Celal Bayar University Journal of Science
Volume 15, Issue 3, 2019, p 287-292
Doi: 10.18466/cbayarfbe.556936

Markov Model Based Real Time Speaker Recognition using K-Means, Fast Fourier Transform and Mel Frequency Cepstral Coefficients

Emin Borandağ*

Department of Software Engineering, Faculty of Technology, Manisa Celal Bayar University, Manisa, Turkey
*emin.borandag@cbu.edu.tr

Received: 22 April 2019
Accepted: 16 September 2019
DOI: 10.18466/cbayarfbe.556936

Abstract

In this study, which was carried out using a combination of machine learning and sound processing methods, a speaker recognition system and application were developed using real-time Mel Frequency Cepstral Coefficients (MFCC) features and Markov chain model classifier. A sound sample was taken from each speaker for the training of the system and these sound samples were processed in Fast Fourier Transform and MFCC feature extraction algorithms. The MFCC features were clustered using the k-means clustering algorithm. A Markov chain model was created for each speaker by using the outputs obtained after clustering. By deducting the characteristic features of the voice of the speaker, the person who was talking in the society and how long and at which time intervals they spoke during the conversation was determined in real time with high accuracy.

Keywords: Real time speaker recognition, mel-frequency, k-means, machine learning, Markov chain, fast fourier transform.

1. Introduction

Automatic Speech Recognition (ASR) technologies, which convert acoustic sound signals into forms that can be read by the computer, have been used for a long time for voice recognition [1]. These technologies allow speakers' words to be transferred to a computer environment as text or by assessing the characteristics of the speaker's voice, it is possible to estimate who is speaking. In this context, speech recognition is divided into two: speaker identification and speaker verification [2]. Various commercial software such as Amazon (Alexa Voice Services) [3], Kaldi ASR [4], Dragon Speech Recognition Solutions [5], Google Voice [6], and Sphinx 4 [7] are available in this area. These software process the human voice with Feature Extraction and Feature Matching methods.

Over the past 20 years, with the development of technology, various academic studies have been conducted on voice recognition. In a study conducted by Douglas A. Reynolds (1995) titled "Automatic Speaker Recognition Using Gaussian Mixture Speaker Models", the Mel Frequency Cepstral Coefficients (MFCC) feature, which is based on the characteristics of the human voice, was used. In his study to obtain the MFCC feature, the sound was divided into proceeding tracks of a length of approximately 20 ms with 10 ms increments, and parts with no human voice or any noise were removed, only then, the MFCC feature vectors were obtained using the MFCC algorithm. These sequences were used in the training of models by the Gaussian Mixture Model (GMM) method. In the test process, the feature vectors of the tested sounds were compared with each model to determine the highest probability model. The accuracy rates varied according to the number of speakers and the database where the sounds were received. The accuracy rates ranged between 78% and 100% for 100 speakers [8].

In a study conducted by Tahira (2015) speaker identification was determined by using GMM with the MFCC Methods in addition to Framing, Fourier Transform (FTT), Vector Quantization (VQ), Hidden Markov Model (HMM). The k-means clustering algorithm was used for the VQ step. The hidden Markov models were trained using the outputs of this clustering algorithm. A total of 20 sound samples (14 training and 6 test samples) per speaker were used for 40 speakers. An accuracy rate of 100% was obtained for up to 4 people in the test stage. It was found that as the number of speakers increased, the success rate decreased to 80% [9].

In a study carried out by Roman Bharti et al. (2015), titled "Real Time Speaker Recognition System using MFCC and VQ technique", they used a real time speaker recognition system using MFCC features based on human voice characteristics. The application was developed in a MATLAB environment. A total of 120 speakers' data from TIDIGIT was used for testing and a 91% accuracy was acquired [10].

In a study conducted by Srivastava (2016) speaker identification was determined by using phase based

Celal Bayar University Journal of Science
Volume 15, Issue 3, 2019, p 287-292
Doi: 10.18466/cbayarfbe.556936

mel-frequency cepstral coefficients (MFCCs). Gaussian mixture model (GMM) based voice conversion with vector quantization was used as the definer for the classification of the speakers. The performance of the PMFCC features were found to be better than the MFCC features and phase features [11].

In a study carried out by Chandar Kumar et al. (2018), titled "Analysis of MFCC and BFCC in a Speaker Identification System", GMM based MFCC and bark frequency cepstral coefficients (BFCC) speaker identification system were analyzed in terms of identification rate. They determined that the GMM based MFCC was the optimum feature extraction technique compared to the BFCC [12].

Jadhav et al. (2018), proposed a speaker recognition system using GMM. This system used the voice recordings of 28 people and achieved an accuracy rate of 96.42% [13].

Speaker recognition is used for identification of a person from characteristics of voice. The present study aimed to predict the speaker who is speaking at the moment and give information about the length and time of speech. This application that was developed using machine learning and sound processing algorithms consists of a system based on the extracts of the voice characteristic properties of a person and used this data in machine learning training. Voice samples of every speaker in the group were taken and a model was generated for each speaker. Using these models, predictions were made regarding which speaker spoke at what time. The application took the training sound samples and the dialog sounds for prediction as input and gave graphical output about which speaker spoke when and for how long.

In the second section of this study, information about the methods and algorithms used in the voice recognition system is given. In the third section, the details of the real time speaker recognition system are explained and the developed system is described. In the fourth and final section, the obtained results are given.

2. Materials and Methods

In this section, the materials, algorithms and methods used in the present study are given.

2.1. Markov Chain Model

The Markov chain model is a graph with nodes that represents states and edges which represent transitions between states. In the Markov chain model, a state can change based on a statistical value or can stay without a change. Past states do not have effect on current state but current state can affect future states. Since Markov models are statistic, it's possible to represent every stochastic event's probability [14].

2.2. Fast Fourier Transform

Fast Fourier Transform (FFT) is a statistic-based mathematical process used in vibration analysis. It separates mixed signal clusters and shows the severity and frequency of a vibration. The audio file is used to convert wave values into frequency size. An example of wave to frequency spectrum transformation is given in Figure 1 [15].

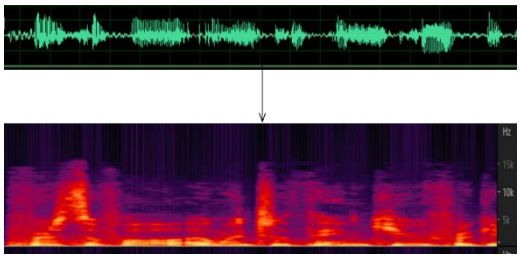

Figure 1. Wave to Frequency Spectrum [15].

2.3 MFCC Feature Extraction

In sound processing, the mel-frequency cepstrum (MFC) represents the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency. Mel-frequency cepstral coefficients (MFCCs) are coefficients that collectively make up an MFC. The MFCC processes frequency values take into account the sensitivity of human perception and is therefore widely used in speech/speaker systems [16].

2.4. K-Means

K-means, a machine learning clustering algorithm, was used to aggregate the MFCC features. In classification, it is the process of determining the classification of objects by considering their features. The k-nearest neighbors (KNN) algorithm is one of the most known algorithms among the machine learning algorithms. Classification is made using the proximity of a selected feature to the feature nearest to it. The formula in Equation 1 was used in the determination of distances between objects. The K value in the equation is considered primarily when a new object that must be defined according to the defined data comes. Methods such as Cosine, Euclidean or Manhattan distance are used when calculating the distances between the incoming data and the other data [17].

$$d_{(i,j)} = \sqrt{\sum_{k=1}^{p} (X_{ik} - X_{jk})^2} \qquad (1)$$

3. Developed System

The system that was developed within the scope of this study is composed of 2 main parts: model training and speaker recognition.

Celal Bayar University Journal of Science
Volume 15, Issue 3, 2019, p 287-292
Doi: 10.18466/cbayarfbe.556936

3.1. Model Training

As seen in Figure 2, the modeling of the voice was performed in 6 sub-stages in total. Detailed information about these steps is given in the sections below.

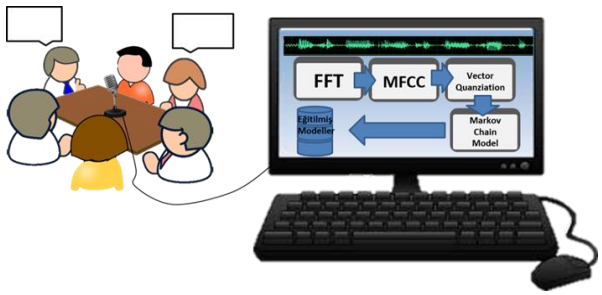

Figure 2. Model Training.

3.1.1. Extraction of Sound Wave Values

All bytes of each wave file were read and the sample rate, bits per sample, number of channels, subchunk2size values of the files were gathered and these values were used to read the data in the data chunk.

3.1.2. Splitting Sound Data into Frames

In order to perform the frequency analysis on the gathered sound data, the sound data must be split into small pieces. There is no specific rule regarding the length of these pieces. In this study, the frame lengths were chosen as 1024.

3.1.3. Frequency Analysis using Fast Fourier Transform

FFT is a faster version of Discrete Fourier Transform (DFT). Fourier transform is used to represent a wave in the frequency domain and calculated with the following equation [18]:

$$X_k = \sum_{n=0}^{N-1} X_n . e^{-i2\pi kn/N}$$
$$= \sum_{n=1}^{N-1} X_n . \left(\cos(\frac{2\pi kn}{N}) - i.\sin(\frac{2\pi kn}{N})\right), \quad (2)$$

3.1.4. MFCC Feature Extraction

In sound processing, the mel-frequency spectrum represents a short-timed power spectrum of voice. MFCC processes sounds considering human ear's sense which why it's commonly used in speech/speaker recognition systems. The sample of mel-frequency filter banks is given in Figure 3.

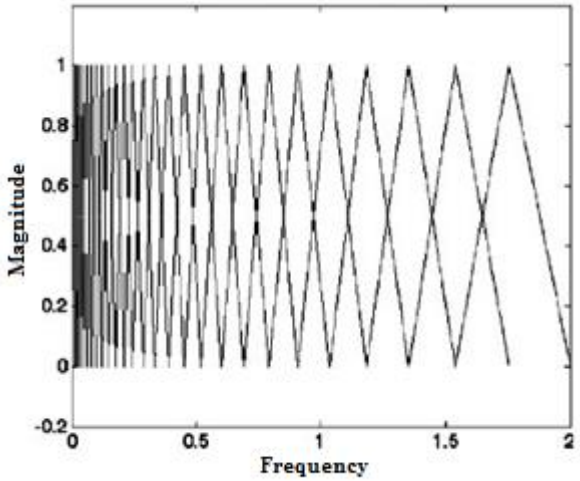

Figure 3. Mel Frequency Filter Banks [19].

3.1.5. Vector Quantization (K-Means Clustering)

When MFCC vectors of a sound sample are combined, they make a 2D matrix and when these matrices are combined for each speaker, they make a 3D dataset. It's difficult to use this dataset in training. For this reason, a vector quantization step is applied. In this step, the MFCC vectors are sent to the k-means clustering algorithm ad data set and cluster assignments are collected from the algorithm. The VQ sample is given in Table 1.

Table 1. Sample of Vector Quantization.

	Frame 1	**Frame 2**	**Frame 3**	**…**
Voice Sample 1	Set_5	Set_7	Set_8	…
Voice Sample 2	Set_23	Set_2	Set_13	…
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.

3.1.6. Generating Sound Model with Markov Chain Model

Every cluster assignment from the vector quantization step represents a frame that is a 11.6 millisecond part of a sound. Naturally, a cluster assignment vector makes a time based sequence. The cluster assignment transitions in these sequences are used to make Markov chain model. The Markov chain pattern probability model equation is given in Equation 3 below [20].

$$A_{jk} = \frac{\sum_{n=2}^{N} Z_{n-1}, j^{Z_{nk}}}{\sum_{l=1}^{K} \sum_{n=2}^{N} Z_{n-1},\ j^{Z_{nl}}} \quad (3)$$

When creating a transition probability matrix from a sequence, the formula shown in Equation 2 was used. The equation states the possibility of transition from

Celal Bayar University Journal of Science
Volume 15, Issue 3, 2019, p 287-292
Doi: 10.18466/cbayarfbe.556936

state j to k. This calculation was made for the transition probabilities between all situations. As a result, a state transition probability matrix was obtained. The model creation process was performed for each speaker and thus, as the output, the audio model was obtained up to the number of speakers. The pseudo code of the study is given below.

```
Input
Defination wave_file[], data_extraction[], Fft[],Mfcc[],
Vq[],Peoples[],i,voice;
Training Process
For i =1 to Peoples[].Count
//Get Wave files for every voice samples
    wave_file[i] = Sys.microphone;
    data_extraction[i]= Split wave_file[i];
// Feature extraction for every voice samples
   Ftt[i] = Transformation(data_extraction[i])
   Mfcc[i] = Extraction( Ftt[i])
   Vq[i]= Quantization(Mfcc[i])
End For
Prediction Process
While (not at end of voice stream) do
     If KnnBasePrediction (Sys.voice) = People[1]
       "People 1 is speaking"
     Else if KnnBasePrediction (Sys.voice) = People[2]
       "People 2 is speaking"
     ...
     Else if "Voice is not recognized"
End While
```

3.2. Prediction Step

The details of the prediction step are given in Figure 4. A sound training data was submitted to the system. This sound data can be a WAVE file or real time sound data from a microphone. The system applies the first 5 steps of the training process which are Sound Wave Data Extraction, Splitting Data Into Frames, Fast Fourier Transform, MFCC Feature Extraction and VQ to this sound data. At this step, all of the 81 consecutive frames are grouped and sent to the k-means model that is stored in memory at training step. As output cluster vectors are obtained. The probability of these vectors were computed for each speaker Markov model. System predicts the model with the highest probability as the owner of the sound.

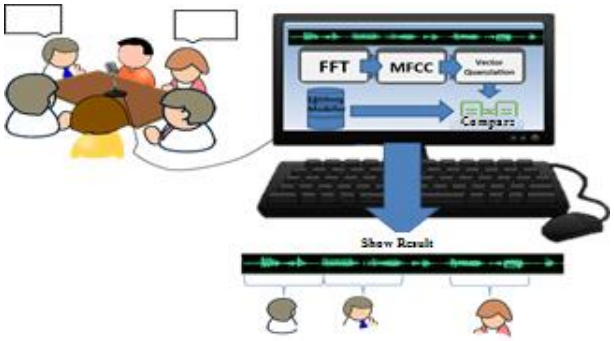

Figure 4. Prediction Step.

3.2.1. Training Screen

The profiles list is a list containing all the profiles created and saved in the system. User can create a new profile by clicking on the “New Profile” button or edit a selected profile by clicking the “Edit Profile” button. By clicking the “Delete Selected Profile” button, user can delete the selected profile. In order to train the system, user should move profiles into the Training Profiles list. The system was able to initiate audio data reception after training was carried out. It determined the speakers by using the training data. Information related to who speaks at what time interval can be displayed on the sound wave graph, while the speech intensity information can be displayed on the pie graph. The screenshot from the software developed is given in Figure 5.

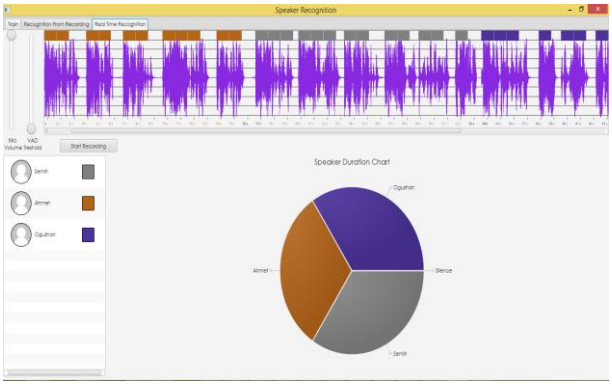

Figure 5. Prediction on Sound File Screen.

3.2.2. Accuracy Performance

When creating a transition probability matrix from a sequence, the formula shown in Equation 2 was used. This calculation was made for the transition probabilities between all situations. As a result, a state transition probability matrix was obtained. The model creation process was performed for each speaker and thus, as the output, the audio model was obtained up to the number of speakers. The accuracy performance chart is given in Figure 6.

Celal Bayar University Journal of Science
Volume 15, Issue 3, 2019, p 287-292
Doi: 10.18466/cbayarfbe.556936

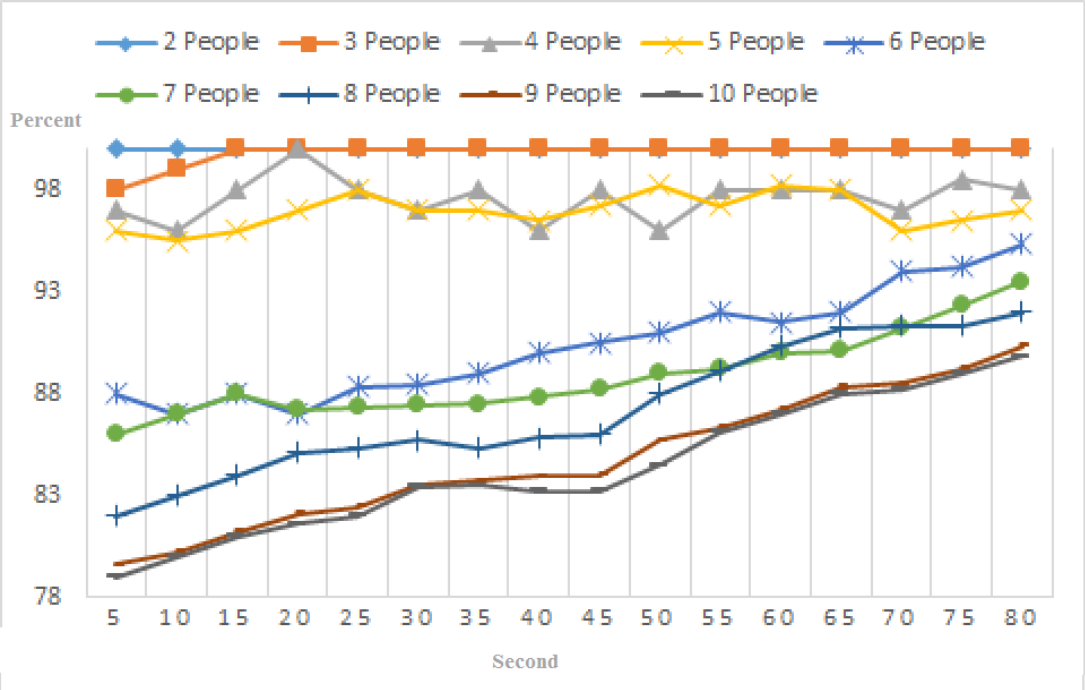

Figure 6. Accuracy Performance Chart by Number of Persons.

Table 2. Test Data

Test Data					
2 People	2 Male	**3 People**	2 Male 1 Female	**4 People**	2 Male 2 Female
5 People	3 Male 2 Female	**6 People**	4 Male 2 Female	**7 People**	4 Male 3 Female
8 People	5 Male 3 Female	**9 People**	6 Male 3 Female	**10 People**	7 Male 3 Female

For the accuracy performance test, sound data from noise-free interviews conducted with 10 people were used. Information regarding the number and sex of the people in the test data are given in Table 2. In the graph displayed in Figure 6, the values in the horizontal axis give the number of sets in the VQ step, and the values in the vertical axis are the percentages of accuracy. According to the laboratory tests, the accuracy rate of prediction for the system is up to 90% when it is tested with 2 to 10 speakers. Voice estimation conducted with two speakers shows a 100% success.

4. Conclusion

With the current state of the system, real time speaker recognition can be made in noiseless environments and speaker predictions and speaker speech density can be visualized once the system has been trained. The accuracy rate of prediction for the system was determined to be up to 90% when it was tested with 2 to 10 speakers. It is planned to research different model creating methods, focus on different feature extraction techniques and improve voice activity detection approaches and graphical user interface to increase the accuracy rates in the future studies.

Author's Contributions

Emin Borandag drafted and wrote this manuscript and performed the experiment and result analysis.

Ethics

There are no ethical issues raised by this study.

References

1. Khosravani A, Homayounpour M, 2017. A PLDA approach for language and text independent spaker, *Computer Speech & Language;* 1(1):457-474.

2. Hana H, Baeb KM, Honga SK, Parkb H, Kwakd JH, Wanga HS, Joea DJ, Parka JH, Junga YH, Hurc S, Yoob CD, Lee KJ, 2018. Machine learning-based self-powered acoustic sensor for speaker recognition. *Nano Energy*; 658-665.

3. Alexa Voice Service, Alexa Voice Information Report. https://developer.amazon.com/alexa-voice-service (accessed at 26.01.2019).

4. Asas Kaldi's code. http://kaldi-asr.org/ (accessed at 26.01.2019).

5. Dragon Speech Recognition Solutions, Information Web. https://www.nuance.com/dragon.html (accessed at 26.01.2019).

6. Google Voice. https://www.google.com/voice (accessed at 26.01.2019).

7. Open Source Speech Recognition Toolkit. https://cmusphinx.github.io/ (accessed at 26.01.2019).

8. Reynolds A, 1995. Automatic speaker recognition using Gaussian mixture speaker models, *The Lincoln Laboratory Journal*.

9. Mahboob T, Khanum, M, Sikandar M, Khiyal H, Bibi R, 2015. Speaker Identification Using GMM with MFCC, *IJCSI International Journal of Computer Science;* 2.

10. Bharti R, Bansal P, 2015. Real Time Speaker Recognition System using MFCC and Vector Quantization Technique.

11. Srivastava S, Chandra M, Sahoo G, 2015. Phase Based Mel Frequency Cepstral Coefficients for Speaker Identification, Springer India*;* 1(1):57-64.

12. Kumar C, Rehman F, Kumar S, Mehmood A, Shabir G. Analysis of MFCC and BFCC in a speaker identification system, International Conference on Computing, Mathematics and Engineering Technologies (iCoMET), 2018, 174-179.

13. Jadhav A, Dharwadkar N, 2018, A Speaker Recognition System Using Gaussian Mixture Model, EM Algorithm and K-Means Clustering, I.J.Modern Education and Computer Science 11(1):19-28.

14. Hunter J, 2018. Kemeny's function for Markov chains and Markov renewal processes. *Linear Algebra and its Applications;* (559):54-72.

15. Strain J, 2018. Fast Fourier transforms of piecewise polynomials. *Journal of Computational Physics;* (373):346-369.

16. Jokinena E, Saeidia R, Kinnunenb T, Alkua P, 2019. Vocal effort compensation for MFCC feature extraction in a shouted versus normal speaker recognition task. *Computer Speech & Language*; 53:1-11.

17. Ismkhan H, 2018. K-means: An iterative clustering algorithm based on an enhanced version of the K-means. *Pattern Recognition*; (79):402-413.

18. Tan, L, Jiang, J. Chapter 4 - Discrete Fourier Transform and Signal Spectrum, Digital Signal Processing 3th edn. Fundamentals and Applications, 2019, 91-142.

19. Taherisadra M, Asnania P, Galsterb S, Dehzangib O, 2018. ECG-based driver inattention identification during naturalistic driving using Mel-frequency cepstrum 2-D transform and convolutional neural networks. *Smart Health*, (9):50-61.

20. Breena J, Crisostomib E, Faizrahnemoonc M, Kirklanda S, Shorten R, 2018. Clustering behaviour in Markov chains with eigenvalues close to one. *Linear Algebra and its Applications*, (555): 163-185.