Open Access **Baghdad Science Journal** **P-ISSN: 2078-8665**
2020, 17(3) Supplement (September):1019-1028 **E-ISSN: 2411-7986**

DOI: http://dx.doi.org/10.21123/bsj.2020.17.3(Suppl.).1019

Voice Identification Using MFCC and Vector Quantization

***Bassel Alkhatib*[1]** ***Mohammad Madian Kamal Eddin*[2*]**

[1]Web Master Director- Syrian Virtual University - Damascus Syria and the Faculty of Information Technology Engineering-Damascus University. Syria
[2]Student at the Web Science program-Syrian Virtual University. Damascus Syria
[*]Corresponding Author: k.madian123@gmail.com
[2]ORCID: https://orcid.org/0000-0003-3806-2920
[1]drbasselalkhatib@gmail.com

Received 19/7/2019, Accepted 2/2/2020, Published 8/9/2020

Abstract:

The speaker identification is one of the fundamental problems in speech processing and voice modeling. The speaker identification applications include authentication in critical security systems and the accuracy of the selection. Large-scale voice recognition applications are a major challenge. Quick search in the speaker database requires fast, modern techniques and relies on artificial intelligence to achieve the desired results from the system. Many efforts are made to achieve this through the establishment of variable-based systems and the development of new methodologies for speaker identification. Speaker identification is the process of recognizing who is speaking using the characteristics extracted from the speech's waves like pitch, tone, and frequency. The speaker's models are created and saved in the system environment and used to verify the identity required by people accessing the systems, which allows access to various services that are controlled by voice, speaker identification involves two main parts: the first part is the feature extraction and the second part is the feature matching.

Key words: MFCC, Recording and signal processing, Speaker Identification, Vector Quantization.

Introduction

The voice is a signal made of tone or a number of tones connected together to mean something and used to communicate between humans or any living organism, through which they express what they want to say or do consciously or unconsciously, and the sensation caused by that waves called hearing. Because of the voice, people get many experiences in life. In the past, the sound that made by humane (by their throats) was not the only way they used to communicate with each other but also they used many things that make noise and vibration like drums and flutes. As known that the speed of the sound in natural is 343 meter per second or 1224 kilometer per hour, and it is related to the density of the object that the sound is moving through The audio signal is changing over time, assumed on short time periods the sound signal changes slowly (the samples change continuously even on short time scales). In addition, to understand voice recognition, two things may be confused: voice recognition and speech recognition. The first aims to identify who is speaking and the second is to determine what is being said, the term "voice recognition" used for each of them. The main concern is the declaration of the authentication process that indicates the verification of the speaker's identity and the automatic recognition of speech that indicates the identification of the spoken words. The speaker's understanding would make the process easier to implement the rules of the system's environment and the task of the authentication that being built where the system translating the user's voice and consider it as part of a security operation.

The main objective of voice-based systems is to customize security operations to suit the needs of

users and to contribute to the development of security operation performance. It is therefore critical to create individual files based on the analysis of the speaker's speech (Fig.1). These data should be used and effectively exploit in the security environment. In this area of working artificial intelligence techniques are useful for several reasons; including the ability to develop and mimic decision-making processes, many artificial intelligence techniques used by security systems are based on physical properties like Gaussian Mixture Models (GMM), Hidden Markov models (HMM), Artificial Neural Network (ANN) and Vector quantization (VQ).
The goal of this paper is to build an index model in the system environment that can be used as a part of a security operation that identifies the users from the unique physical properties of each voice and moves from traditional methods to more rigorous and highly reliable user identification patterns. For this reason, the digital processing speech signal and the way that the system will deal with the incoming voices and the procedure of analyzing the data play an important part in fast identification and recognition. In order to recognize the speaker in the speaker identification system, the signal must pass through several phases, which present briefly:

1. Recording & signal processing
2. Feature Extraction
3. Feature Matching

The system's structure shows in Fig.1:

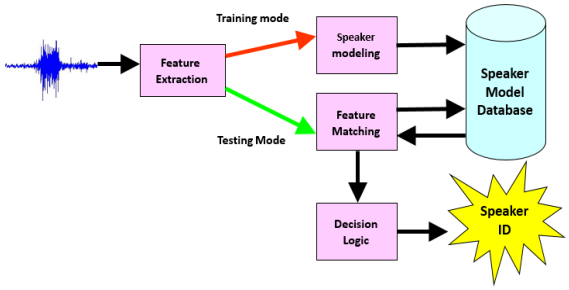

Figure 1. Speaker identification system Basic structure

Speaker Identification

The speaker identification systems usually require several operations and phases in which the voice signal must pass through to reach the result. There are two classes of these systems and our system is classified as text-independent identification systems (Fig.2).

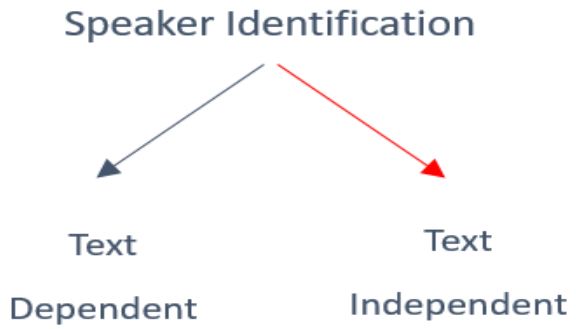

Figure 2. Speaker recognition types

The first phase in such systems is to train the system environment on the new speeches to form knowledge and create the reference models from these speeches, where each speaker must provide samples of his speech so that the system can create the reference model for the speaker (Fig.3). It consists of two main parts. Part 1: consists of processing the speech sample provided by the speaker to condense and summarize the properties of the acoustic tract and it is the first phase of the system and calling the training phase, Part II: Collecting the data of each speaker together in one matrix that can be easily processed. The second phase is the test phase where it reflects the structure of the training phase. First, analyze the input signal, and then compare the data stored in the Codebook (Fig.4). The difference between the input speech and the stored one is used to make the decision in the system.

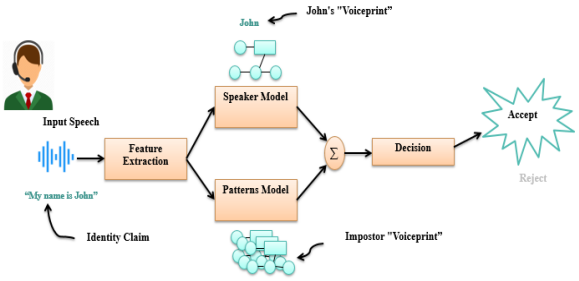

Figure 3. Speaker identification

The nature of the system determines the choice of technology used in the application. The application that is built does not depend on a specific sentence or words to allow the user to access to the system, in other words, the user does not have to say a specific sentence already stored in the database, he will say anything and the system will analyze his voice and make the decision based on that analysis.

Generally, all speaker recognition systems have two basic units: extract features and match these features. The first part is the extract features,

Baghdad Science Journal
P-ISSN: 2078-8665
E-ISSN: 2411-7986

which extract the physical characteristics from the user talking speech and analyzing that physical characteristics later to represent if the user authenticates to get in or not, and for determining who is speaking by comparing features extracted from the input speech with forms saved or known by the system, which already explained.

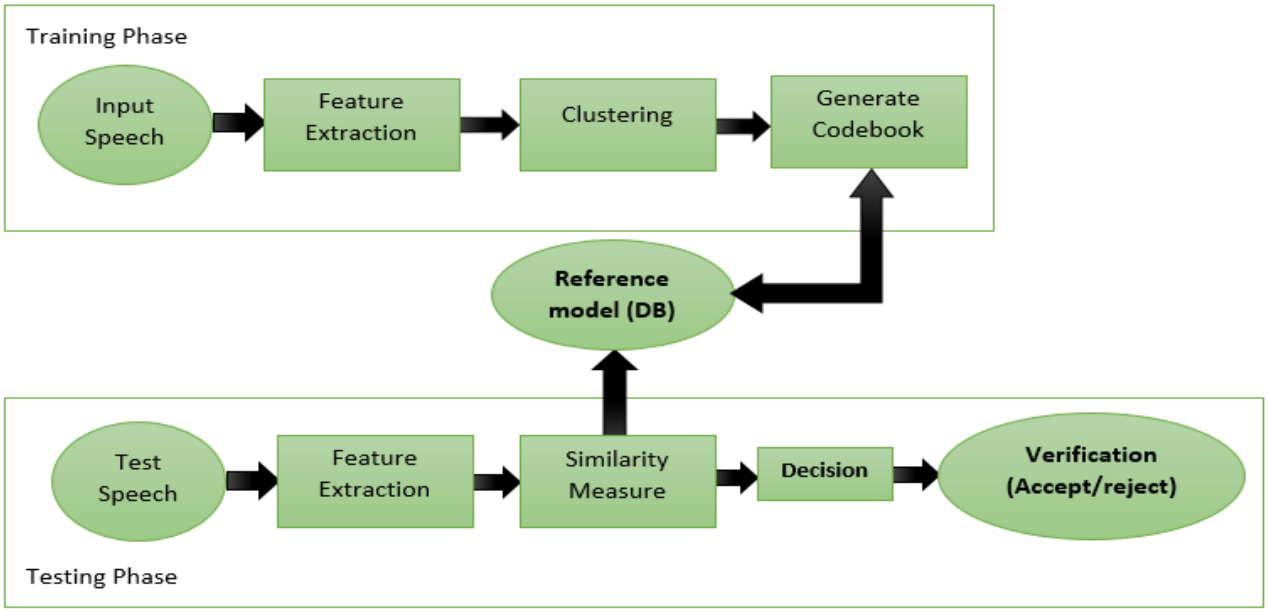

Figure 4. Speaker's identification block diagram

Recording & Signal Processing

Mean Correction

The first process that can be applied is to modify the signal values according to the average as the purpose of this process is to reduce the effect of any continuous frequency (1) produced by the recording devices. A certain threshold is selected from the mean and subtract it from the signal values; this process does not change the shape of the signal but modifies the frequencies slightly as follows:

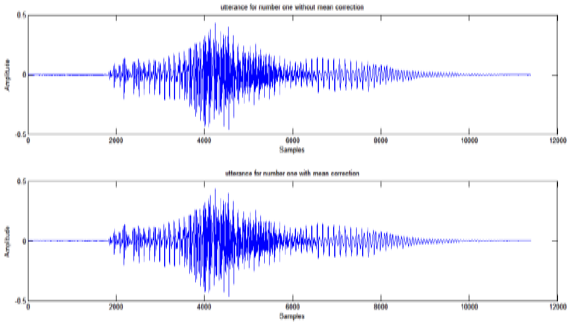

Figure 5. Number one with a threshold of 0.05

As shown in (Fig.5), there is no clear difference between the two signals, the primary objective of this process is to try to mitigate the impact of the continuous frequency produced by the recording devices.

Speech Boundary Detection

Often, moments of silence may pass before and after recording. These moments may affect the quality of the sound sample in distinguishing the content of this sample. Therefore, these static

samples removed from the signal and the samples containing the operative sound information must be removed. Because of that, a short-term energy measure is used and this method is one of the most widely used methods in edge detection algorithms because it gives us the power to distinguish between sounds and silence, as this method relies on signal energy (Fig.6).

$$E_{log} = \sum_{i=1}^{n} \log\left(s(i)^2\right)$$

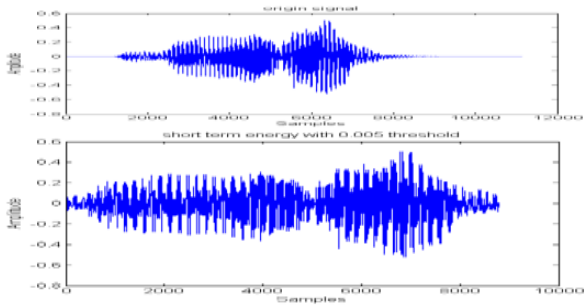

Figure 6. The signal after cutting the silence edges

Feature Extraction

At this stage, techniques that contribute to the feature extraction of the audio signal that helps to distinguish the content of the signal from others are applied. These features are used to train a probabilistic system or a neural network to distinguish speech content. The first step in any system for speaker identification is feature extraction, which aims to identify voice signal components to recognize the language content and cutout anything else like noise and emotion. So to understanding voices that the audio forms including the tongue, teeth, etc. filter the sounds generated by the human, and this shape determines the resulting sound. If the voice data could be modeled, that should give a better prediction of the sound made by the user. The modeling of the voice data channel (audio channel) shown by the short spectrum energy (envelope) and the MFCC can represent this state accurately. The design of the system and the environment required by the speaker to pass through deferent operations to recognize the voice. Moreover, the first one is getting the input speech through the microphone from the user speech signal and then apply the process steps on that signal to identify the user.

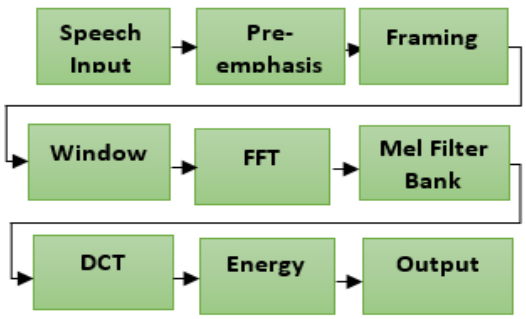

Figure 7. Block diagram of MFCC

Why MFCC's steps are necessary: The voice is continuous over time; the voice signal is changing slowly on short time periods and the samples are changing with it.

That is why the voice signal has to be cut into frames (20-40ms) witch shown in Fig.7. So the point of understanding is, if the section is short, there will not be enough samples and the signal will change in a fast way which prevents to get the spectral energy that is reliable. Then the energy spectrum has to be calculated for each frame. From here, the Mel filter-bank is obtained, and then the Mel filter-bank is calculated.

Why would the system use a logarithm rather than a cube root: The logarithm allows us to use subtraction, the channel normalization, and to imitate human cognition of sound because experiments have shown that humans recognize sounds on a logarithmic scale (2).

Pre-emphases

Pre-emphases refers to focusing the filtering process on the high frequency and it aims to equilibrate the whole frequencies of the sounds and it is defined as:

$$z(i) = y(i) - a * y(i-1)$$

Output: $z(i)$.
Value of (a): Between 0.9 and 1.0.
as shown in (Fig.8), it raises the frequencies of high frequencies versus low frequencies, to increase the predictability of the sounds represented by these frequencies (3).

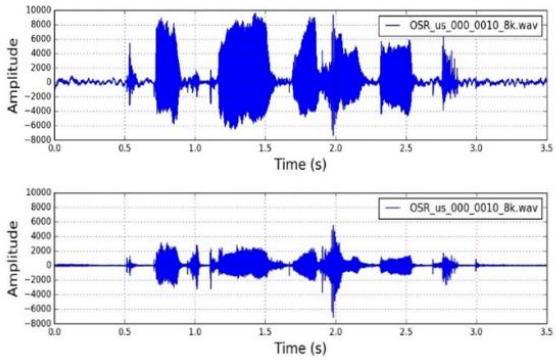

Figure 8. The signal before and after Pre-emphasis

Blocking the Signal & the Window Function

Framing the signal is the process of dividing the reference into a number of sections, each called a frame, and each frame is treated independently from the rest of the frames. In most cases, the signal cannot be handled simultaneously because this may cause unsatisfactory results, since the Fourier

transform does not specify the time frames corresponding to a given frequency, so the chipping is the ideal solution to calculate frequencies in a semi-local manner. Timing frequencies. Windowing is the process of calculating the similarity between two mathematical subjects. The first is the sign or frame, the second is the mathematical object that expresses the window, and the window function is the sum of the circumference between the frame signal and the window. As mentioned before the speech signal is continuous with the pass of time or semi-static slowly. For stable audio properties, the speech signal must be examined within enough short periods. Speech analysis should be performed on small frames where the speech signal assumed to be static. The process usually performed over 20 milliseconds, and progress every 15 milliseconds. Increasing the hope size (step) every 15 milliseconds (Fig.9) enables tracking the physical characteristics of the voice data, and moving on the frames of 30 milliseconds is good to provide valid spectral analysis of the given voice signal, while short enough to solve important temporal characteristics (3). In general, Hamming windows are used to surround the signal as shown in the following figure (Fig.9). This was made on the signal to prevent losing the spectral energy from the edges of each frame after the edge cut in the blocking step and to make the value at both ends of the frame near to zero to maintain the harmony of the frame (4).

$$Y(n) = x(n)W(n)$$

Hamming window used for speaker recognition task is defined as:

$$W(n) = 0.54 - 0.46\cos(\frac{2\pi}{N-1})$$

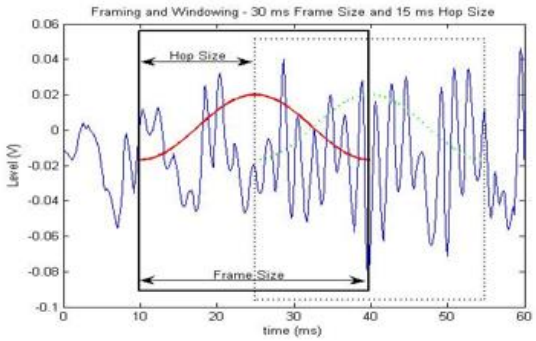

Figure 9. Framing & windowing

Fast Fourier Transform FFT

It is an algorithm that collects a signal over a period of time (or space) and divides it into its frequency components, these components are single sinusoidal oscillations at different frequencies, each with its own amplitude and frequency. Fourier transform converts the data from the time domain to the frequency domain (5).This transform descript in the following figure (Fig.10):

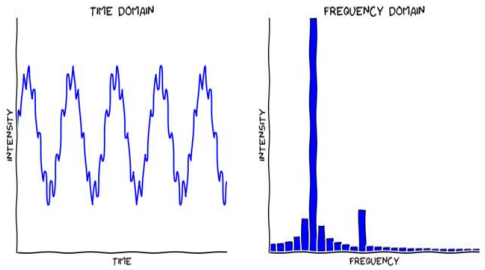

Figure 10. Fast Fourier Transform

Mel-Frequency Wrapping

It is calculated by passing a Fourier signal that converted in the last step to a number of filters called Mel filter bank. The Mel is a measurement scale shown in Fig.11 depends on the frequencies that the human can hear, and does not have close similarity in linear scale for the frequency tone; it seems that the human hearing system does not recognize linear vibrations. The Mel scale is approximately a linear spacing below 1 kHz and a logarithmic distance above 1 kHz and a series of triangular bandpass filters designed to simulate the bandpass filtering believed to occur in the audible system. Therefore, the following approximate formula could be used to compute the Mel for a given frequency in Hz (5).

$$Mel(f) = 2595 * \log(1 + f/700)$$

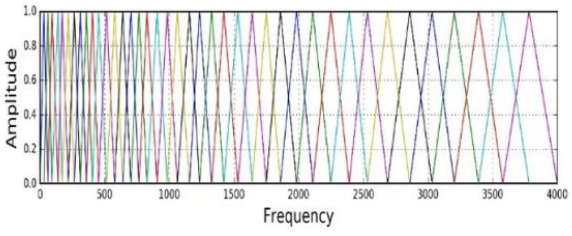

Figure 11. Mel-filter bank

Discrete Cosine Transform

In this step, the objects that represent each Cepstral Coefficients are calculated. Where objects

are the characteristic features of the signal (after experimenting, 13 objects were selected to represent each frame).

Because of the smooth of the audio channel, the energy levels of the near data points connected together, The DCT performs on that data points to convert the Mel frequency coefficients resulting in a set of Cepstral Coefficients. Before DCT calculation, the Mel spectrum is typically represented on a logarithmic scale this results in a signal in the Cepstral domain with the frequency peak corresponding to signal vibration and a number of formulas representing low-frequency peaks. Since most signal information represented by the first few MFCC transactions, the system can be more accurate by extracting transactions that ignore or interrupt the higher-order DCT components. Finally, the DCT transform can be performed on the output of the Mel filters by performing the following equation:

$$C(n) = \sum_{m=0}^{M-1} \log_{10}\big(s(m)\big) \cos\left(\frac{\pi n(m-0.5)}{M}\right)$$

Where n = 0, 1, …, C-1

$C(n)$: Cepstral coefficients.
C: features number of MFCC.

In this project, there are n = 40 and C = 13. As FFT is applied on the signal, DCT transforms the domain backward.at the end of this process the features needed for the next step of matching and comparing the voices are saved in the system environment.

Vector Quantization

At this stage, the features extracted in the feature extraction phase, the values are converted into a form used as an input for the probabilistic model. Similar values grouped together and given a common value. This called clustering.

Until this point, the characteristic features of each speech signal is obtained in the identification system, but the direct use of these features is not possible for a number of reasons:

1. In Pattern Recognition systems (which our system is a part of) specific data values are needed witch should mark the data point with a value.
2. Because of the large number of data where its values are close, these data can be Summarizing by a table representing an index of these data and its values.

For these reasons, data cannot be directly manipulated, as the data must be converted into the right shape to be suitable for later processing. This conversion process called indexing. It is a process whereby each data point represents the speech signal given a value that distinguishes it from other elements so that this value is used to denote this data point and this value becomes the input of subsequent algorithms.

The idea of recognition based on the clustering process witch separate the data object into multiple categories or classes so that the objects in a cluster are similar, but they are quite different from the objects in other clusters. Similarities evaluated based on the values that describe the objects and often involve distance scales. The latest techniques for feature matching used in speaker identification are Hidden Markov Modeling (HMM), Gaussian Mixture Model (GMM) and Vector Quantization (VQ). The main idea of the VQ is that it is the process of shorthand the whole data object from large data points in the space to a region contains a set of numbers of data that clustered in that region which can be represented by its center and that center called Codewords. All Codewords called Codebook; VQ is the technique used in this system because of the fast and high accuracy of the comparison. The VQ basically created as a compression algorithm witch indexing the Codewords in an index called Codebook and it's from the type of lossy data (6).

Vector Quantization is one of the most common techniques used in the field of indexing for audio samples, because of its multiple uses in the areas of voice compression and speech recognition, the use of this technique has increased since the start of the use of linear prediction technology LPC in the sixties of the last century. The idea of the Vector Quantization is to give the signal a unique value from a set of values so that two different signals do not have the same value, and the converged and similar signals have the same value. The process of VQ is to limit the repetition of similar objects with a single value represents these objects.

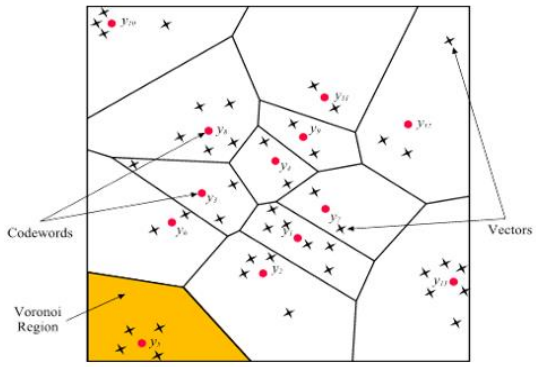

Figure 12. Two Dimensional VQ

The previous Fig.12 shows how the algorithm works by distributing data points to clusters so that data within the same cluster as mentioned before are as close as possible and as different as possible from data in other clusters. The following form shows two separated speakers the first speaker is shown as the green circles surrounding the black one, these are the Codevector and the centroid of that speaker (speaker 1), the second speaker is shown as the red circles surrounding the black one, these are the Codevector and the centroid of that speaker (speaker 2). The distance between the center of each region and the data point in that region called distortion and it is based on the Euclidean distance measurement as shown in Fig.13.

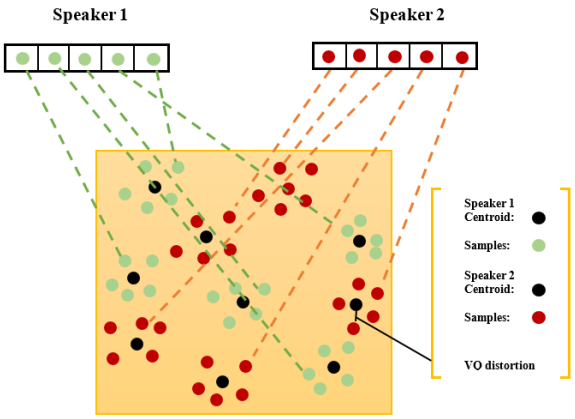

Figure 13. Index modeling

In the training phase from our project, The VQ Codebook has generated for each speaker and it aims to cluster its Codevectors in the given database (7).

The distance of the Euclidean's calculations of the speaker's Codevector that is closest to the Codebook is called distortion, in the process; this distance is used by the system to recognize the speaker with the lowest distortion. For a new speech signal of unknown speaker, the system will calculate the distance of its voice and if there is no match and the distortion is too high from its calculation, the system will decide that the speech signal does not belong to any of the speakers and ask the user to role in the system (8).

LBG Clustering Process

To build the index or Codebook, the system needs a clustering algorithm that assembles the converged vectors and finds the Codeword that represents them. The best and most widely used algorithm is the LBG algorithm (Fig.14), which divided Training data space into clusters to achieve the following two-optimization conditions:

1. Each training data (vector) must be close to a particular cluster and away from the rest of the clusters. In other words, the vector must belong to only one cluster so that the distance between this vector and the cluster center is smaller than all other distances or other cluster centers.
2. The center of the cluster is selected so that the distance between the center and all vectors is as low as possible. The average error (distance from center) for each cluster reduced. Taking into consideration, that the center of the cluster is the average of the vector's value that belongs to this cluster.

LBG steps

Initialization: At this point, the number of clusters K selected and initial vectors selected to represent these clusters, as these vectors selected randomly from training data.

Search the nearest neighbor: Calculate the distance between each vector and the center to find the nearest center for that vector.

Compute Euclidean distance: The total of the distances is calculated for this iteration, which represents the amount of error in this iteration.

Update centroids: The centers are updated where the average values of the closest vectors of this center, which resulted from step 2, are calculated. This average is the new center of the cluster.

Termination: Repeat the previous steps (except initialization) to get an error in step 3 below a specified limit previously selected or even reaching the greatest number of sessions.

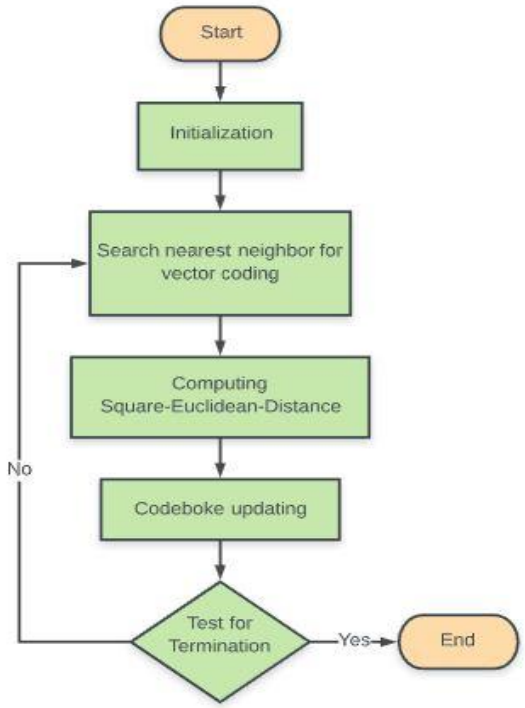

Figure 14. LBG process steps

At the end of this algorithm, the final clusters and the vectors they represent were obtained and achieve the best distance and the lowest possible error rate.

Using the VQ, the Codebook index is generated, as after extracting the training vectors, these vectors become frames and each of these frames has its own attributes so that a matrix is obtained in which each column represents a frame. In the beginning, the size of the Codebook index must be determined, the number of clusters K in which the frames will be clustered, and then the clustering algorithm is applied as shown previously. The index by K vectors is configured and selected from the training matrix randomly. The algorithm looks at each cluster in the training matrix for the nearest vectors from the index. This vector (located in the training matrix) gives the value indicating the location of the cluster that follows it. The distances obtained at this stage are combined to represent the total error for this iteration. After the attribution, the average values of the vectors of the same value are calculated from the index (the vectors that belong to the same cluster). This means a new center is selected. All previous steps are repeated until the difference between the old center and the new one is equal to zero as shown in the previous figure (Fig.14). At the end of the algorithm implementation, the final positions have been obtained that will convert the series of frames represented by an audio signal to a series of vectors representing the Codebook to that entered into the next step of matching.

Experiment Results & Analysis

As described earlier in the previous sections, in this paper, the idea is to build a voice identification system and test it, and to build this system the user must pass several steps from defining the system to the user's voice, where the physical characteristics of each voice were analyzed to identify the speaker.

Enrollment Phase or Training Phase

At this stage, our goal is to train the voice model (training the VQ model on MFCC extracted features) for all speech files in the training phase. The following point is that the system will have a good knowledge of the voices and its characteristics for each user, so for later steps, the system will determine the voice signals and identify them.

Speech Signal Processing

At this point, the speech signals are analyzed and converted to a series of MFCC features using the speech processing steps described earlier. After getting the speech signal and cut it to overlay (30 MS) frames. Then the frames stored, and its content in a matrix to represent each frame and what its hold of information from the original voice and apply the process steps of handling the signal where it is converted from its basic domain to another using Fast Fourier Transform, which is commonly used in many systems. Thus, the matrix of Cepstrum Coefficients is obtained that leads to MFCC treatment.

Building the Model

The previous process result is converting the voice data into several characteristics features, in this part the VQ approach techniques will be applied to those features. To achieve this, the system builds the previously described Codebook that trains the VQ models. Comparisons performed by calculating the Euclidean distance between the input speech signal and the models stored in the database (Table 1).

Euclidean distance between Codebook patterns for some users:

Table 1: Euclidean space distortion

	Sp1	Sp2	Sp3	Sp4	Sp5
Sp1	**4.7141**	8.91	10.7859	10.3863	11.4722
Sp2	6.7652	**5.5148**	6.3454	7.4315	7.7702
Sp3	5.1283	5.5148	**3.3465**	4.0861	4.3574
Sp4	3.5836	3.5063	3.2954	**2.404**	3.3519
Sp5	1.404	1.4154	1.3641	1.2587	**1.206**

The result of matching some users in the system:

Table 2: Matching results

	Sp1	Sp2	Sp3	Sp4	Sp5
Sp1	**Match**	No	No	No	No
Sp2	No	**Match**	No	No	No
Sp3	No	No	**Match**	No	No
Sp4	No	No	No	**Match**	No
Sp5	No	No	No	No	**Match**

The Codebook for some users that shown in the previous table (Table 2) witch the system has calculated the Euclidean distance for their voice signal.

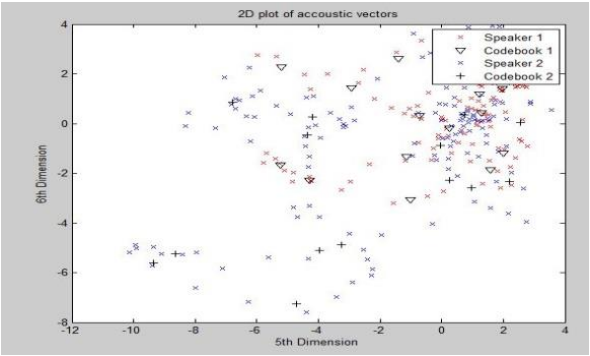

Figure 15. VQ codebook for some users

Conclusion

The purpose of this paper is to identify the speaker where a number of techniques have been used together to achieve the desired results from the system. The speaker's speech processed with a number of operations discussed in detail in the previous sections. Mel Frequency Cepstral Coefficient and Vector Quantization used together to extract the features and matching them. Moreover, both gave good performance and accuracy results. MFCC used to analyze the voice and extract the tone, pitch and frequency features of that speech. VQ (Fig.13) used to encode these features and matching the voices and speaker's Codebook must be updated to get satisfactory results in the environment of the system. Thus, the more the system is used, the faster the recognition of speakers and the more accurate the system becomes. And it differs from traditional methods such as Hidden Markov Models (HMM) because it is based on measures the similarity between two sequences that differ by time or speed (Table 1 and 2), like when the speaking speed changes. (Fig.15) shows that each speech could be concluded within the system can be indexed into a model and saved in the Codebook index. Using the Vector quantization can reduce the time of comparison and reduce the amount of the audio samples because as mentioned in the previous sections that this algorithm is originally designed for sound compression processes.

Authors' declaration:

- Conflicts of Interest: None.
- We hereby confirm that all the Figures and Tables in the manuscript are mine ours. Besides, the Figures and images, which are not mine ours, have been given the permission for re-publication attached with the manuscript.
- The author has signed an animal welfare statement.
- Ethical Clearance: The project was approved by the local ethical committee in Syrian Virtual University.

References

1. Pejman M, Josef Ku, Johannes S, Florian M. Single Channel Phase-Aware Signal Processing in Speech Communication: Theory and Practice. Wiley. 2016; 53-55. Available from: https://onlinelibrary.wiley.com/doi/book/10.1002/9781119238805
2. Niemann H. Klassifikation von mustern. springer-Verlag; 2013 Mar 13: Available from: https://www.springer.com/de/book/9783540126423
3. Mostafa E. Advanced Intelligent Systems for Sustainable Development: (AI2SD 2018). Springer, Mar 2019; Available from:https://www.springer.com/gp/book/9783030119270
4. Durgesh K M, Malaya K N, Amit J. Information and Communication Technology for Sustainable Development: Proceedings of ICT4SD Springer 2016; Available from: https://www.springer.com/gp/book/9789811039195.
5. Karpov E. Real-time speaker identification. University of Joensuu, Department of Computer Science, Master's Thesis. 2003 Jan 15.
6. Linde Y, Buzo A, Gray R. An algorithm for vector quantizer design. IEEE Trans. commun. 1980 Jan;28(1):84-95..

7. Soong FK, Rosenberg AE, Juang BH, Rabiner LR. Report: A vector quantization approach to speaker recognition. AT&T tech. J. 1987 Mar;66(2):14-26.
8. Rudresh MD, Latha AS, Suganya J, Nayana CG. Performance analysis of speech digit recognition using cepstrum and vector quantization. In2017 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECCOT) 2017 Dec 15 (pp. 1-6). IEEE.Available From:https://ieeexplore.ieee.org/document/8284580 DOI:10.1109/ICEECCOT.2017.8284580

التعرف على الصوت باستخدام MFCC و Vector Quantization

د. باسل الخطيب [1] محمد مدين كمال الدين [2]

[1]مدير برنامج الماجستير في علوم الوب، الجامعة الافتراضية السورية، دمشق، سوريا
مدير قسم الذكاء الصنعي، جامعة دمشق، دمشق، سوريا
[2]طالب في برنامج علوم الوب في الجامعة الافتراضية السورية، دمشق، سوريا

الخلاصة:

يعد التعرف على المتحدث أحد المشكلات الأساسية في معالجة الكلام ونمذجة الصوت. تتضمن تطبيقات التعرف على المتحدث المصادقة في أنظمة الأمان ودقة الاختيار. تشكل تطبيقات التعرف على الصوت تحديًا كبيرًا على نطاق واسع حيث يتطلب البحث السريع في قاعدة بيانات الاصوات تقنيات حديثة سريعة وتعتمد على الذكاء الاصطناعي لتحقيق النتائج المرجوة من النظام. تم بذل العديد من الجهود لتحقيق ذلك من خلال إنشاء أنظمة قائمة على المتغيرات وتطوير منهجيات جديدة لتحديد المتحدثين. التعرف على المتحدث هو عملية التعرف على من يتحدث باستخدام الخصائص المستخرجة من موجات الكلام الخاصة به مثل درجة الصوت والنغمة والتردد ويتم إنشاء نماذج المتكلم وحفظها في بيئة النظام وتستخدم لاحقا للتحقق من الهوية المطلوبة من قبل الأشخاص الذين يصلون إلى النظام، والذي يسمح بالوصول إلى الخدمات المختلفة التي يتم التحكم بها عن طريق الصوت، ويشمل تحديد المتحدث جزأين رئيسيين: الجزء الأول هو استخراج الميزات الصوتية أما الجزء الثاني فهو مطابقة ومقارنة هذه الميزات.

الكلمات المفتاحية: التعرف على المتحدث، التسجيل ومعالجة الإشارات، MFCC، Vector Quantization