AMSE JOURNALS-AMSE IIETA publication-2017-Series: Advances B; Vol. 60; N°3; pp 581-592
Submitted Nov. 08, 2017; Revised Dec. 26, 2017; Accepted Jan. 09, 2018

MFCC for Voiced Part Using VAD and GMM Based Gender Recognition

Hema Kumar Pentapati, Srinivas Vasamsetti, Madhu Tenneti

Department of ECE, SIET, Narasapur, W.G.Dt, Andhra Pradesh, India
(hemapentapati@gmail.com; srinivas.siet@gmail.com; tennetimadhu@yahoo.com)

Abstract

For many applications, identifying the gender information of a speaker is important. In this paper, we implemented the system which identifies the speaker and also gender of the speaker by using MFCC and GMM in an uncontrolled environment. In this text independent system, we aim on the classification using GMM for the extracted features using MFCC and also the speech signal is processed with Voice Activity Detector (VAD). In the experiments using locally recorded database, the system without voice activity detector (VAD) does not provide accurate results. So, the main aim of this paper is to develop a text independent speaker identification and also gender identification using MFCC along with VAD and GMM which improves the performance further relatively when compared with the system without VAD. The performance of the proposed system tested for 70 speakers with 100 percent recognition rate is achieved based on the log likelihood scores.

Key words

Mel frequency cepstral coefficients (MFCC), Vector quantization, Gaussian mixture model (GMM), Voice Activity Detector (VAD).

1. Introduction

One of the biometric applications is speaker authentication that can verifies the unknown's voice. If the gender of the speaker is known, that can be useful in many fields and also it is possible to define gender specific models. So that, the gender dependent models give more accurate results than the gender independent models due to the reduction of number of speakers to model for each class. The motive of this System is to identify the speaker as well as the gender

of the person from MFCC features through modelling with GMM which can be useful in most security applications.

The drawback in the existing system is its classification of speakers which effects the accuracy in identifying the gender. The process of identifying the gender here in this system is improved by using VAD before extracting the speaker specific features in the MFCC feature extraction phase. For this, the non-parametric model like vector quantization may not give the accurate results. So, the GMM which is a parametric model for speaker identification is the empirical observation that a linear combination of Gaussian basis function is capable of representing a large class of sample distributions. we recorded the speech signals in adverse environment but in the same matched conditions for all speakers. If it is in mismatched conditions, this leads to degrade in the performance of the system drastically. In this study, we analysed the gender recognition for large set of speakers which are recorded in the same environment. It gives almost 99 to 100 percent success rate.

Also it is developed as a text independent system using MFCC and GMM such that it does not depend on the particular text to be spoken in training and testing phases and also to work with large data set with large number of speakers such that it uses the log likelihood detector algorithm for making decision of accepting or rejecting the speaker at the final stage. These experiments are performed in different environmental states such as class rooms, office etc. Moreover, the GMM does not work properly under the mismatch conditions of training and testing. So, developing the speaker recognition system in real time through microphone is the major challenging task. The complete system involves MFCC and GMM model is developed through MATLAB.

The complete paper describes about the survey of literature on speaker recognition system using GMM in the next section. Also in the next subsequent sections, it deals about the importance of Voice Activity Detector (VAD). Thereafter, the description about GMM modelling with the proposed system and its implementation. Finally, described about the experiments conducted and the performance evaluation with results.

2. Review of Literature

S.G. Bagul, R.K.Shastri the paper was based on statistical model like Gaussian Mixture Model (GMM) and Features extracted from the speech signals (MFCC) and they concluded with FrFT based algorithms are better compared to Mel scale based methods and Gaussian mixture models provides robust speaker recognition,also computational less expensive in real time.

Ergün Yücesoy, Vasif V.Nabiyev were proposed a paper for automatic identification of gender information from a speech signal for the development of gender dependent models which gives more accurate results than gender independent models. it also focuses on the effect of number of mixture models used and its effect on the performance of the system.

Rania Chakroun, Leila Beltaïfa Zouari1, Mondher Frikha, and Ahmed Ben Hamida were proposed reduced feature vector employing new information detected from the speaker's voice for performing text-independent speaker verification applications using GMM. They concluded that this will decrease the error rate and avoided the complicated calculations and gives the better results compared to baseline systems with GMM Models.

Sourjya Sarkar and K.Sreenivasa Rao proposed "Speaker Verification in noisy environment using GMM super vectors". They uses combined approach of GMM- SVM. For that hybrid systems, the result for the noisy data was studied and they seen significant improvement in performance of the system was observed.

M.S.Sinith, AnoopSalim, GowriSankar K, Sandeep Narayanan K V, Vishnu Soman proposed "A Novel Method for Text-Independent Speaker Identification Using MFCC and GMM". They focussed on text independent speaker recognition system using GMM. The experiments conducted for various speech time durations and achieved high recognition rate.

"Improved Recognition Rate of Language Identification System in Noisy Environment" was a paper presented by Randheer Bagi, JainathYadav, K. SreenivasaRao.In this paper, they focussed on language identification in noisy environment and for that they used GMMs to train the models. Also, they used Spectral Subtraction and Minimum Mean Square Error (MMSE) methods to reduce the noise from the speech signal. They compared the recognition rate of the system for the clean signal and noise suppressed signal.

"Speaker identification and verification using Gaussian mixture speaker models" was a paper proposed by Douglas A. Reynolds. This paper was focussed to achieve the high performance speaker identification and verification, these were evaluated using different speech databases such as TIMIT, NTIMIT and YOHO. It uses different levels of degradations, speech quality with noisy signal, clean speech, telephonic speech.

3. Gaussian Mixture Model(GMM)

There are two types of methods, Deterministic methods and statistical methods. Here we used the statistical method that is parametric model called Gaussian Mixture Model (GMM) for the identification of speakers.

The weighted sum of M component densities gives the Gaussian mixture density.

The Gaussian mixture density have the parameters mean vectors, covariance matrices and mixture weights, such that these can be represented by

$\lambda = \{ P_i, \bar{\mu}, \sum i \}$ i=1,……..M.

There are two main reasons for using the GMM as representation of speaker identity.

Individual densities of a multi modal density like GMM may model underlying set of acoustic classes also such that these set of acoustic classes are used to characterize the speech utterance of the particular speaker that represents some phonetics such as vowels, nasals and fricatives.

Another important reason involves, GMM is able to represent large class of samples or its distributions by means of linear representation of Gaussian functions.

GMM is a parametric model, the maximum likelihood estimation is one of the method available for estimating the parameters of GMM. The algorithm called Expectation Maximization (EM) can be used for estimation of ML parameter iteratively and these parameters maximize the likelihood of the GMM.

For to guarantee the improvement in model's likelihood value, the following formulas can be used,

$$\overline{p_1} = \frac{1}{T}\sum_{t=1}^{T} p(i/\bar{x}_t, \lambda)$$

$$\bar{\mu} = \frac{\sum_{t=1}^{T} p(i/\bar{x}_t,\lambda).\bar{x}_t}{\sum_{t=1}^{T} p(i/\bar{x}_t,\lambda)}$$

$$\bar{\sigma}^2 = \frac{\sum_{t=1}^{T} p(i/\bar{x}_t,\lambda).\bar{x}_t^{\,2}}{\sum_{t=1}^{T} p(i/\bar{x}_t,\lambda)} - \bar{\mu}^2$$

The M component Gaussian mixture density forming a GMM can be represented by $p(\vec{x}/\lambda)$.We have T independent training vectors for a given sequence, for these vectors calculate the log likelihood scores and search for the maximum likelihood. The log likelihood can be computed as

$$\log p(X/\lambda) = \frac{1}{T}\prod_{t=1}^{T} p(\vec{x}/\lambda)$$

The system uses a log-likelihood scores to know whether the claimed speaker is true or false. Thus for an input vector and a claimed speaker model λ_1, the likelihood score can be given by $p(x/\lambda 1)$.

So, based upon the log likelihood scores of the unknown speaker model and the trained speaker models, the decision of accept or reject of the speaker can be made by considering a given threshold.

Selecting the order M of the mixture and initializing the parameters before EM algorithm plays a dominant role in the speaker modelling.

By using likelihood algorithm, likelihood scores for all speakers are calculated for the corresponding GMM models which forms the trained database. Finally, the likelihood score of unknown speaker compared with the trained database.

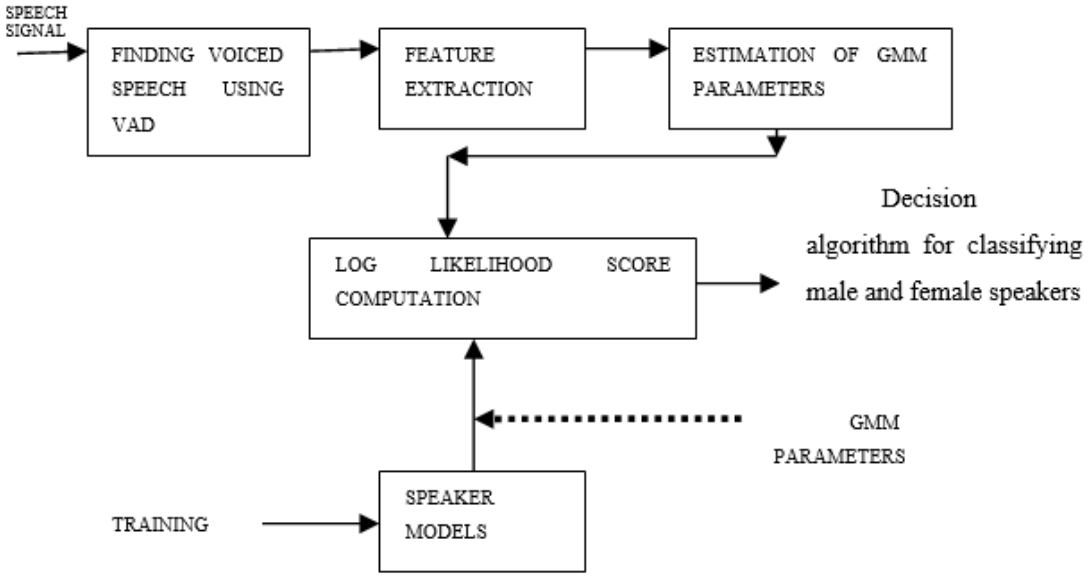

Figure 1. Block diagram of the proposed system

4. Voice Activity Detector (VAD)

By extracting the information from the speech signal, the voiced and unvoiced decisions of speech signal can be performed. Here the methods used to separate voiced and unvoiced parts is zero crossing rate (ZCR) and. The zero-crossing rate is low and energy is high for voiced part while ZCR is high and energy is low for unvoiced part. Voiced part, it is periodic in nature and can be identified and extracted, while unvoiced part, it is non-periodic in nature. Zero cross count indicates the frequency at which the energy is concentrated. Energy of a speech is another parameter for classifying the voiced/unvoiced parts. The voiced part of the speech has high energy because of its periodicity and the unvoiced part of speech has low energy. We already the

amplitude of the speech signal continuously varying with respect to time. unvoiced speech segments have less strength compared to strength of the voices segments. The speech signal's energy gives these variations in the amplitude of the speech utterance. Short-time energy can define as:

$$En = \sum_{m=-\infty}^{\infty}[x(m).w(n-m)]^2$$

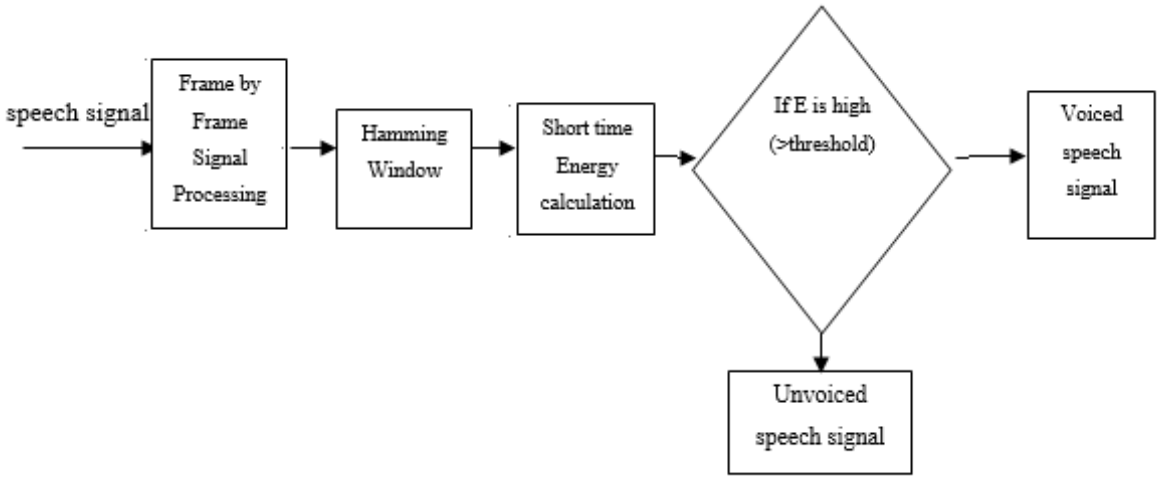

Figure 2. The flow chart of VAD for detecting the voiced and unvoiced part of speech signal

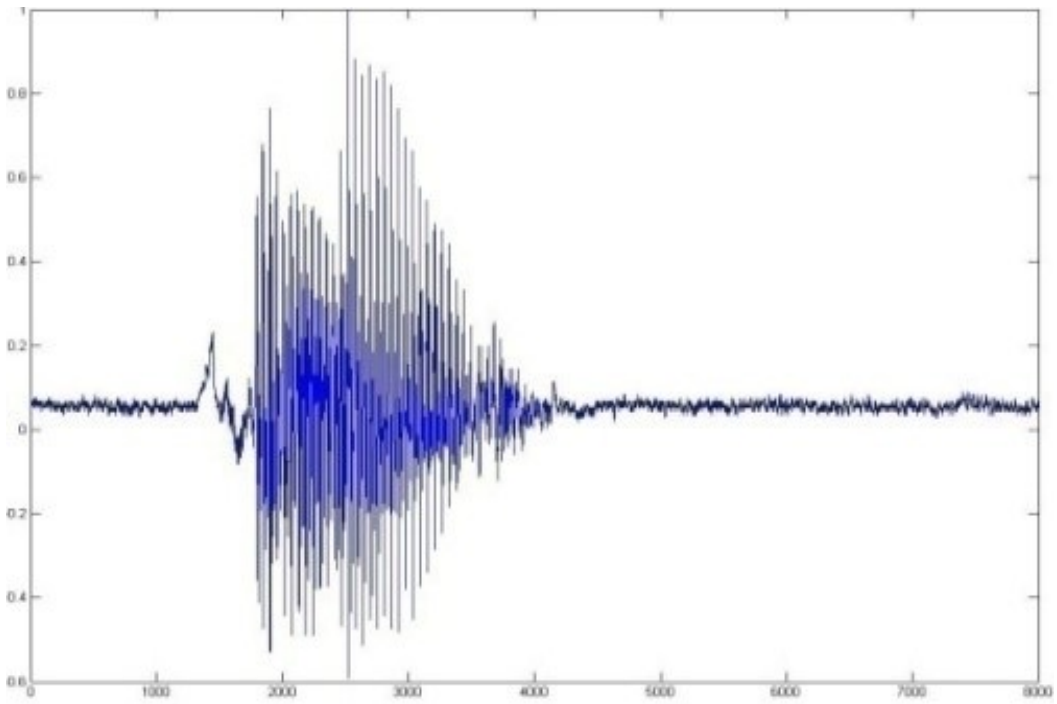

Figure 3. Original speech signal for the word "hello"

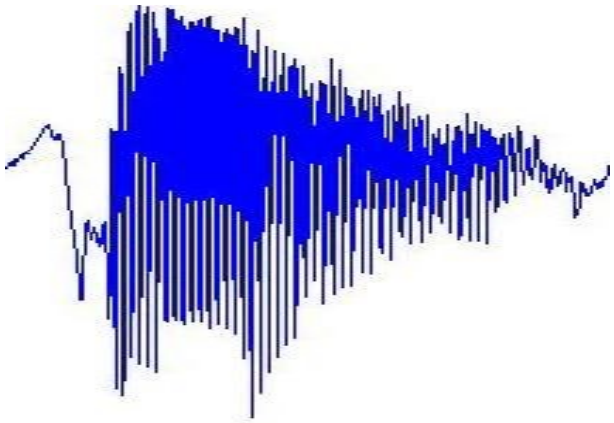

Figure 4. Speech signal after applying VAD

For the below speech utterance of the word “hello” for 5sec duration, the voice activity detector is used to find only the voiced part of the speech utterance and remaining unvoiced is eliminated, such that features obtained in the feature extraction phase will be only for voiced part which improves the speaker modelling and more better speaker classification process, thus improves the identifying the speaker and gender.

Thus, the modified speech utterance such that to get only the voiced part of the speech signal can be obtained by using Voice activity detector (VAD). The sample utterance is shown in the below figure.

5. Experiments

For this Text independent system using Gaussian mixture models, the experiments conducted using Matlab 7 language environment. Here about 50 number of speakers were trained and uses the locally recorded database. Out of these 50 speakers, 40 were female and 10 were male. Also, most importantly, the experiments are conducted such that training and testing in the same environment and compute recognition rate by calculating the log-likelihood scores

%Recognition rate = [(scores obtained in the testing) / (scores obtained in the training)]*100

Here, Mel Frequency Cepstral Coefficients (MFCC) are used to extract features from the speech utterance. These features are widely used in speaker recognition to obtain speaker specific information which gives frequency distribution of sounds along with vocal tract shape and length. For each frame and Every 10 ms,12 MFCC together with log energy were calculated.

In this, the adaptive filter can also be used to reduce the noise effects in the speech signal. The effect of using the adaptive filter on the input speech utterance can be observed below. The

plots of the speech signal with reduced noise and the speech signal with noise can be obtained using Matlab.

6. Results

In this, the speakers are collected through microphone, thus called it as locally recorded database and the speech signal was sampled at 8KHz sampling frequency. In these experiments, the adaptive filters namely LMS, NLMS and RLS filters may be used to slightly reduce the effect of noise in the speech signal. In identifying the gender of the speaker, we compute the log-likelihood scores for each speaker, which are appeared in the modelling of the speaker through Expectation Maximization (EM) algorithm. Also, here the accuracy or recognition rate can be used to measure the performance of the system. The log-likelihood scores for all the speakers are computed and this can be useful in classifying the gender of the speaker, that is, the variations in scores for male and female can be used to classify the gender and thus finally identifies the gender and speaker simultaneously.

The recognition rate of 96.96% was achieved for the system with adaptive filter, also this can be computed based on log likelihood scores. Moreover, the unknown speaker can be accepted as a true speaker when the log likelihood ratio is below the predetermined threshold, otherwise the speaker was rejected and does imply as a imposter.

The threshold for log-likelihood score is selected as 0.3, inorder to identify the gender of the speaker. The log-likelihood score in the testing and training are used as a measure of recognition rate. This is considered for system using GMM but MFCC coefficients are obtained for speech signal without Voice Activity Detector (VAD) in the feature extraction phase.

Table 1. The loglikelihood scores for a system without VAD involves different speakers to identify gender

Testing	**Given unknown speaker for testing**	**Log likelihood ratio**	**Gender identified as**	**Recognition rate**
1st time	S1	0.40	Male	76.73%
2nd time	S2	0.50	Female	Incorrectly identified
3rd time	S3	0.70	Female	Incorrectly identified

The below table shows the log likelihood ratio and recognition rate for the speaker recognition system that uses VAD in the feature extraction phase, that is MFCC obtained for only the voiced part of the speech signal.

Table 2. The log-likelihood scores for a system with VAD involves different speakers to identify gender

Testing	Given unknown speaker for testing	Log likelihood ratio	Gender identified as	Recognition rate
1st time	S1	0.77	Male	96.73%
2nd time	S2	0.23	Female	95.56%
3rd time	S3	0.15	Female	96.96%

The below figure 5 shows the maximum likelihood score for the particular speaker. Similarly, we can plot the maximum likelihood scores for remaining speakers also such that the speaker can provide any text.

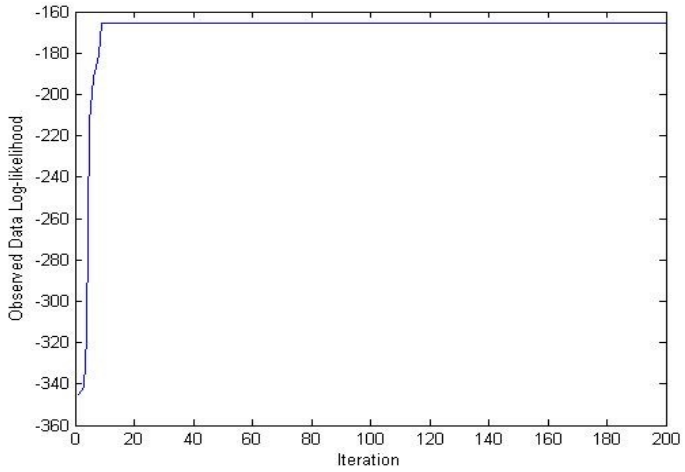

Figure 5. Plot of iteration vs log likelihood for the unknown speaker

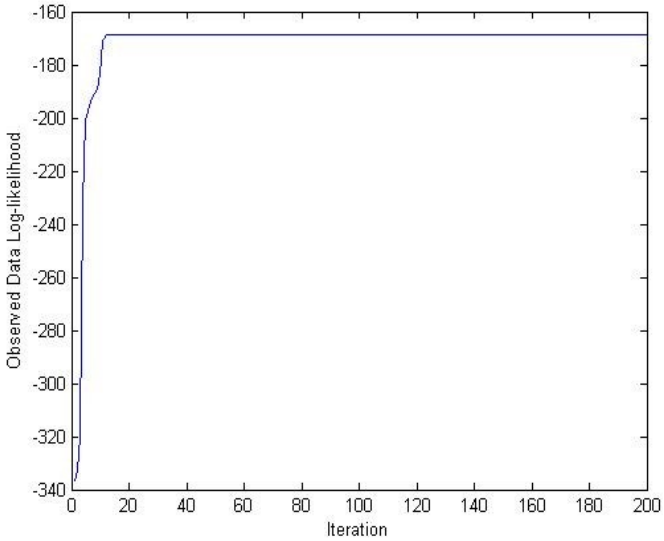

Figure 6. Plot of Iteration vs log likelihood for one of the speaker in the trained database.

The same test was conducted under different environmental conditions in both testing and training. The recognition rate was decreases and also log likelihood ratio was also not within the threshold was observed, so there was significant degradation of the performance of the system in many mismatched conditions for testing and training. But, if the speaker recognition system uses the same database for training and testing rather than mismatching then the recognition rate was improved.

Table 3. The identification of speaker under matched conditions

Speaker	Log likelihood ratio	Recognition rate
1st time	5.59	96.73%
2nd time	7.68	95.56%
3rd time	5.18	96.96%

The below table shows the performance of the system in real time without any adaptive filter in terms of recognition rate.

Table 4. Output for the system under mismatched conditions

Speaker	Log likelihood ratio	Recognition rate
1st time	>8	57 % with the result “Speaker does not found”
2nd time	>8	65% with the result “Speaker does not found”
3rd time	>8	55 % with the result “Speaker does not found”

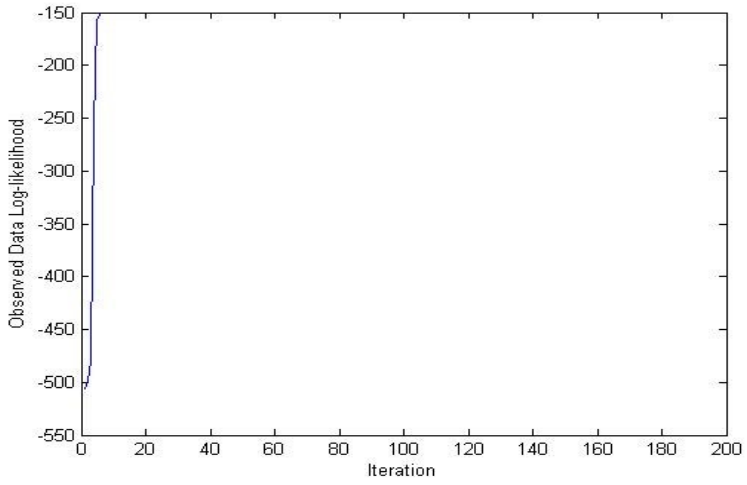

Figure 7. Plot of iteration vs log likelihood for an imposter

Here, the recognition rate for the system with enhanced speech signal and system with noisy speech signal are compared as shown in table 1 and table 2. Also, the speech utterance of

duration of 5sec is used for each speaker (3 times) and computed the log likelihood scores. So the task of recognizing the person in real time was somehow improved, that is shown in above tables.

The log likelihood score for each iteration was computed. So for all 200 iterations, computed log likelihood scores and GMM parameters called mean, variance also computed. The plot of log likelihood scores for the 200 iterations is shown in the below figure.

7. Conclusion

This paper shows the text independent system using GMM and MFCC along with Voice activity detector was able to identify the gender of the speaker such that this information about the speaker may useful in many security applications. The system can also use adaptive filter as pre-processing stage and also the recognition rate of 97.36% was achieved using this proposed method, while identifying the unknown speaker. However, the accuracy drastically reduces under mismatched conditions. Further, the performance of the system can be improved by optimization techniques for feature selection after the feature extraction phase.

References

1. Ergün Yücesoy, Vasif V. Nabiyev, Gender Identification of a Speaker Using MFCC and GMM, 2013, ELECO 8th international conference.
2. Lantian Li, Thomas Fang Zheng, Genderdependent Feature Extraction for Speaker Recognition, China SIP 2015.
3. Huang Ting, Yang Yingchun, Wu Zhaohui, Combining MFCC and Pitch to Enhance the Performance of the Gender Recognition.
4. M.S. Sinith, AnoopSalim, K GowriSankar., K.V. Sandeep Narayanan, Vishnu Soman, A Novel Method for Text-Independent Speaker Identification Using MFCC and GMM ICALIP 2010.
5. P. Hema Kumar, V. Srinivas, T. Madhu. Improved Dynamic Speaker Recognition System using NLMS Adaptive Filter, 2016, *International Journal of Computer Applications*, vol. 148, no. 10.
6. R. Bachu, B.K. Adapa, S. Kopparthi. Barkana Buket, Separation of Voiced and Unvoiced Speech Signals using Energy and Zero Crossing Rate, 2008.
7. Wenyong Lin, An improved GMM based clustering algorithm for efficient speaker identification, 2015, 4th International Conference on Computer Science and Network Technology (ICCSNT 2015).

8. Michael Lutter, Mel Frequency Cepstral Coefficients (feature extraction/MFCC). The Speech Recognition Wiki25 November 2014.
9. J.P. Campbell, Speaker Recognition: A Tutorial, 1997, Proc. Of the IEEE, vol. 85, no. 9, pp. 1437-1462.
10. Vibha Tiwari, MFCC and its applications in speaker recognition, 2010, International Journal on Emerging Technologies, vol. 1, no. 1, pp. 19-22.
11. Rania Chakroun, Leila Beltaïfa Zouari, Mondher Frikha, Ahmed Ben Hamida, Improving Text-independent speaker recognition with GMM, 2nd International Conference on Advanced Technologies for Signal and Image Processing - ATSIP'2016 March 21-24, 2016, Monastir, Tunisia.
12. Vaishali M. Karne, Akhilesh Singh Thakur, Vibha Tiwari, Least Mean Square (LMS) Adaptive Filter For Noise Cancellation, International Journal of Application or Innovation in Engineering & Management (IJAIEM), ISSN 2319–4847.
13. Sourjya Sarkar, K. Sreenivasa Rao, Speaker Verification in Noisy Environment Using GMM Supervectors, 2013.
14. Sheng Zhang, Jiashu Zhang, Hongyu Han, Robust Variable Step-Size Decorrelation Normalized Least-Mean Square Algorithm and its Application to Acoustic Echo Cancellation IEEE/ACM Transactions on Audio, Speech, and Language Processing.
15. Xin-xing ling, Ling Zhan, Hong Zhao, Ping Zhou Speaker Recognition System Using the Improved GMM-based Clustering Algorithm.
16. Jayant M. NaikSpeaker Verification: **A** Tutorial January 1990 - IEEE Communications Magazine.
17. Yuan Liu, Tianfan Fu, Yuchen Fan, YanminQian, Kai Yu Speaker Verification with Deep Features2014 International Joint Conference on Neural Networks (IJCNN), July 6-11, 2014, Beijing, China.
18. SourjyaSarkar, K. Sreenivasa Rao, Significance of Utterance Partitioning in GMM-SVM Based Speaker Verification in Varying Background Environment, A Novel Windowing Technique for Efficient Computation of MFCC for Speaker Recognition Md Sahidullah, Student Member, IEEE, Goutam Saha, Member, IEEE.
19. RandheerBagi, JainathYadav, K. Sreenivasa Rao, Improved Recognition Rate of Language Identification System in Noisy Environment.
20. Douglas A. Reynolds Speaker Identification and verification using Gaussian Mixture speaker models speech communication, 1995, vol. 17, ELSEVIER.