ASR Features Extraction Using MFCC And LPC: A Comparative Study

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
drbasselalkhatib@gmail.com

[2]Student at the PhD program
Syrian Virtual University
Damascus. Syria
*k.madian123@gmail.com
ORCID: https://orcid.org/0000-0003-3806-2920

Journal of Digital Information Management

ABSTRACT: *The field of Automatic Speaker Recognition (ASR) is an important and open field for researchers and scientists, especially as it has become essential in facilitating the work we do in our daily lives. Such as digital authentication and electronic transactions, and consider It as a secure environment to authenticate users' access to their accounts. Many technologies have been developed in the field of recognition but so far, there is no complete tool or method for speaker identification, the most important step in ASR is the extraction of voice features. Many methods and tools can be used to extract the speaker's vocal characteristics (voice features), which in turn will identify the user and recognize his voice spectrum through the phonetic linguistic message. In this paper, two methods will be studied, each using a different technique MFCC, which uses a logarithmic scale, and LPC, which uses a linear scale. The method used in ASR should have minimal error because it is an important authentication technology like a fingerprint, where two different people cannot have the same voice spectral range (voiceprint).*

Subject Categories and Descriptors: [I.2.7 Natural Language Processing]; Speech recognition and synthesis: [H.5.1 Multimedia Information Systems]

General Terms: Automatic Speaker Recognition, Speech Recognition Technology, Voice recognition

Keywords: ASR, Speech Recognition, MFCC, LPCASR, Speech Recognition, MFCC, LPC

Received: 19 December 2022, Revised 28 February 2023, Accepted 9 March 2023

Review Metrics: Review Scale: 0/6, Review Score 4.85, Inter-reviewer Consistency: 88.5%

DOI: 10.6025/jdim/2023/21/2/39-49

1. Introduction

Speech recognition technology is a growth skill and it represents an important part of our daily lives, but now it is still limited to relatively simple commands.

As technology advances, researchers seek to create smarter systems for understanding speech and sounds (remember the people who do robot job interviews...!). One day, you will be able to talk to your computer like you are talking to any human being, and it will be able to send you logical responses.

All this will be possible through signal processing techniques. More and more researchers want to be a part of it. Processing, interpreting, and understanding speech signals are keys to many powerful new technologies and methods of communication. Given current trends, speaker recognition technology will be a rapidly growing (and world-changing) subset of signal processing for years to come.

Voice recognition automatically identifies who is speaking using the speaker's information contained in voice waves [1].

Verifying the identity claimed by people who access the systems allows controlling access to various services by voice and identifying users by recognizing their voices.

Because the voice cannot be the same for two persons as the voice characteristics differ between individuals like fingerprints [2] [3]. In addition, it represents a protection model that is analyzed by biometric patterns or artificial intelligence techniques and deep learning. it can be used to authenticate and verify the identity of the speaker as part of a security process Applicable services include access to user accounts through their voices, database access services, security control of confidential information, voice calling, telephone banking, telephone shopping, information and reservation services, voice mail, and remote access to computers.

With the increasing development of Internet technologies and electronic communication with the large spread of business conduct over the Internet, which in turn provides smooth and simple methods for carrying out all the tasks that took time in the past to implement, and the increase in electronic outlets. It is necessary to find modern ways and methods that adapt security operations to preserve the privacy of users and block those who try in one way or another to access their information and use it in malicious and illegal ways to achieve goals that harm the end-user or the target user.

Some Common Advantages of ASR:

1. ASR is easy to use for everyone where the user has to enroll his voice into the system and then he only needs to speak to verify his identity.

2.When the system is used many times, that leads to an increase in the accuracy of the verification and the system will be more effective.

3.ASR provides a security layer that preserves the users with more privacy, as it can be used with smart devices as a means to access their accounts using their voices (biometric patterns) instead of the old and traditional methods (Username/Password).

4.Voice recognition technologies in all their forms provide an easy way for people to access and control their accounts, as they can direct commands to devices or send e-mails using their voices.

5.This technology is giving some help and making things easier for people with disabilities or visual impairment [4] and putting them in the position of command to control things around them that work by voice.

6.The voice is representing a special character where it can't be the same for two persons as the fingerprint [2][3].

7.When you use ASR technology, you can do anything while you are talking to your device [4] Like driving a car and talking to your smartphone for directions.

Features Extraction:

MFCC:

Speech recognition is under the class of supervised learning. in ASR the first problem that the system will deal with is the input which will be represented as an audio signal, and it has to predict the speech signal from that input. So, we cannot process the raw audio signal input to the model because there will be a lot of extra signals that the system does not want [6].

Here where the MFCC algorithm takes action and plays its part in extracting the features from the audio signal and using it as input to the base model which will produce better accuracy than directly considering raw audio signal as input to the model. MFCC is widely used in the task of ASR for extraction features from the given input (audio signal).

The process steps of MFCC:

1. Digitalization.

2. Pre-emphasis.

3. DFT.

4. Mel filterbank.

5. IDFT.

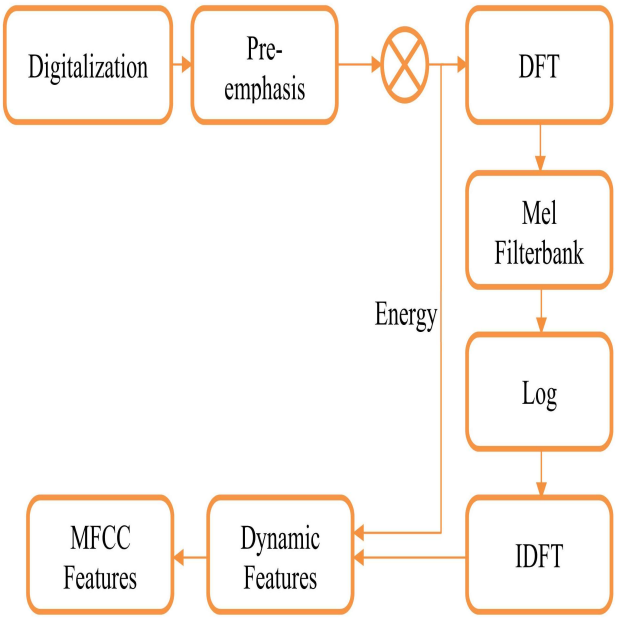


Figure 1. MFCC steps

Digitalization

In this step, the signal will be converted from analog into digital format with a sampling frequency of 8 kHz or 16 kHz [6]

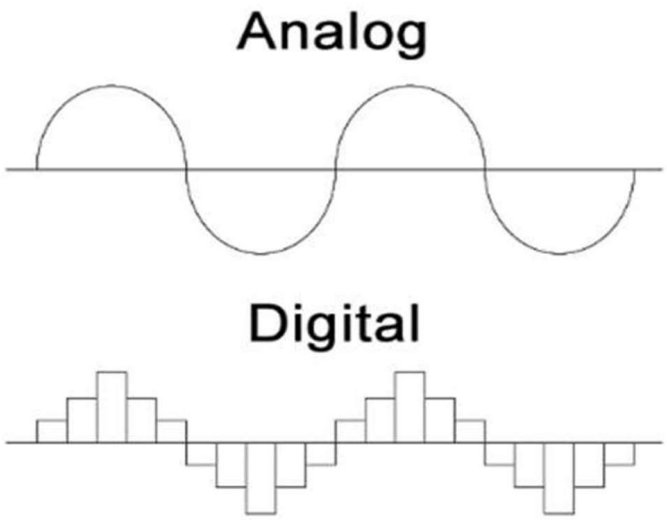


Figure 2. Converting the signal from analog to digital

Pre-emphasis:
In this step of processing, the magnitude of the energy will be increased in the higher frequencies' domain of the input audio signal, which will improve the detection accuracy and performance represented by these frequencies and balance the spectrum of voiced sounds that have a steep roll-off in the high-frequency region [7].

Because the energy at the higher frequency is lesser than the energy at the lower frequency, it is focused to equilibrate the whole frequencies of the audio signal.

It is done using the first-order high-pass filter 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.

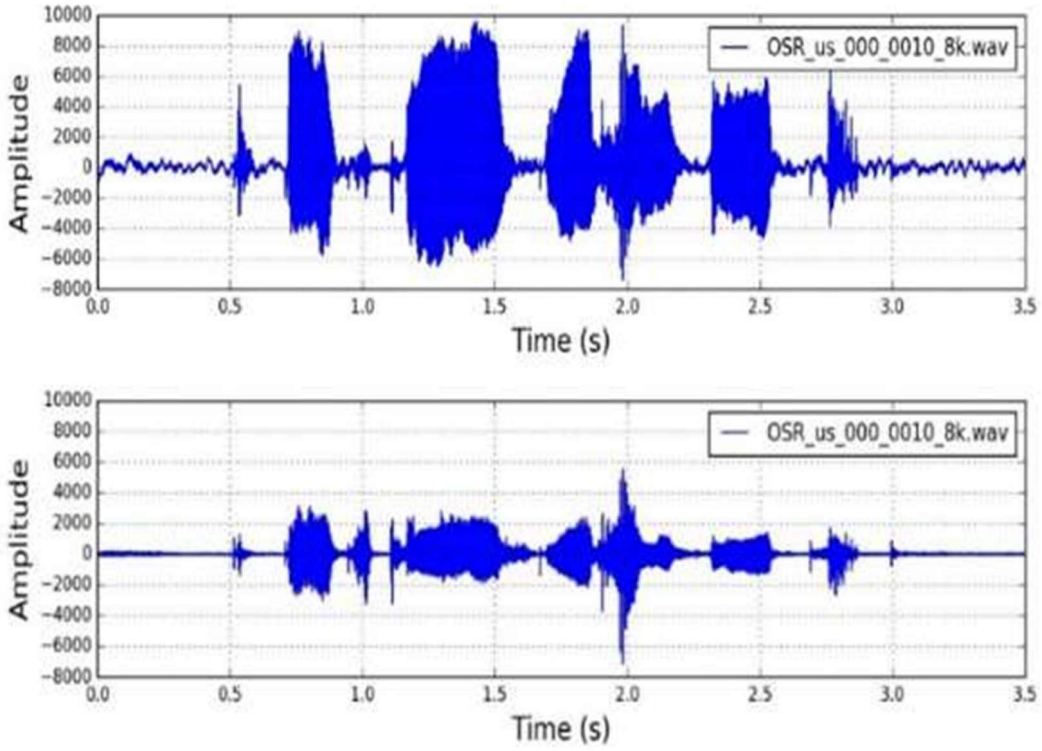


Figure 3. The signal before and after Pre-emphasis

Windowing audio signal:
The main objective of this step is framing the given signal, where the processing divides the audio signal into several sections, each section called a frame, and each frame is processed independently from the rest of the others.

In most cases, the signal cannot be processed simultaneously because this may lead to unsatisfactory results, in the other hand MFCC is based on the Fourier transform, so it is hard to determine the periods (time frame) corresponding to a particular frequency. Thus, slicing is the best solution for calculating frequencies in a semi-local manner, where the values are defined within one frame and correspond to a specific frequency.

Since the audio signal is a continuous signal over time, this continuous signal must be divided to obtain semi-static characteristics as each frame is processed within a short period. The audio signal is processed and performed on a frame ranging from 20 to 25 milliseconds and a frameshifting (overlapping) every 15 milliseconds [8], which allows tracking of the physical properties of the individual speech sound and is short enough to resolve significant temporal characteristics.

The main purpose of the overlapping process is that each frame of the input sequence would be in the region of the center at some frame. During framing and cutting of the signal, if the signal is clipped directly at the edges, the sudden decrease in the amplitude range at the edges will create a lot of noise in the high-frequency field [6]. Therefore, we have to make this cut in a smooth way that prevents noise in the high-frequency field as we cut it using Hamming/Hanning window instead of the rectangular window

Hamming window used for speaker recognition task defined as:

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

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

DFT (Discrete Fourier Transform):
Analyzing and processing the signal is easier in the frequency domain [6], The main objective of this step is to convert the signal from the time domain to the frequency domain (magnitude spectrum) by applying the DFT transform.

Mel-Filter Bank:
The purpose of this step is to compute the Mel spectrum bypassing the signal, which is converted in the last step to a set of band-base filters (several filters) known as the Mel-filter bank shown in the following figure.

The Mel scale is found tostimulate the hearing frequency that a human can hear, how to perceive sound is different between humans and machines where the human ears do not perceive tone linearly, and therefore linear vibration

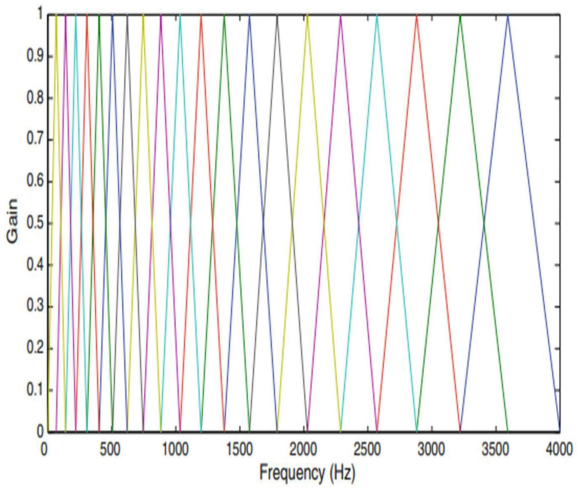


Figure 4. Mel-filter bank

will not be recognized, so the Mel scale is approximately a linear frequency spacing below 1kHz and logarithmic spacing above 1kHz [9].

To determine the actual frequency that a human can recognize, the Mel scale must be used.

For a given frequency (f) in Hz, the following formula for mapping the Mel filterbank:

$$mel(f) = 1127\, ln\left(1 + \frac{f}{700}\right)$$

Automatic speech recognition (ASR) systems normally use a logarithm scale rather than a linear scale. It does so because the logarithm allows the system to use subtraction (the channel normalization) to mimic the human cognition of speech and sound (human hearing system), as humans are less sensitive to changes in the spectrum (audio signal energy) at higher energy compared to lower energy. Because studies showed that, humans recognize sounds on a logarithm scale [10].

IDFT:
The main purpose of this step is to compute the Cepstral that split the glottal source and the filter by inversing the result of the last step by using inverse Fourier transformation to separate the pitch from the formants. The word Cepstral is derived from the spectrum and it is the inverse of the signal's magnitude.

As the spoken words of human speech are inverted at the moment when it transforms from the throat to the tongue so the period in the time domain and frequency domain is inverted, and that lead to the frequency domain with the lowest rate will have the highest frequency in the time domain [6].

The following figure shows the changes in the signal before and after using IDFT:

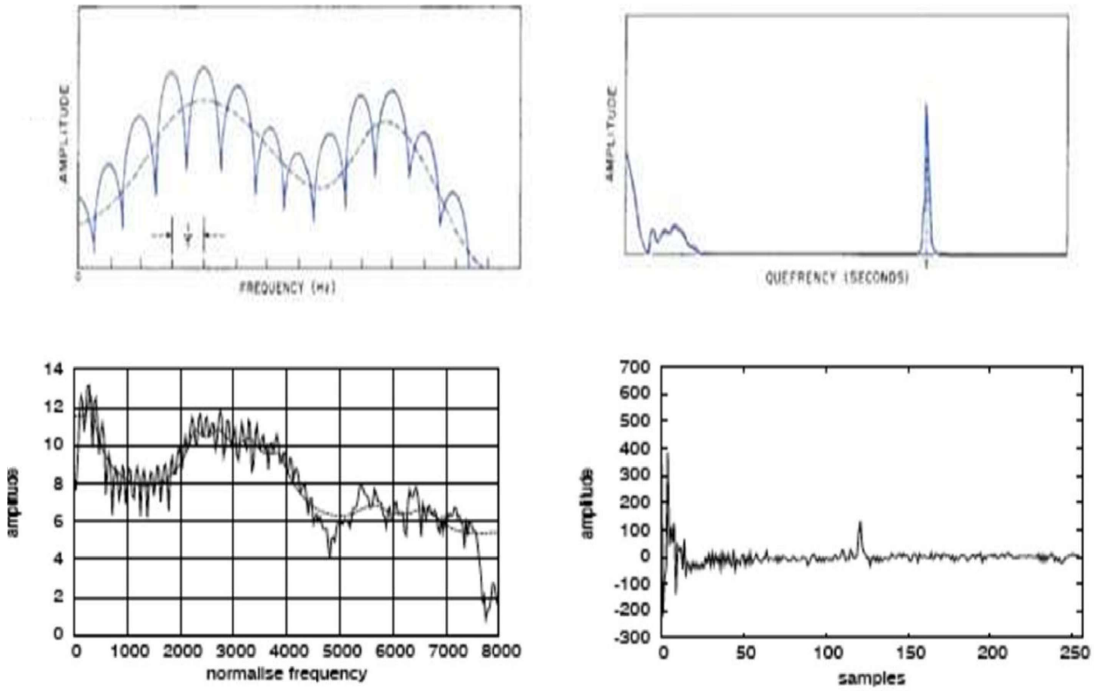


Figure 5. Signal before and after IDFT

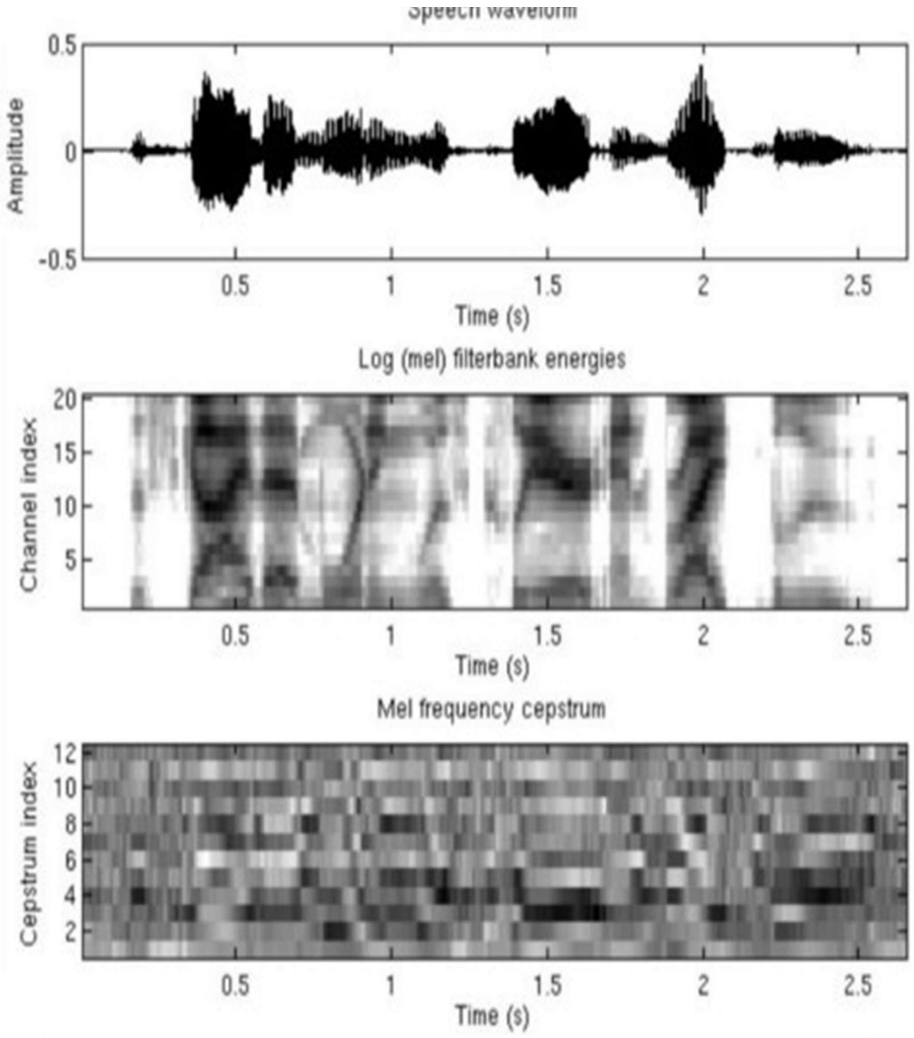


Figure 6. The 12 Cepstral coefficients

Usually for ASR, in the MFCC algorithm, the model only takes the first 12 coefficients of the signal after applying the inverse step (IDFT) and discards the other values. With the 12 Cepstral values, it will take the energy of the signal (Log power spectrum) as a feature [6] and this process is like using the other method called discrete cosine transform (DCT), which will help by identifying the speech and the formula of the energy of the samples is defined as:

$$\text{Energy} = \sum_{t=t1}^{t2} x^2[\text{t}]$$

The following figure shows the 12 Cepstral coefficients (values):

MFCC has 39 features (static features); we looked at the first 13 ones so what is the rest of them? The other features have information about the temporal dynamics of the signal which is calculated by using the first-order (delta coefficients) and second-order (delta-delta coefficients) derivatives of Cepstral coefficients [10]. The first Cepstral coefficients give information about speech rate and the second one gives information about the similarity acceleration of the voice.

The difference in the coefficients between the audio signal samples calculated by performing the first-order and second-order derivatives will help to understand how the transition occurred. The MFCC 39 features will be used later as input for the ASR model [11].

The first and second Cepstral coefficients are defined as:

$$\Delta c_m(n) = \frac{\sum_{i=-T}^{T} k_i c_m(n+i)}{\sum_{i=-T}^{T} |i|}$$

C_m is the feature's number, n is the number of time frame

K_i is the ith weight; T is the computation successive frames number and usually $T = 2$.

The delta-delta coefficients are calculated by computing the first-order derivative of the delta coefficients.

LPC: LPC is a tool used for sound processing to simulate the human vocal apparatus and gives a strong advantage in the speech by focusing and repeating residues from the speech signal after removing the effects of evaluating the audio signal by approximating formulas [12]. LPC is used in voice recognition systems where its main objective is to extract the voice features involved in speech. It is known for its speed and accuracy where it excellently represents stable and consistent source behaviors as it gives accurate estimates of speech parameters and is relatively effective in computation, however, it has a high sensitivity to quantization noise and may not be suitable for generalization [12].

LPC is a powerful method of speech analysis, which has gained fame as an explicit estimation method [13]. This technique is used to predict the positions of the speech signal parameters by calculating the linear predictive parameters above the windowed signal, and finding the peaks in the filter spectrum beyond the linear prediction, the frequencies that occur at the resonance peaks called audio frequencies. LPC is widely used for medium or low bitrate [13].

LPC works in a pattern assuming that the result of each voice sample is shown as direct integration of the previous samples. The difference equation coefficients characterize the formulas; Therefore, LPC needs to approximate these coefficients.

Generally, LPC is used to reconstruct speech. The companies that work in music and electric fields usually used LPC to create sounds and analyze tones and string instruments [12]. Several characteristics can be derived from the LPC algorithm like reflection coefficients (RC), linear prediction cepstral coefficients (LPCC), and line spectral frequencies (LSF) [12].

The algorithm reduces the squared error between the input signal and the estimated speech by applying a linear prediction method to obtain the equivalent filter coefficients for the audio channel. The algorithm analysis of speech signal prediction of any speech sample during a specified period as a weighted linear aggregation of the prior samples [12], there are many advantages of using LPC:

- Better approximate coefficient spectrum.
- The shorter and more efficient time calculation for signal parameters.
- Get important characteristics of the input signals.

To understand the logic behind LPC, we must first understand the Auto-regressive (AR) model of speech, which can be modeled as pth order AR process, where each sample is given by:

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

The LPC works by linking all samples where the current samples depend on the previous samples added with Gaussian noise u(n), in other words, each sample at nth process step depends on the last 'p' sample, this model comes from the assumption that When we speak and make sounds, a buzzer is produced at the end of the tube which generates a vibration that can be heard by the human ear (sounds) and this is the speech signal, with occasional added background noise.

In the equation above (a) represents the LPC coefficients

where the Yule-Walker which connects auto-regressive parameters to auto-covariance for a random process at X [16], and it is used to estimate these coefficients by using the auto-correlation function (ACF) Rx, the correlation at each lag is scaled by the sample variance by using the Box-Jenkins method while calculating the ACF, at lag l the auto-correlation is given by:

$$R(l) = \sum_{n=1}^{N} x(n)x(n-l)$$

The Yule-Walker equation final form:

$$\sum_{k=1}^{n} a_k R(l-k) = -R(l)$$

$$\begin{bmatrix} R(0) & R(1) & R(n-1) \\ R(1) & R(0) & R(n-2) \\ R(n-1) & R(n-2) & R(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_n \end{bmatrix} = - \begin{bmatrix} R(1) \\ R(2) \\ R(n) \end{bmatrix}$$

Where (α) is given by:

$$a = -R^{-1}r$$

In the case of LPC, the normalized LPC coefficients estimated at a range lie between [-1, 1] to give more accurate results where the speech at the beginning is divided into small frames of 20 to 25 milliseconds with 15 milliseconds of frameshifting (overlapping) as described earlier in the preprocessing steps in the previous section.

for one speech frame, the auto-correlation is shown in the figure below:

The linear combination that defines the values of the signal is expressed as:

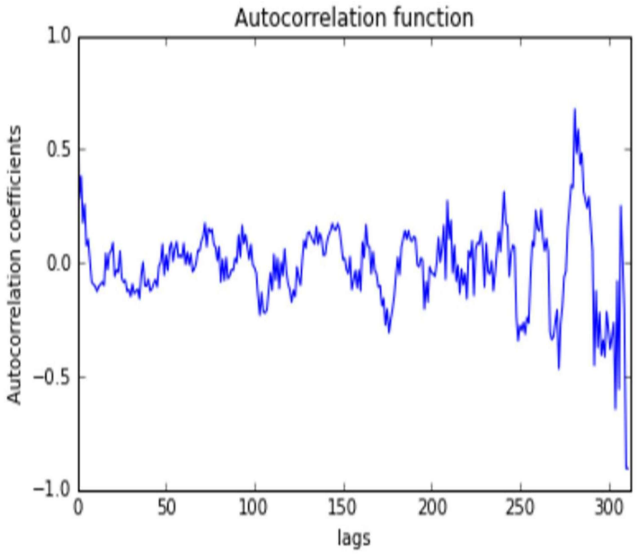


Figure 7. ACF of LPC processing for one speech frame

$$z(i) = x1 * z(i\text{-}1) + x2 * z(i\text{-}2) + \ldots + xp * z(i\text{-}p)$$

where: z(i) is the amplitude at the time i.

p is 10 for normal LPC, and 12 for improved LPC [14].

As described in the section when a speech is given to the model, it becomes a series of linear combinations of several previous samples. that can be solved to determine the values of $x1$ to xp where it is used to produce the signal after clearing it from noise to identify the speaker [15]. The model is based on the process of automatically linking the entire windowed signal where the highest value of the auto-correlation is the order of the linear prediction analysis, followed by the analysis of the algorithm where each frame of the windowed signal is converted into LPC parameters that contain the LPC coefficients [14].

The following figure shows the procedure for the algorithm:

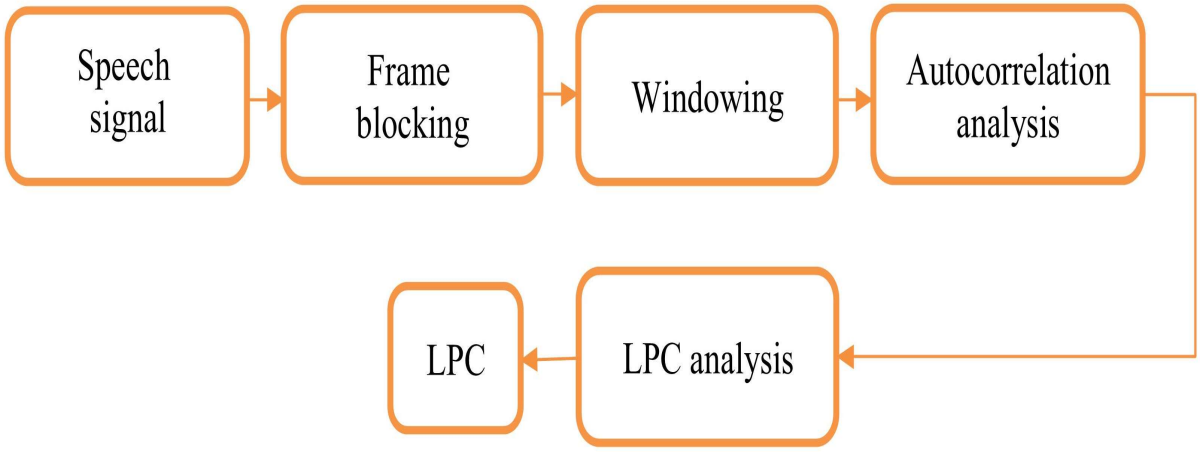


Figure 8. Block diagram of LPC processing

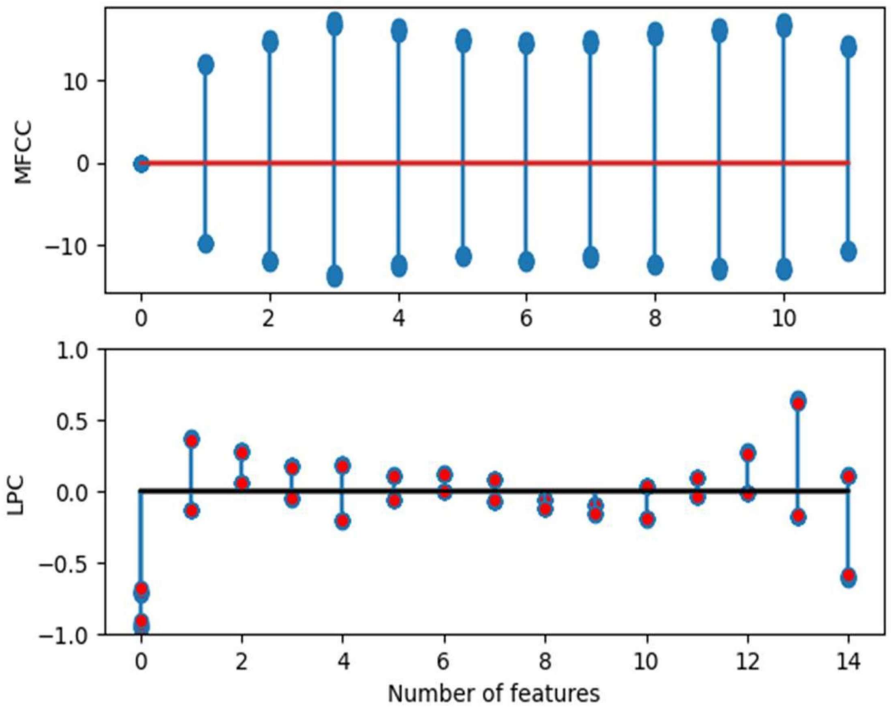
Test and Result:
The main objective of this stage is to test both methods described earlier in this paper that had been used for feature extraction of a given voice. Where the test stage is based on two steps the first one is to enroll the voice of the user into the system, the second one is to compare the given voice with the stored one in the system. The system recorded a database of 50 speakers while they are speaking, the recording phase worked under a 22050-sampling frequency and 8 bits per sample. The model used a 20 Mel filter bank, and the test phase was made using 10, 20, and 30 filters the result is as follows:

Mel filter banks			
No of filters	10	20	30
Error	32	16	16

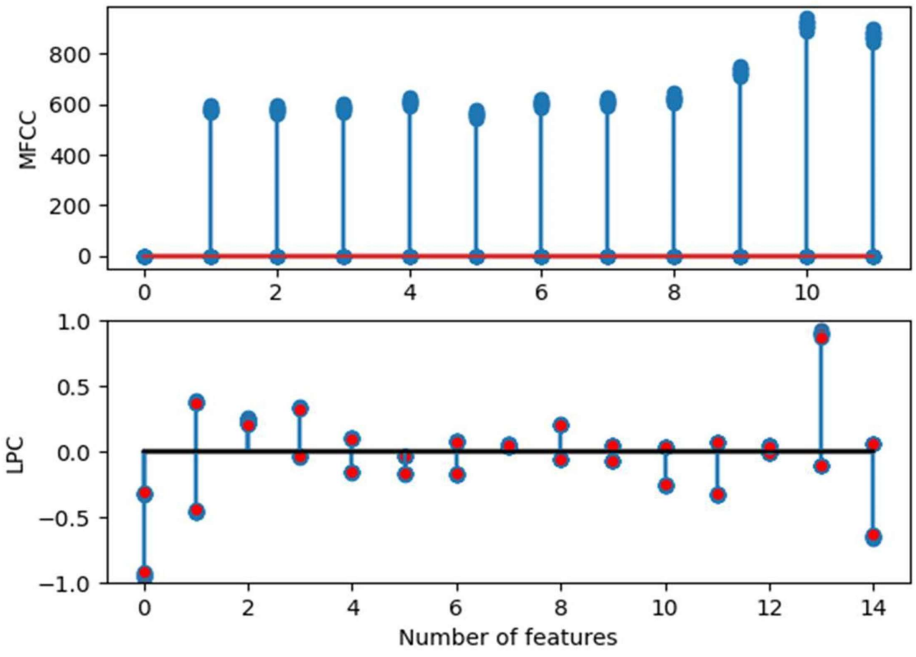
Table 1. Performance comparison with several filters in the model

Figure 9 shows the voice samples and features selection for some users in the system in the training phase:

Speaker 7



Speaker 8



Speaker 9

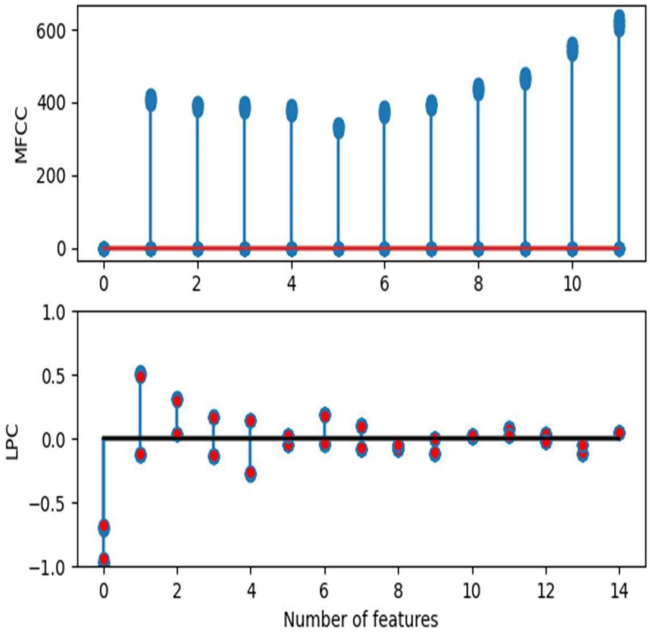


Figure 9. Voice samples and features

Some epochs of the training and testing phases for some speakers:

Speaker	**Identified as (MFCC)**	**Identified as (LPC)**
S 1 S 2 S 3 S 4 S 5 S 6 S 7 S 8 S 9 S 10	S 1 S 2 S 3 S 4 S 5 S 6 S 7 S 8 S 9 S 10	S 1 S 2 S 3 S 2 S 9 S 2 S 2 S 8 S 9 S 8
#Speaker = 10	Accuracy = 100%	Accuracy = 50%

Table 2. Verified speakers for some epochs in the model

#speakers	Error rate (MFCC)	Error rate (LPC)
8	0%	50%
10	0%	50%
15	13.4%	40.1%
20	14%	45%
25	14.3%	48.2%
30	14.9%	48.8%
35	15. 72%	49.47%
40	16.51%	49.79%
45	16.72%	49.82%
50	17%	50%

Table 3. Overall accuracy in the training and testing phase

As shown in Table 2, in the first row speaker1 is verified by MFCC and LPC, but speaker 7, for example, is identified by MFCC but not by LPC. So the total speakers that are identified by MFCC are 10 out of 10 but for LPC is 5 out of 10, with an accuracy of 100% for MFCC and 50% for LPC.

Table 3 shows the accuracy rate between both methods (MFCC and LPC) for the same number of speakers:

Figure 10 shows the centroid distribution of the user's voice samples that had taken in the training phase in the system for some speakers:

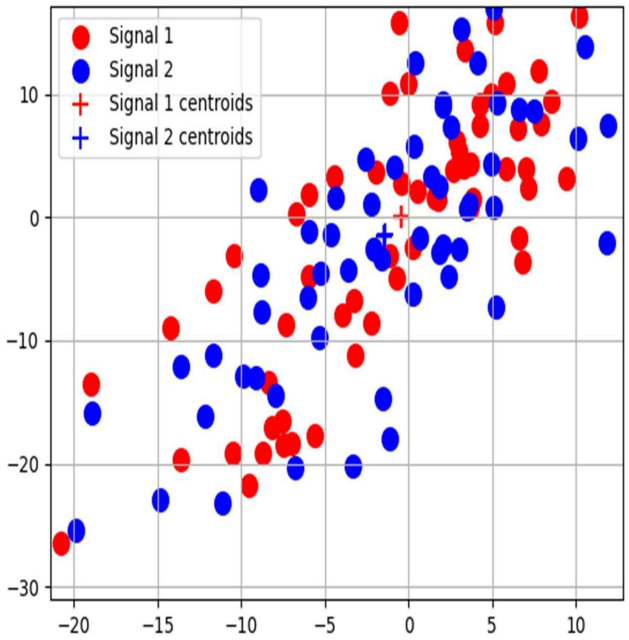


Figure 10. Centroid distribution for some speakers

The system achieved an overall accuracy of 81% percentage using the MFCC method by using a database of 50 speakers and then comparing it with the normal algorithm of LPC where the system got a 52% percentage.

Conclusion

In this paper, we invest the effect of using features extraction methods for the ASR system that explore two approaches and compare the results for each of them, where the model is inducted on several recorded audio files for several speakers, both methods trained by the same speeches and files. According to the experimental results, MFCC was better than LPC by recognizing the speakers and that may be because of the way it works where it uses a logarithmic scale because the logarithm allows the system to use subtraction (the channel normalization) to mimic the human cognition of speech and sound (human hearing system) and that is the normal hearing way for humans. The overall accuracy for MFCC was 81% and for LPC was 52%. For future works to improve the model a complex classification algorithm like ANN, DNN and SVM should give a better result where it could be added to train the model on the speaker speeches where the MFCC works for features extraction and the classification algorithm do the job of identifying the speaker after the training and testing phases.

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 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 at Syrian Virtual University.

References

[1] Poddar, A., Sahidullah, Mhd., Saha, G. (2018) Speaker verification with short utterances: A review of challenges, trends, and opportunities. IET Biometrics, 7, 91–101 .

[2] Aizat, K., Mohamed, O., Orken, M., Ainur, A., Zhumazhanov, B. (2020) Identification and authentication of user's voice using DNN features and i-vector. Cogent Engineering, 7.

[3] https://en.wikipedia.org/wiki/Speaker_recognition. Wikipedia.

[4] https://espanol.verizon.com/articles/speech-recognition-technology/.

[5] Pew Research Cntr (2017) Voice assistants topline and methodology. https://www.pewresearch.org/fact-tank/2017/12/12/nearly-half-of-americans-use-digital-voice-assistants-mostly-on-their-smartphones/.

[6] Kiran, U. (2021) MFCC techniques for speech recognition. https://www.analyticsvidhya.com/blog/2021/06/mfcc-technique-for-speech-recognition/,June13.

[7] Rao, K.S., Manjunath, K.E. (2017) "Speech Recognition Using Articulatory and Excitation Source Features", SpringerBriefs in speech. Technology.

[8] Benesty, J., Sondhi, M.M., Huang, Y.A. (2008). Handbook of Speech Processing. Springer: New York, USA.

[9] Karpov, E. (2003). Real-Time Speaker Identification [The University of Joensuu, Department of Computer Science Master's Thesis].

[10] Niemann, H. (2013) Klassifikation von mustern. Available from: https://www.springer.com/de/book/9783540126423. Springer-Verlag.

[11] Rabiner, L., Juang, B.-H., Yegnanarayana, B. (2008). *Fundamentals of Speech Recognition*. Pearson Education: London.

[12] Alim, S.A., Rashid, N.K.A. (2018). *From Natural to Artificial Intelligence – Algorithms and Applications*.

[13] Buza, O., Toderean, G., Nica, A., Caruntu, A. (2006) Voice signal processing for speech synthesis, *In*: IEEE International Conference on Automation, *Quality and Testing Robotics*, Vol. 2, pp. 360–364.

[14] Shrawankar, U., Thakare, V. (2013). Techniques for Feature Extraction in Speech Recognition System: A Comparative Study, [arXiv:1305.1145v1].

[15] Kurzekar, P.K., Kurzekar, P.K., Waghmare, V.B., Shrishrimal, P.P. (2014) A comparative study of feature extraction techniques for speech recognition system. *International Journal of Innovative Research in Science, Engineering and Technology*, 3.

[16] Mathuranathan Viswanathan, Y.W. (2014) Estimation and simulation in MATLAB. Found at: https://www.gaussianwaves.com/2014/05/yule-walker-estimation.