of N samples shown below

$$X(k) = \sum_{n=0}^{N-1} Y(n) e^{\frac{-j2\pi kn}{N}} \qquad (4)$$

Where k= 0, 1, 2 ….. N-1

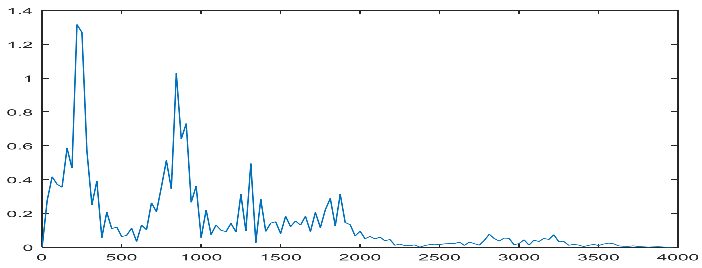


Fig.1.4 (FFT)

Basically the definition for FFT and DFT is same, which means that the Output for the transformation will be the same; however, they differ in their computational complexity. Thus it is in digital processing or other area instead of directly using DFT, FFT is used for applying DFT[13].

5- **Mel scale** The human auditory system doesn't interpret pitch in a linear manner. The sole purpose of the experiment were to describe the human auditory system on a linear scale.The formula to convert frequency f hertz into Mel mf is given by Eq.

$$\mathrm{mf} = 2595\log_{10}\left(\frac{f}{700} + 1\right) \qquad (5)$$

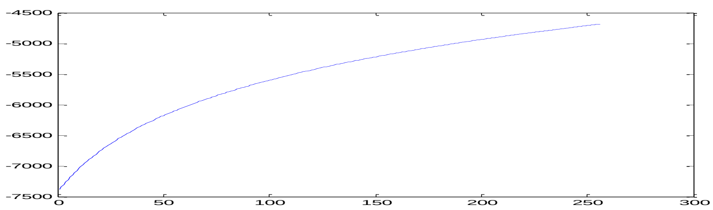


Fig1.5 Mel scale

As shown in Fig1.5 which describe the relationship between the real frequency f in Hz and the mel scale frequency.

6- Calculate Mel Filter bank : The filter bank is a set of overlapping triangular bandpass filter, that according to mel-frequency scale**[5].**

مجلة التربوي
Journal of Educational
ISSN: 2011- 421X
Arcif Q3

معامل التأثير العربي 1.63
العدد 23

- let The FFT size =256. We will use only the half of the FFT size i.e Nfft =128.see fig(1.6)

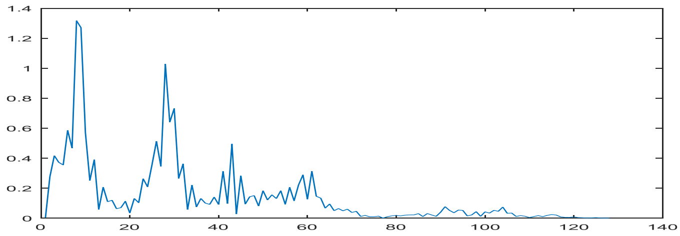
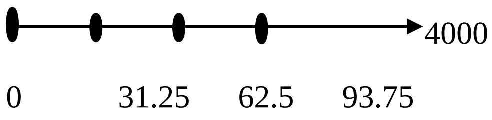
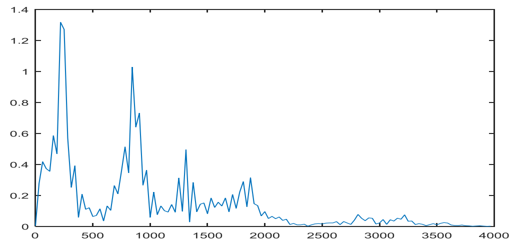


Figure 1.6: The FFT sample Manner value

- Now we will convert the X-Axis from samples to Hz frequency.by multiplying each sample by Δf ,see equation (6) and Fig 1.7(a,b)





(a) (b)

Figure 1.7: The FFT values in Hz manner

$$\Delta f = \frac{4000}{128} = 31.25 \qquad (6)$$

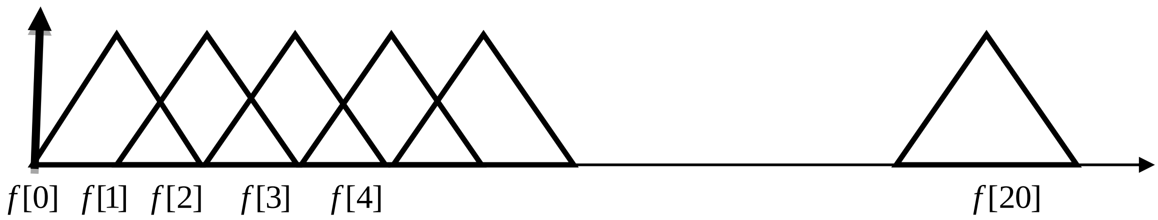
- Now Convert the minimum and maximum from Hz frequency to mel scal frequency by eq (5) for example when the min=0 then:

$$Minmel = 2595\log 10(\frac{0}{700}+1) = 0$$

When the max =4000 then

$$Maxmel = 2595\log 10(\frac{4000}{700}+1) = 2146mel$$

$$f = [0\ 34.96\ \ 69.92\ \ 108.37\ \ 146.82]$$



$$filter2 = [0,0,0,0.78,0.38,0,.........0]$$

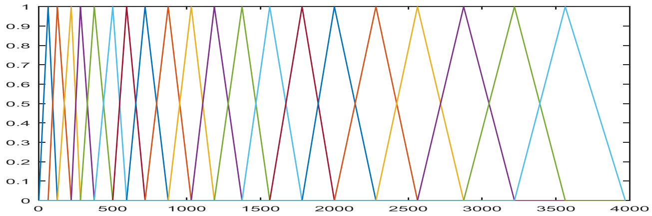


Figure 1.8: -filter bank

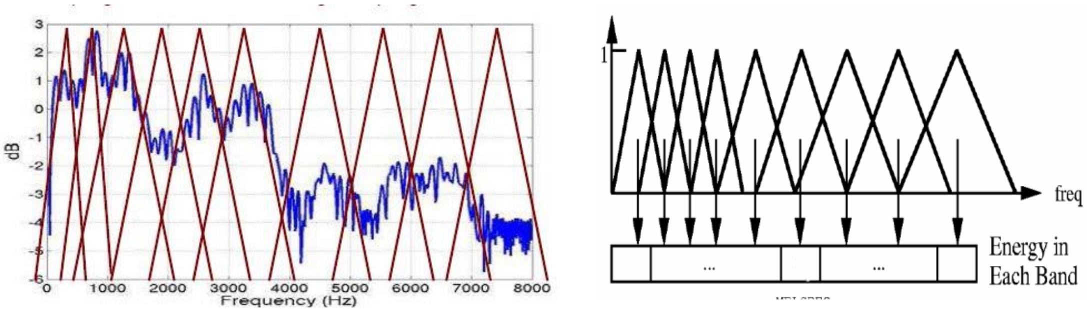


Figure 1.9: -filter bank× FFT

These filters are non-uniformly spaced on the frequency scale, with more filters in the low frequency regions and less filters in the high frequency regions.

-Apply the bank of filters according Mel scale to the spectrum

-Each filter output is the sum of its filtered spectral components

Mel spectrum can be used for calculating first 13 coefficients using DCT. Hence, first 13 coefficients are calculated using DCT and higher are discarded.

6- **Discrete cosine Transform (DCT)**: This allows for better processing of data (Namrata Dave 2013). This is the process to convert The log Mel spectrum into time domain using Discrete Cosine Transform (DCT). Each input utterance is transformed into a sequence of acoustic vector [11].

$$C_n = \sum_{K-1}^{K} (\log mf)\cos[m(K-\frac{1}{2})\frac{\pi}{K}] \quad (5)$$

Where m = 0, 1… k- 1 .Where Cn represents the MFCC and m is the number of the coefficients here m=13 so, total number of coefficients extracted from each frame is 13 (Desai and etal 2013).

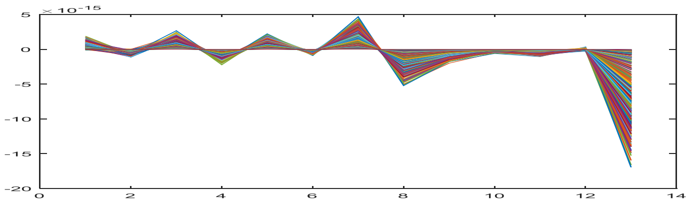


Figure 1.10:MFCC

The MFCC algorithm
Read the wave file (X)
Determine the wav length (N)
Determine the overlapping length (M)
Determine the sample length (n)
K=0 where k is number of frames
1) Pre–emphasis step
$X1(n) = X(n) - 0.95X(n-1)$
2) framing step
For I = 1 to overlapping length to wav length - sample length do
K=K+1
F1(n)=X1(Mn+ N)
End do
3) hamming windowing step
For I = 1: sample length do
W(n)= hamming windowing.*F1(n)
End do
4) Fast Fourier Transform (FFT) step
$f = FFT(W(n))$
5) filter bank step
$ws = call\, fun\, filterbank$
$fet = ws \times f$
6) DCT step
$for\, i = 1 \quad to\, 13\, do$
$MFCC = DCT(fet)$
$end\, do$
Database

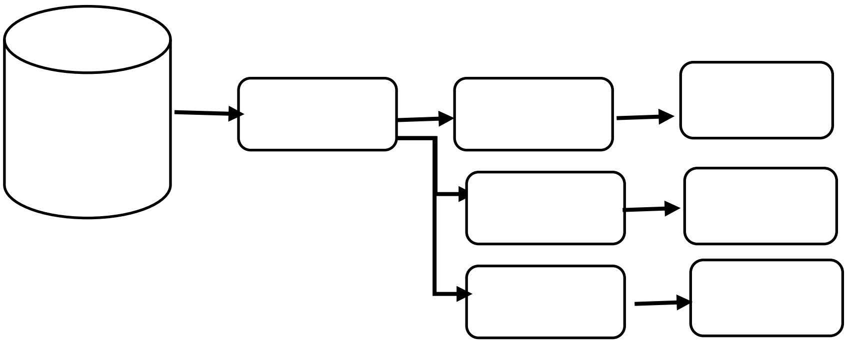


Fig3.13
Each file is framed into number of frames different for each file, we calculated 13 MFCC for each frame.

After extract 13MFCC for each wave. So, total number of coefficients extracted from each frame is 13. There for speech signal contain (number of frames multiplied number of MFCC).

After extract the selected 13MFCC for each frame for each wave. Calculate the average of each MFCC Coefficient so we have 13 values per signal. For example if the signal contains 200 frames then we have 200 multiplied 13 features. The arithmetic average is not enough because it does not reflect the specific values of each speaker.

In this work we proposed adding some statistics measures to extend the number of features. Each dimension with seven statistical characteristics. These statistics are average (avg), the standard deviation (std), variance (var), maximum (max), minimum (min), rang and Medium (mid) of MFCC Per coefficient of MFCC. Therefore we obtained 91 dimension vector as parameters of speech features for each wave. The feature vector of 91dimension, consisting of the following Coordinates From 1to 13are the average,14 to 26 are the standard deviation, 27 to39 are variance, 39 to 52 are maximum, 52 to 65 are minimum,65 to 78 are rang and from 78 to 91 are Medium.

2 Gaussian mixture models

GMMs are often used to generalize models from sparse data [6]. It is considered as one of the methods for clustering that helps in building soft clustering boundaries [21]. The model has two main applications. The first one is to use it as initiating data when creating models of particular speakers[29]. In traditional speaker recognition system usually use 13 MFCC features which extracted for each frame and then the calculate arithmetic mean for each frame and after use GMM classification model for recognition speaker.

The paper is organized as follow. Introduction are presented in section 1, Used Database is introduced in section2, proposed system in section 3, .section 4 present Results, and Discussion Conclusion section5. The proposed method will be explained in more details in following section.

3. The Database

The first step to build a speaker recognition system is to select or create database. The database must include record by different utterance speakers. In this study the Berlin database is chosen. This database was

recorded at the Technical University of Berlin [24]. The Berlin database is widely used in speaker recognition and emotional speech recognition. It contains about 500 utterances spoken. Ten professional German actors (five female and five male) Five of the ten sentences consisted of one phrase, the other five consisted of two phrases, which include the ten actors (5 female and 5 male) producing 10 German utterances (5 short and 5 longer sentences). As the recordings were intended for phonetic analysis of emotions and emotional speech synthesis they were conducted under very controlled conditions and so are marked by a very high audio quality. 200 utterances where selected. Table (1) show the information about the distribution of sentences for each speaker form berlin database. Where S1,S2,…,S10 represent speakers, a01,a02,a04,a05,a07 represent the five short utterance ,b01,b02,b03,b09,b10 represent the five long utterance.

Table (1) the distribution of sentences and speakers of the used Database

3. **The proposed system**

This section present the proposed system. The main parts of the system are feature extraction, classification. These parts will explained in more detail in the following subsection. Figure (3) descried the proposed system.

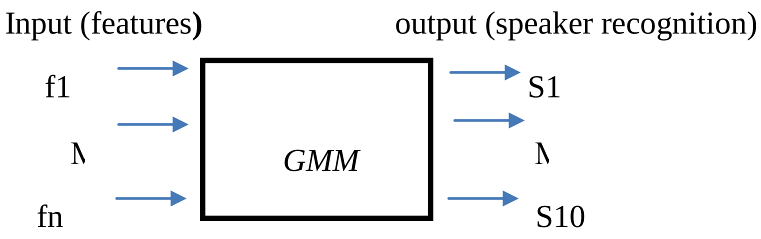


Fig (3)

The flowchart in figure (4) is partitioned into 5 parts. This flowchart introduced the system partitions in very clear ways.

Speaker / Utterance	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	Total
a01	3	3	1	2	2	3	3	1	3	3	24
a02	3	2	-	3	4	2	1	2	3	2	22
a04	4	3	1	2	3	1	2	1	3	2	22
a05	2	2	3	1	3	1	2	3	1	3	21
a07	4	4	3	2	2	1	2	2	3	3	26
b01	2	2	2	1	1	1	2	3	-	3	17
b02	3	2	1	1	4	3	2	2	3	-	21
b03	1	2	2	1	1	1	2	1	1	2	14
b09	1	2	1	-	2	1	1	1	2	1	12
b10	1	5	1	1	3	1	3	1	3	2	21
Total	24	27	15	14	25	15	20	17	22	21	200

مجـلة الـتربـوي
Journal of Educational
ISSN: 2011- 421X
Arcif Q3

معامل التأثير العربي 1.63
العدد 23

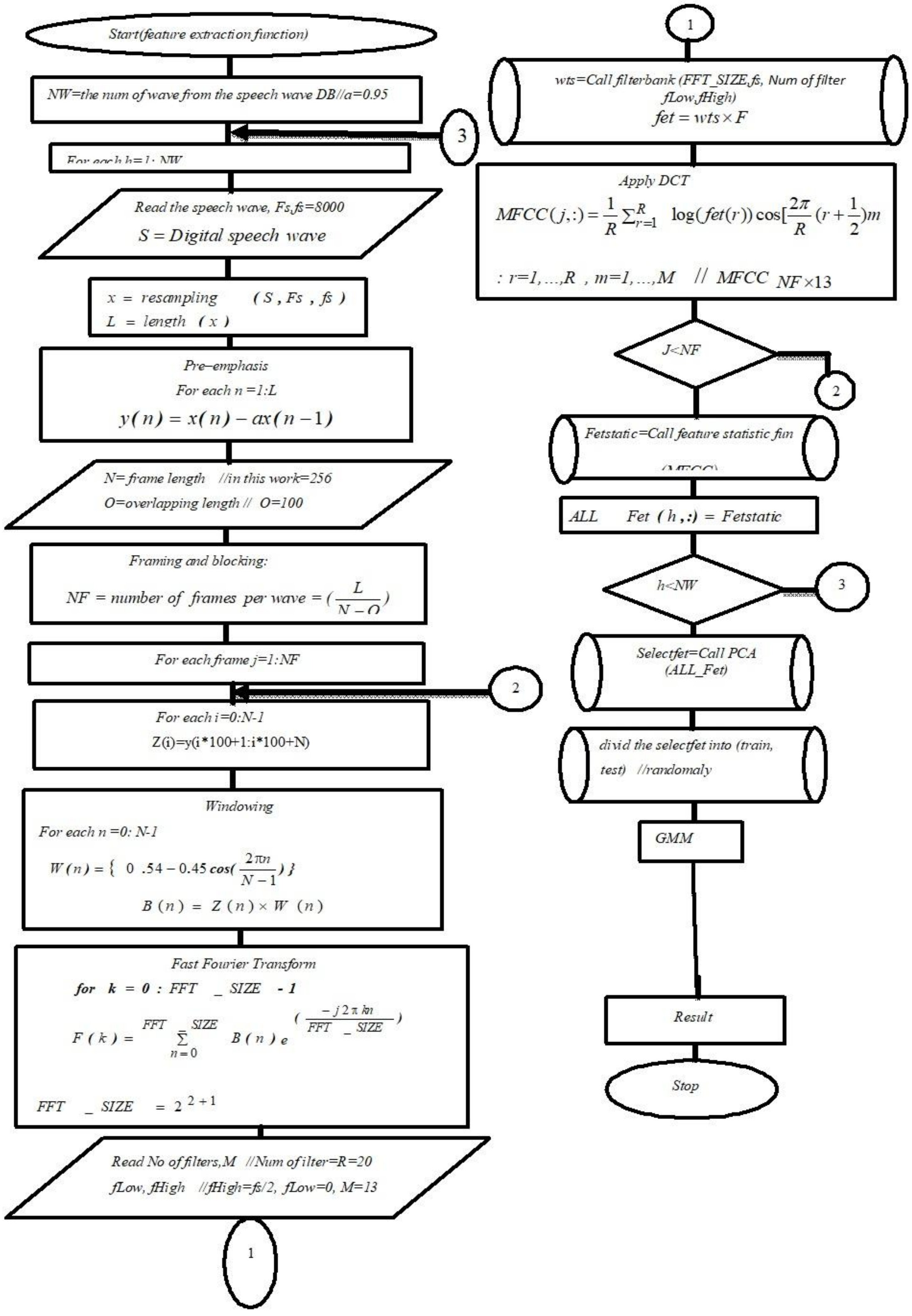


Fig (4). The main proposed system

مجلة التربوي
Journal of Educational
ISSN: 2011- 421X
Arcif Q3
معامل التأثير العربي 1.63
العدد 23

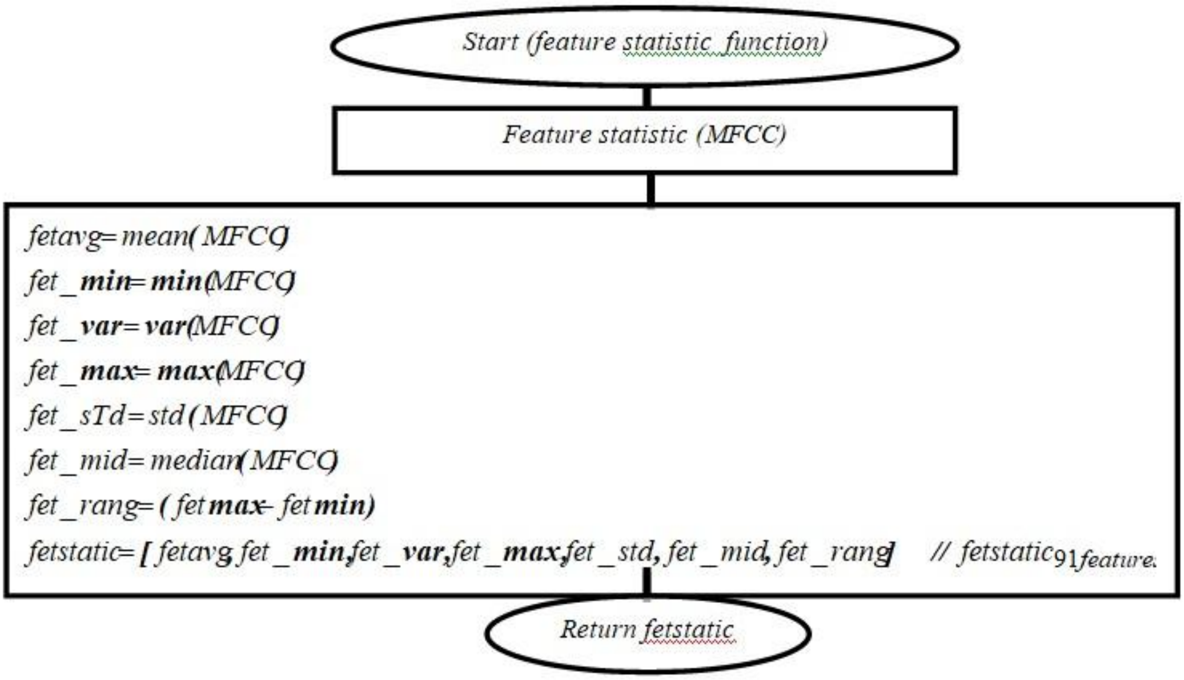


Fig (4.d). Feature statistic function

3.1 feature extraction

After extract the selected 13MFCC for each frame for each wave. Calculate the average of each MFCC Coefficient so we have 13 values per signal. For example if the signal contains 200 frames then we have 200 multiplied 13 features. The arithmetic average is not enough because it does not reflect the specific values of each speaker.

Therefore, this article proposed an adoptive MFCC by adding some statistics measures to extend the number of features. Each dimension with seven statistical

	Avg	Std	Var	Rang	Max	Min	Mid
MFCC1	MFCC1-avg	MFCC1-std	MFCC1-var	MFCC1- rang	MFCC1- max	MFCC1- min	MFCC1- mid
MFCC2	MFCC2-avg	MFCC2-std	MFCC2-var	MFCC2-rang	MFCC2-max	MFCC2-min	MFCC2- mid
MFCC3	MFCC3-avg	MFCC3-std	MFCC3-var	MFCC3- rang	MFCC3-max	MFCC3-min	MFCC3- mid
MFCC4	MFCC4-avg	MFCC4-std	MFCC4-var	MFCC4- rang	MFCC4-max	MFCC4-min	MFCC4- mid
MFCC5	MFCC5-avg	MFCC5-std	MFCC5-var	MFCC5- rang	MFCC5-max	MFCC5-min	MFCC5- mid
MFCC6	MFCC6-avg	MFCC6-std	MFCC6-var	MFCC6- rang	MFCC6-max	MFCC6-min	MFCC6- mid
MFCC7	MFCC7-avg	MFCC7-std	MFCC7-var	MFCC7- rang	MFCC7-max	MFCC7-min	MFCC7- mid
MFCC8	MFCC8-avg	MFCC8-std	MFCC8-var	MFCC8- rang	MFCC8- max	MFCC8-min	MFCC8- mid
MFCC9	MFCC9-avg	MFCC9-std	MFCC9-var	MFCC9- rang	MFCC9-max	MFCC9-min	MFCC9- mid
MFCC10	MFCC10-avg	MFCC10-std	MFCC10-var	MFCC10-rang	MFCC10-max	MFCC10-min	MFCC10-mid
MFCC11	MFCC11-avg	MFCC11-std	MFCC11-var	MFCC11-rang	MFCC11-max	MFCC11-min	MFCC11-mid
MFCC12	MFCC12-avg	MFCC12-std	MFCC12-var	MFCC12-rang	MFCC12-max	MFCC12-min	MFCC12-mid
MFCC13	MFCC13-avg	MFCC13-std	MFCC13-var	MFCC13-rang	MFCC13-max	MFCC13-min	MFCC13-mid

مجـلة الـتربـوي
Journal of Educational
ISSN: 2011- 421X
Arcif Q3
معامل التأثير العربي 1.63
العدد 23

characteristics. These statistics are average (avg), the standard deviation (std), variance (var), maximum (max), minimum (min), rang and Medium (mid) of Per coefficient of MFCC. Therefore we obtained 91 dimension vector as parameters of speech features for each wave, as shown in table (2).

The feature vector of 91dimension, consisting of the following Coordinates from 1to 13are the average,14 to 26 are the standard deviation, 27 to39 are variance, 39 to 52 are maximum, 52 to 65 are minimum,65 to 78 are rang and from 78 to 91 are Medium.

PCA technique is used to select the best significant feature vector from 91 dimension.

The selected features are divided into training and test data in the classification system.

The features extracted are reduction to 3 dimension by using PCA technique

3.2 Gaussian Mixture Model technique

From the previous section we have a dataset (D) with n (200) utterance in a d-dimensional, $D=\{x_i\}_{i=1}^{n}$ i, where d=3, Given i=1, and given the number of desired clusters k, the cluster her as speaker, the goal of representative-based clustering is to partition the dataset into k groups or clusters, which is called a clustering and is denoted as $C=\{C_1,C_2,...,C_k\}$ Where k=1,…,10. Further, for each speaker C_i there exists a representative utterance that summarizes the cluster, a common choice being the mean.

Let X_a denote the a^{th} random variable corresponding to the a^{th} features. We also use X_a to denote the features vector, corresponding to the n data samples from X_a . Let $X=(X_1,X_2,...,X_d)$ denote the vector random variable across the d-features, with x_j being a data sample from X. It is assume that each speaker C_i is characterized by a multivariate normal distribution, that is,

$$f_i(x)=f(x\backslash\mu_i,\sigma_i)=\frac{1}{(2\pi)^{\frac{d}{2}}|\sigma_i|^{\frac{1}{2}}}e^{-\left(\frac{(x-\mu_i)^T\sigma_i^{-1}(x-\mu_i)}{2}\right)} \quad (1)$$

Where the speaker mean $\mu_i \in R^d$ and covariance matrix $\sigma_i \in R^{d\times d}$ are both unknown parameters. $f_i(x)$ Is the probability density at x attributable to speaker C_i

It is assume that the probability density function of X is given as a Gaussian mixture model over.

3.2.1Train stage

The purpose of this stage is calculate the parameters for each speaker by several steps: Initialization, expectation, and maximization.

Initialization Step

For each speaker C_i , with i = 1,2, . . . ,k, we randomly initialize the mean μ_i by selecting a value μ_{ia} for each dimension X_a uniformly at random from the range of X_a . The covariance matrix is initialized as the d ×d identity matrix $\sigma_i = I$. Finally, the cluster prior probabilities are initialized to $P(C_i)$ = 1/k, so that each speaker has an equal probability.

Expectation Step

In the expectation step, we compute the posterior probability of speaker c_i given utterance x_j with i = 1, . . . ,k and j = 1, . . . ,n. As before, we use the shorthand notation $w_{ij} = p(c_i \mid x_j)$ to denote the fact that $w_{ij} = p(c_i \mid x_j)$ can be considered as the weight or contribution of utterance x_j to speaker c_i, and use the notation $w_i = (w_{i1}, w_{i2}, \ldots, w_{in})^T$ to denote the weight vector for speaker c_i, across all the n utterance.

$$w_{ij} = p(c_i \mid x_j) = \frac{f(x_j \mid \mu_i, \sigma_1^2)\, p(c_i)}{\sum_{a=1}^{k} f(x_j \mid \mu_a, \sigma_a^2)\, p(c_a)} \quad (2)$$

Maximization Step

Given the weights w_{ij}, in the maximization step, we re-estimate σ_i, μ_i and $P(C_i)$. The mean μ_i for speaker c_i can be estimated as

$$\mu_i = \frac{\sum_{j=1}^{n} w_{ij} \cdot x_j}{\sum_{j=1}^{n} w_{ij}} \quad (3)$$

Considering the covariance between dimensions X_a and X_b is estimated as

$$\sigma_{ab}^{i} = \frac{\sum_{j=1}^{n} w_{ij}(x_{ja} - \mu_{ia})(x_{jb} - \mu_{ib})}{\sum_{j=1}^{n} w_{ij}} \quad (4)$$

Expectation-Maximization Clustering

Where x_{ja} and μ_{ia} denote the values of the a^{th} dimension for x_j and μ_i, respectively. The prior probability $P(C_i)$ for each speaker is, given as

$$P(C_i) = \frac{\sum_{j=1}^{n} w_{ij}}{n} \quad (5)$$

Finally, after the training we have all parameters $\theta = \{\mu_i, \sigma_i, P(C_i)\}$,i=1:k.

3.2.1Test stage

From the previous section we have the parameters GMM model for each speaker as shown in figure (5).

This stage, we have a speech wave y_j and its categories by equation (6) speaker recognized

$$w_{ij} = p(c_i \mid y_j) = \frac{f(y_j \mid \mu_i, \sigma_1^2)\, p(c_i)}{\sum_{a=1}^{k} f(y_j \mid \mu_a, \sigma_a^2)\, p(c_a)} \quad (6)$$

EM Clustering Algorithm In the multivariate EM clustering algorithm .After initialization of μ_i, σ_i and $P(C_i)$ for all i = 1, . . . ,k, the expectation and maximization steps are repeated until convergence. For the convergence test, we check whether $\sum_i \left\| \mu_i^t - \mu_i^{t-1} \right\|^2 \leq \varepsilon$, where $\varepsilon > 0$ is the convergence threshold, and t denotes the iteration. The iterative process continues until the change in the speaker means becomes very small.

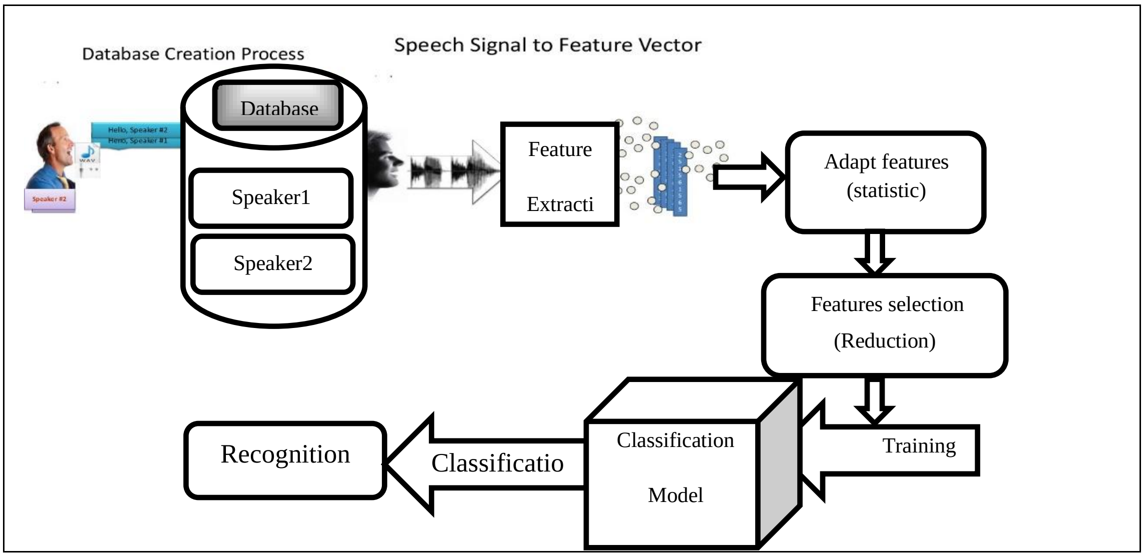


Fig (6): The proposed speaker recognition system

Table (4) the recognition accuracy of the proposed

Target	Accuracy
	GMM
S1	75%
S2	76%
S3	82%
S4	72%
S5	82%
S6	73%
S7	74%
S8	82%
S9	73%
S10	75%
Total accuracy	76.4%`

3.4 Discussion

The article presented an adoptive MFCC by adding some statistical method such as the standard deviation, variance, maximum, minimum, rang and Medium.

In order to evaluate the performance of the proposed feature extraction technique and the designed classifier, experiment is carried out. The results gained after applying the proposed system through the five partitions shown in the proposed system methodology in Fig. 4 many times in different experiments. The efficiency of the proposed recognition system is judged from two different perspective: recognition rate and network performance, which are illustrated in Table (5).

Table 5. The best neural networks and GMM experimental results

Feature extraction technique	Classifier technique	Feature vector length	Feature selection	Training performance	Training recognition rate	Testing recognition rate	Predicting recognition rate	Epochs number
Traditional MFCC	GMM	13	3		73.2%		71.3%	
	NNT	13	10	E-5	94.69%	80.79%	64.26%	1000
Adoptive MFCC	GMM	91	3		76.4%		74.1%	
	NNT	91	10	E-5	100%	99.27%	96.10%	1000

The result in table (4) show that the adoptive MFCC technique is more effective than the traditional MFCC.

3.5 Conclusion

This paper proposed a speaker recognition system based on newly proposed and powerful extracted features from the speech signal. MFCCs coefficients were commonly used in most of the previous studies for speaker recognition. However the traditional systems only used the average value of these coefficients for all frames in the speech signal. The number of extracted features using MFCC usually 13 feature. This number of extracted features are not enough to distinguish speakers in their systems. Therefore, this paper proposed an adoptive method for extracting more expressive and distingue features form MFCCs coefficients. Six statistics were calculated for each of the 13 coefficients for all frames, the obtained 78 new features in addition to the 13 traditional features. Therefore our features vector consists of 91 features. In order to select the most important features for the proposed system PCA method is applied to reduce the features dimensions from 91 to 3 discriminate features. GMM classification methods were proposed for speaker classification based on the 3 discriminated features. The results of the speaker recognition with the proposed system outperforms the traditional system.

Reference

[1] Tomi Kinnunen, Haizhou Li, "An overview of text-independent speaker recognition: From features to supervectors", Speech Communication Vol 52 P12–40 2010.

مجـلـة الـتربـوي
Journal of Educational
ISSN: 2011- 421X
Arcif Q3

معامل التأثير العربي 1.63
العدد 23

[2] Nilu Sing, R.A.Khan,Raj Shree, "Applications of Speaker Recognition", Procedia Engineering ,Vol 38 (2012),pp 3122 – 3126

[3] Nidhi Desai, .Kinnal Dhameliya, Vijayendra Desai, "Feature Extraction and Classification techniques for Speech Recognition: A Review", International Journal of Emerging Technology and Advanced Engineering (IJETA) Certified Journal, Vol 3, Issue 12,p367-371, December 2013.

[4] Jia-Ching Wang, Jhing-Fa Wang, Yu-Sheng Weng, "Chipdesign of MFCC extraction for speech recognition", INTEGRATION, the VLSI journal 32 (2002) 111–131, 2002.

[5] Bhargab Medhi and P.H. Talukdar, " Different acoustic feature parameters ZCR, STE, LPC and MFCC analysisof Assamese vowel phonemes ", International Conference on Frontiers in Mathematics (ICFM), p39-43 2015.

[6] Francesc Alías, Joan Claudi Socoró and Xavier Sevillano, "A Review of Physical and Perceptual Feature Extraction Techniques for Speech, Music and Environmental Sounds", applied sciences (Appl. Sci.), vol 6, Issue143, p1-44 2016.

[7] Najiya Omar," Speaker Identification System Enhanced By Optimized Neural Networks And Feature Fusion Techniques Evaluated By Cochlear Implant-Like Spectrally Reduced Speech", the degree of Master of Applied Science at Dalhousie University Halifax, Nova Scotia February 2017.

[8] Navnath S Nehe and Raghunath S Holambe,"DWT and LPC based feature extraction methods for isolated word recognition", EURASIP Journal on Audio, Speech, and Music Processing 2012,,p1:p7

[9] Alfie Tan Kok Leong, "A Music Identification System Based on Audio Content Similarity", Oct-2003.

[10] Lei Xie, Zhi-Qiang Liu, "A Comparative Study of Audio Features For Audio to Visual Cobversion in MPEG-4 Compliant Facial Animation", Proc. of ICMLC, Dalian, 13-16 Aug-2006.

[11] Namrata Dave, "Feature Extraction Methods LPC,PLP and MFCC In Speech Recognition", International Journal for Advance Research in Engineering and Technology, Vol 1, Issue VI, pp1-5 2013.

[12] RajivChechi, Reetu, "Performance Analysis of MFCC And LPCC Techniques In Automatic Speech Recognition", International Journal of Engineering Research & Technology (IJERT), Vol. 2 Issue 9 ,pp 3142-3146 2013

[13] Dr E.Chandra, K.Manikandan,M.S.Kalaivani, "A Study on Speaker Recognition System and Pattern classification Techniques", International Journal Of Innovative Research In Electrical, Electronics, Instrumentation And Control Engineering Vol. 2, Issue 2, pp 772-775, 2014.

[14] Md Jahangir Alam , Tomi Kinnunen , Patrick Kenny , Pierre Ouellet and Douglas O'Shaughnessy, "Multitaper MFCC and PLP features for speaker verification using i-vectors", 2012 Elsevier B.V. All rights reserved, SciVerse ScienceDirect.

[15] Vladimir Fabregas Surigué de Alencar and Abraham Alcaim ,"Transformations of LPC and LSF Parameters to Speech Recognition Features", S. Singh et al. (Eds.): ICAPR 2005, LNCS 3686, pp. 522 – 528, 2005© Springer-Verlag Berlin Heidelberg 2005.

[16] Wei Guo, Liqing Zhang, and Bin Xia," An Auditory Neural Feature Extraction Method for Robust Speech Recognition", 2007 Ieee, Icassp 2.
[17] Juan A. Morales-Cordovilla, Antonio M. Peinado," Feature Extraction Based onPitch-Synchronous Averaging for Robust Speech Recognition", Ieee Transactions On Audio, Speech, And Language Processing, Vol. 19, No. 3, March 2011.
[18] Chulhee Lee, Donghoon Hyun, Euisun Choi, Jinwook Go, and Chungyong Lee, "Optimizing Feature Extraction for Speech Recognition", Ieee Transactions On Speech And Audio Processing, Vol. 11, No. 1, January 2003,P80:P87.
[19] Athira Aroon, S.B. Dhonde, "Speaker Recognition System using Gaussian Mixture Model", International Journal of Computer Applications (0975 – 8887) Vol 130 – No.14, November 2015.
[20] Douglas A.Reynolds, and Richard C Rose, "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models", IEEE Transactons on Speech and Audio Processing, Vol 3, No 1, January 1995.
[21] Sreenivasa Rao Krothapalli, Shashidhar G. Koolagudi, Emotion Recognition using Speech Features, Springer Science+Business Media New York 2013.
[22] Zhenhao Ge, Ananth N. Iyer, Srinath Cheluvaraja, Ram Sundaram, Aravind Ganapathiraju," Neural Network Based Speaker Classification and Verification Systems with Enhanced Features", Intelligent Systems Conference 2017, IEEE,P1:P6.
[23] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, B. Weiss, " A Database of German
[24] Talieh Seyed Tabtabae, "Speech-based human emotion recognition", Ryerson University 2007.
[25] Mohammed J. Zaki, Wagner Meira Jr, Data Mining And Analysis, Cambridge University Press, First published 2014.
[26] Fred Richardson, Douglas Reynolds, Fellow, and Najim Dehak, "Deep Neural Network Approaches to Speaker and Language Recognition",Ieee Signal Processing Letters, Vol. 22, No. 10, October 2015.
[27] Yanick Lukic, Carlo Vogt, Oliver D¨urr, Thilo Stadelmann, "Speaker Identification And Clustering Using Convolutional Neural Networks", IEEE International Workshop On Machine Learning For Signal Processing, Sept. 13–16, 2016, Salerno, Italy
[28] V.srinivas, Ch.Santhi rani and T.Madhu, "Neural Network based Classification for Speaker Identification, "International journal of SignalProcessing and Pattern Recognition, Vol7, No.1 (2014), pp109:120.
[29] Kamil A. Kami ´nski 1,2,* and Andrzej P. Dobrowolski 3"Automatic Speaker Recognition System Based on Gaussian Mixture Models, Cepstral Analysis, and Genetic Selection of Distinctive Feature" 2022.
[30] Ugur Ayvaz, Hüseyin Gürüler, Faheem Khan, Naveed Ahmed , Taegkeun Whangbo and Abdusalomov Akmalbek Bobomirzaevich, Automatic Speaker Recognition Using Mel-Frequency Cepstral Coefficients Through Machine Learning, Tech Science Press, CMC, 2022, vol.71, no.3

مجلة التربوي
Journal of Educational
ISSN: 2011- 421X
Arcif Q3

معامل التأثير العربي 1.63
العدد 23

الفهـــرس