UTM UNIVERSITI TEKNOLOGI MALAYSIA
INTERNATIONAL JOURNAL OF INNOVATIVE COMPUTING
ISSN 2180-4370
Journal Homepage : https://ijic.utm.my/

Study on Gender Identification Based on Audio Recordings Using Gaussian Mixture Model and Mel Frequency Cepstrum Coefficient Technique

Thurgeaswary Rokanatnam & Hazinah Kutty Mammi
School of Computing, Faculty of Engineering
Universiti Teknologi Malaysia
Email: thurgeas1197@yahoo.com; hazinah@utm.my

Submitted: 1/9/2021. Revised edition: 4/10/2021. Accepted: 26/10/2021. Published online: 15/11/2021
DOI: https://doi.org/10.11113/ijic.v11n2.343

***Abstract*—Speaker recognition is an ability to identify speaker's characteristics based from spoken language. The purpose of this study is to identify gender of speakers based on audio recordings. The objective of this study is to evaluate the accuracy rate of this technique to differentiate the gender and also to determine the performance rate to classify even when using self-acquired recordings. Audio forensics uses voice recordings as part of evidence to solve cases. This study is mainly conducted to provide an easier technique to identify the unknown speaker characteristics in forensic field. This experiment is fulfilled by training the pattern classifier using gender dependent data. In order to train the model, a speech database is obtained from an online speech corpus comprising of both male and female speakers. During the testing phase, apart from the data from speech corpus, audio recordings of UTM students will too be used to determine the accuracy rate of this speaker identification experiment. As for the technique to run this experiment, Mel Frequency Cepstrum Coefficient (MFCC) algorithm is used to extract the features from speech data while Gaussian Mixture Model (GMM) is used to model the gender identifier. Noise removal was not used for any speech data in this experiment. Python software is used to extract using MFCC coefficients and model the behavior using GMM technique. Experiment results show that GMM-MFCC technique can identify gender regardless of language but with varying accuracy rate.**

***Keywords*—Speaker recognition, feature extraction, Mel Frequency Cepstrum Co-efficient (MFCC), Gausian Mixture Model (GMM), gender identification**

I. INTRODUCTION

History shows that speech recognition has seen breakthroughs since the 18th century providing tech experts a platform to develop the current technology known today. A genuine speech recognition was invented in the 1950s with a 90% accuracy rate by Bell Labs. In the mid 1980's IBM released a speech recognition system, IBM Tangora, using hidden Markov model. The speech recognition system was successfully released as a software in 1997. This led to Google Voice Search App by iPhone which was created in 2008 [1]. For years, people have been interested in identifying a person's characteristics through voice recordings. Voice recordings has broad range of use in many fields today, such as, health care industry, service delivery, automated identification, and communication service providers.

Over the years, speaker recognition was developed from speech recognition system to disregard the language and focus on identifying the physical person behind the voice. Now, speaker recognition system has advanced into being able to determine speaker's gender and accent as well.

The key issue this research would like to address is the use of speaker recognition in forensics. The law enforcement team has been trying to find a solution to solve cases using forensic analysis methods based on voice recordings. Forensic voice is used for a vast range of criminal cases such as, rape, human trafficking, murder, and money laundering. Suspects might leave voice recordings in the form of phone conversations, voice mail, ransom demands, hoax calls and calls to emergency or police numbers [2]. Identifying voice using forensic methods is a difficult task, for example, in the past, voice experts evaluated voice recordings using their experience rather than statistical analysis. The results were not always accurate and was not reliable evidence in investigations.

There a quite a number of researches done in gender recognition of speakers based on audio recording around the world; while relevant research done in Malaysia is quite limited.

Findings on gender recognition using Malaysian speakers' voices may be able to strengthen the studies done by other researchers. It will also prove the performance rate and efficiency of the technique used to differentiate the genders based on audio recordings only. In this research, respondents are local Universiti Teknologi Malaysia (UTM) students from different ethnicity groups.

II. RELATED WORKS

Some, reference papers are studied to understand the techniques used in gender identification. Review of some papers are given below:

A. Description of Related Studies

A gender identification system was proposed in a paper where the gender identification of a speaker uses the classification of MFCC with GMM [3]. The experiment was run using TIMIT database for 760 sentences with 76 speakers and 6100 sentences with 610 speakers. 97.67% success rate was gained from determining 5958 sentences correctly out of 6100 sentences. The paper concluded that increase in the Gaussian components and MFCC coefficients increase the success rate of the system.

Research proposed that a gender classification system based on GMM models using clean, noisy speech and multi languages [4]. This study was conducted to study the performance of the feature extractors, pitch and RASTA-PLP, separately and then combined into as one feature extractor. The speech database was obtained from TIMIT comprising of English, German, Japanese and English languages. The PLP was used along with RASTA as feature extraction and selection method. This technique was proved to be more robust in noisy speaker recognition. The classification accuracy was above 98% for all clean speech and remained 95% for noisy speech. It was also proven that the classification performance has shown better performance when the parameters, pitch and RASTA-PLP, are combined rather than using only either one.

Another research proposed text independent speaker identification using MFCC and Deep Neural Networks [5]. The cases were tested using two kinds of hybrid, MFCC-DNN and CFCC-DNN for both genders. CFCC is also another type of feature extraction which are defined based on a recently developed auditory transform and a set of modules to perform the signal processing functions in cochlea. The accuracy of MFCC – DNN hybrid is higher compared to CFFC-DNN hybrid. At 10 utterances and 10 dB noise, MFCC-DNN hybrid achieved 90.24% in both men and women while CFCC-DNN hybrid was 87.84% in men and women. At 10 utterances and 6 dB noise, MFCC-DNN hybrid performance was 83.4% while at 12 dB noise was 73.1%. It was higher compared to CFFC-DNN hybrid. Also, at 20 utterances am 0 dB and 6 dB noise, MFCC-DNN hybrid showed higher speaker identification accuracy than CFCC-DNN hybrid. At 20 utterances and 12 dB noise, MFCC-DNN hybrid was 93.12% in both men and women. However, CFCC-DNN hybrid was only 87.74%. Hence, it was proved that MFCC has a higher performance rate than CFCC.

In this research, a comparison study in text independent speaker identification system was proposed using different feature classifiers [6]. For feature classifiers, four algorithms were applied; ANN, GMM, VQ and DT. Later, fusion method was used to assess the efficiency of the system. The recorded utterances were in English language but with different dialects under different conditions. The study was conducted with different number of speakers' organization database, ranging from 10 to 120 speakers. It was observed that better identification rate was achieved in small group of testing and training speakers. Generally, GMM produced the best identification rate compared with other algorithms.

TABLE I. COMPARATIVE RESULT ANALYSIS OF SPEECH SIGNALS

Author Name	***Feature Extractor***	***Feature Classifier***	***Performance Rate***
Yucesoy *et al.* (3)	MFCC	GMM	97.67%
Zeng *et al.* (4)	Pitch	GMM	96.5%
	RASTA-PLP		93.7%
	Pitch & RASTA-PLP		98.1%
Sarhan *et al.* (5)	MFCC	DNN	83.4%
	CFCC		73.1%
Alhalabi *et al.* (6)	MFCC	VQ	92.0%
		GMM	94.2%
		ANN	85.8%
		DT	84.1%

B. Proposed Solution

A project from GitHub will be used in this experiment. The project, Voice based gender recognition using Gaussian Mixture Model, is developed by Ayoub Malek [7]. In this experiment, the source code and blog will be the reference for MFCC and GMM technique to determine the gender of speakers based on audio recordings.

This GitHub project's main idea is to recognize the gender of the speaker based on pre-generated Gaussian mixture models (GMM). Once the data is properly formatted, the Gaussian mixture models are trained for each gender by gathering Mel-frequency Cepstrum Coefficients (MFCC) from their associated training wave files. After the models have been generated, the speakers' genders are identified by extracting their MFCCs from the testing wave files and scoring them against the models. These scores represent the likelihood that user MFCCs belong to one of the two models. The gender models with the highest score represents the probable gender of the speaker.

III. FEATURE EXCTRACTION USING MFCC

MFCC being the feature extractor is used to convert speech signals from time domain to frequency domain as shown in Fig. 1.

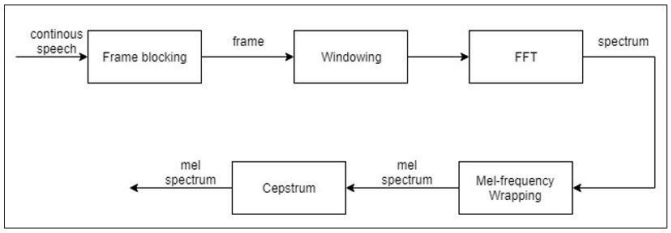

Fig. 1. MFCC Process [8]

A. Frame Blocking

Since speech is a non-stationary signal, its frequency contents are continuously changing with time. In order to do any sort of analysis on the signal it has to become stationary signal as shown in Fig. 2. The speech signal will be segmented into small duration of blocks, 20 – 30ms which are known as frames. The signals will be divided into N samples, with adjacent frames being separated by M(M<N) samples. The frame duration is calculated by dividing the frame size with sample rate. This process goes on until all speech data is converted into frames.

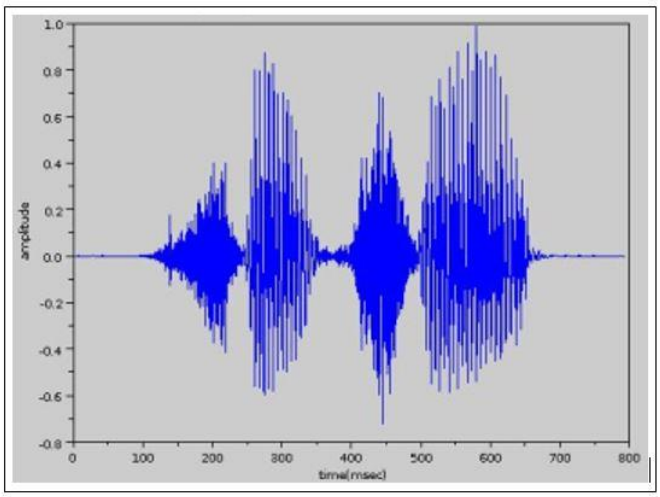

Fig. 2. Non-stationary signal [9]

B. Windowing

Discontinuities can happen towards the endpoints because of raw frames extractions from speech signals. This happens because of non-integer number of periods in the extracted waveform, which will then lead to an erroneous frequency representation. Windowing is done to each and every frame in order to keep the continuity of the beginning and the end of frames like in Fig. 3. The window function is added by tapering the voice sample signal to zero in the starting and ending point of every frame.

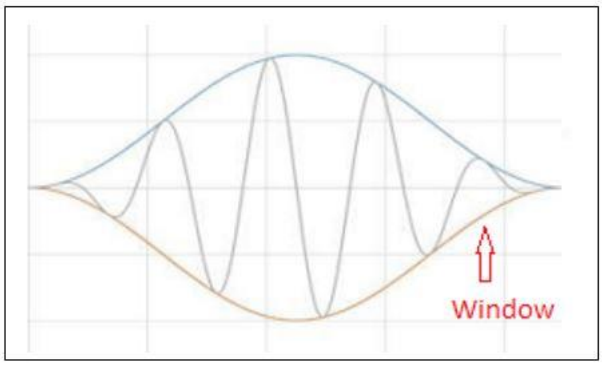

Fig. 3. A windowed signal [9]

C. FFT

FFT algorithm is known as Fast Fourier Transform. This step is to convert all the frame samples from time domain into frequency domain. By performing FFT, the output will be in spectrum.

D. Mel-Frequency Wrapping

There is another step in Mel-Frequency wrapping, triangular band pass filters. The spectrum will be multiplied with a set of 20 triangular band pass filters to get smooth magnitude spectrum. These filters will be placed equally along Mel frequency scale.

E. FFT

Finally, the Mel spectrum will be converted back to time domain. The conversion will be using the calculation of Discrete Courier Transform, DCT. The obtained features are similar to Cepstrum, so will be known as Mel-scale Cepstrum coefficients [9].

IV. Gaussian Mixture Modelling

Gaussian Mixture Model (GMM) is created after extracting the features. The probability density function is used to calculated Gaussian Probability. The multivariate normal distribution is essential in Gaussian distribution statistics. It is used to study the dependence of random variables.

The likelihood of data points which are known as, feature vectors, for a model is given by following equations (Eq. 1 and Eq. 2):

$$P(X|\lambda) = \sum_{k=1}^{K} \omega_k P_k(X|\mu_k, \sum_k) \quad (1)$$

where,

$P_k(X|\mu_k, \sum_k)$ is the Gaussian distribution

$$P_k(X|\mu_k, \sum_k) = \frac{1}{\sqrt{2\pi|\sum_k|}} e^{\frac{1}{2}(X-\mu_k)^T \sum^{-1}(X-\mu_k)} \quad (2)$$

The training data X_i of the class λ are used to estimate the parameters, co-variance matrices, mean, and weights of these k components.

- λ is the training data.
- μ represents the mean.
- Σ is for co-variance matrices.
- ω_k are the weights.
- k refers to the index of the GMM components.

Fig. 4 shows each component density denotes a D-dimensional distribution with co variance and mean vector. These are the parameters used in complete Gaussian mixture density. Each speaker is represented by a GMM and is referred by the respective model λ. The most popular estimation to calculate the optimum model of each speaker is maximum likelihood (ML) algorithm. By maximizing the likelihood of GMM, the model parameters can be found. The ML parameter estimation is obtained by using an iterative algorithm, expectation – maximization (EM) algorithm.

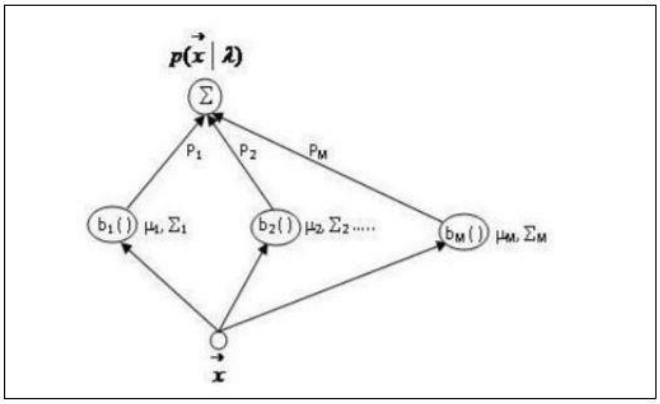

Fig. 4. Gaussian Mixture Model [10]

A. EM Algorithm

The EM algorithm consists of two steps, an E-step, also known as, Expectation step and M-step, or known as, Maximization step. The main goal is to maximize the marginal likelihood of X based on the given parameters (denoted by vector θ). The marginal distribution can be found as the joint of X and Z and sum over all of Z's sum rule of probability).

$$l_n\, p(X|\theta) = l_n\{\sum_z p(X, Z|\theta)\}$$

Eq. 3 Marginal likelihood with latent variables [11]

The above equation usually results in a complicated function that is not easy to maximize. Essentially, to estimate the model, two steps are needed to be carried out. For the first step (E-step), the posterior distribution of the latent variables $\gamma\gamma$ estimated conditionally on the weights, (π), means (μ), and covariance (Σ). The vectors of parameters are denoted as θ. Taking the results from the above equation, Gaussian Distribution, the responsibilities that each Gaussian's posterior distribution has for each data point can be calculated using the formula below in Eq. 4. The equation is Bayes rule where π is the prior weights and the likelihood is normal.

$$\gamma(z_{nk}) = \frac{\pi_k N\,(x_n|\mu_k,\ \Sigma_k)}{\sum_{j=1}^{K} \pi_j\, N\,(x_n|\mu_j, \Sigma_j)}$$

Eq. 4 Posterior Responsibilities using Bayes Rule [11]

After calculating the posterior, an estimate of the parameters of each Gaussian is needed defined by the equations below and then evaluate the log-likelihood.

$$\mu_k^{new} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma\,(z_{nk}) x_n)$$

Eq. 5 Equation for mean of the Gaussians [11]

$$\sum\nolimits_k^{new} = \frac{1}{N_k} \sum_{n-1}^{N} \gamma\,(z_{nk})(x_n - \mu_k^{new})\ (x_n - \mu_k^{new})^T$$

Eq. 6 Equation for covariance of the Gaussians [11]

$$\pi_k^{new} = \frac{N_k}{N}$$

Eq. 7 Equation for weights [11]

$$N_k = \sum_{n=1}^{N} \gamma\,(z_{nk})$$

Eq. 8 Equation for sum of responsibilities in each Gaussian k [11]

$$\ln p(X|\mu, \Sigma, \pi) = \sum_{n=1}^{N} \ln\{\sum_{k=1}^{K} \pi_k N\,(x_n|\mu_k, \Sigma_k\}$$

Eq. 9 Equation for marginal likelihood (11)

V. RESEARCH DESIGN AND IMPLEMENTATION

For this research, there are two phases in an experiment, used to identify gender of speakers based on speech corpus data and audio recordings of UTM students. The whole experiments design and development is described in Fig. 5.

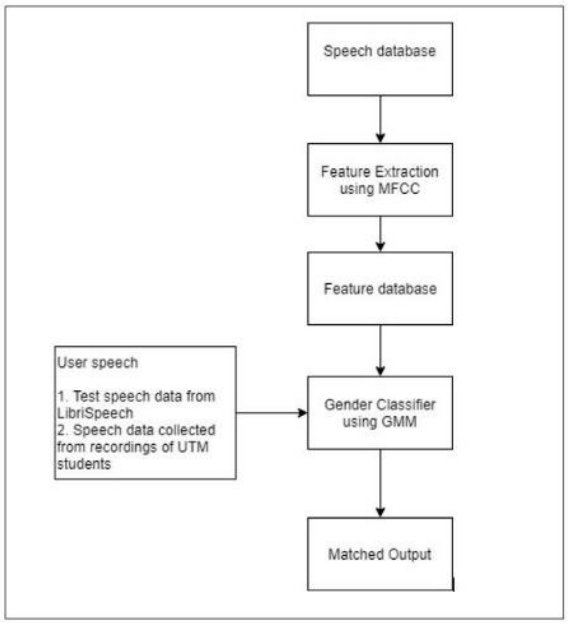

Fig. 5. Experiment Design

A. Training Phase

In this study, training phase is the most crucial part of the experiment. The gender classifier that determines the gender is based on the GMM models trained in this phase of experiment. The features have to be extracted and saved in order to differentiate between both genders. Also, the model trainer has to build models based on these extracted features to identify the gender of speakers from audios that are about to be tested randomly.

B. Testing Phase

During the second phase, training phase, the accuracy of gender identification of speakers will be determined. Also, performance evaluation will be done by comparing the accuracy with by running different data set in altered source codes.

C. Parameter and Testing Method

During the testing phase, GMM models will be made for different speakers and will be loaded and input will be made in waveform from any speakers from the speech corpus. The results from this experiment will be derived according to the number of components of MFCC and GMM are used and the success rate of each training data set speech. The success rate is calculated by using the log-likelihood for each gender after the MFCC trained speech model is passed on to GMM [11]. The scores from all speakers will be computed and will be put into a database. During the training phase, the speakers with maximum scores will be identified as the ones who have spoken. The test phase is explained in Fig. 6 below.

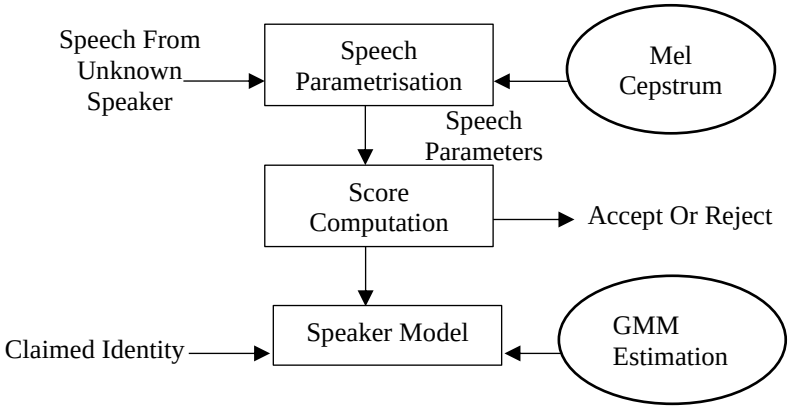

Fig. 6. Test Phase [12]

VI. RESULTS, ANALYSIS AND DISCUSSION

This experiment was divided into two phases; training and testing phase. This testing phase is carried out in two separate parts to determine whether the accuracy rate of gender identification is the same.

The training data consisted of 804 male audio files and 1163 female audio files. The code responsible for the entire gender classification took about 10 minutes 9 seconds to complete the GMM modelling.

A. Testing phase using SLR45 Data Set

This experiment is only carried out in order to ensure the source code of gender is executed correctly. Also, this phase will be set as benchmark for the accuracy of gender identification of the obtained source code.

This experiment was done using a total of 988 audio files consisting of both genders. The analysis is done separately for each gender. Table II shows the results of tested female audio files. Out of 584 audio files. 566 audio files were identified correctly. Only 18 audio files were labelled incorrectly as male. Using the results, the following accuracy rate can be computed to 96.9%.

TABLE II. ACCURACY RATE FOR FEMALE AUDIO TESTING

Testing Result	**Accuracy**
Correctly Identified	566
Incorrectly Identified	18
Female Audio Total	584

Table III shows 375 audio files were recognized correctly out 404 audio files. 29 files were identified incorrectly as female. The gender accuracy percentage for male was 92.8%.

TABLE III. ACCURACY RATE FOR MALE AUDIO TESTING

Testing Result	**Accuracy**
Correctly Identified	375
Incorrectly Identified	29
Male Audio Total	404

The average gender accuracy percentage was 95.2%. The research article stated 95% for gender accuracy percentage for this experiment.

B. Testing phase using UTM Students' Data Set

This is the second part of the testing phase in this experiment. This data set consists of 60 audio files consisting of both genders. 15 speakers from each gender were involved in this data collection. For each gender, there were 5 Indians, 5 Malays and 5 Chinese speakers. Each speaker has two audio recordings; one in English and one in Bahasa Malaysia. The sentences used in these recordings were fixed.

This experiment was done using the same source code with the same feature vectors of GMM models. The analysis is done separately for each gender. However, the analysis will be done for gender and as well as languages, Bahasa Malaysia and English. Table IV shows the results of tested female audio files. The gender accuracy percentage for females was 100% correct for both languages. Table V shows accuracy rate for both languages for male speakers.

TABLE IV. ACCURACY RATE FOR FEMALE AUDIO TESTING

	Bahasa Malaysia	**English**
Correctly Identified	15	15
Incorrectly Identified	0	0
Female Audio Total	15	15

TABLE V. ACCURACY RATE FOR MALE AUDIO TESTING

	Bahasa Malaysia	**English**
Correctly Identified	6	11
Incorrectly Identified	9	4
Male Audio Total	15	15

- Accuracy for male = (6+11) / 30 = 0.566 (56.6%)
- Accuracy for female = (15+15) / 30 = 1.0 (100%)
- Total accuracy = 47/60 = 0.783 (78.3%)

It is observed that unbalanced training data set leads to gender biased speaker recognition. The dimensionality of two different data sets might affect the accuracy rate given that the GMM models are only trained once. Lack of diversity in training data influences gender identification accuracy because of the language spoken during audio recordings. The GMM models created were only based on one language, English. This also explains why the audio files in English language scored higher on average for the accuracy rate compared to Bahasa Malaysia. It was also observed that 100% accuracy rate was achieved for female audio files and the accuracy rate for male audio files are much lower. This is because of gender recognition system in this experiment during the training phase was trained with more female data compared to male data.

Also, background noise affects the accuracy rate of gender identification. All the recordings in UTM students' data set were recorded by the speaker themselves in their own environment. Due to the enforcement of MCO across the country, it was impossible to meet the speakers directly to record the audios. Thus, the collected audio recordings had different environment and microphone conditions. Noises and echoes are unavoidable interference while recording audio.

Other than that, it can be concluded from the results that the experiment conducted using UTM students' data did not have any significant errors in identifying genders regardless of the different ethnicity and accent of the speakers. The identification of gender of speakers based on audio recordings can be done using GMM-MFCC technique but provided the quality of audio used in the dataset should be up to par. Also, the training set used to train the GMM models must be balanced for both genders. Data set obtained from a Malaysian demographic dataset will yield better results.

VII. CONCLUSION

This paper has evaluated the use of GMM models, made using the features extracted using MFCC on two different data sets with two different languages. It is proven that the gender of speakers can be identified through audio recordings based on GMM-MFCC technique. Also, it was found that gender of a speaker can be identified regardless of the language spoken. The improvements that could be done to this study would be to improve the robustness of MFCC to extract features even in noisy speech signals. Future works should be continuously done to determine speakers by race or ethnicity; and different cultural and educational upbringing. In order to achieve this, more Malaysian speakers' data set should be made available online. By having a large data set of Malaysian speakers, speaker models can be trained to accommodate any speaker recognition system made for Malaysia.

REFERENCES

[1] Boyd, C. (2018, January 10). Speech Recognition Technology: The Past, Present, and Future. Medium. https://medium.com/swlh/the-past-present-and-future-of-speech-recognition-technology-cf13c179aaf.

[2] Michele Catanzaro, Elisabetta Tola, Philipp Hummel, Astrid Viciano. (2017, January 25). Voice Analysis should be Used with Caution in Court. Scientific American. https://www.scientificamerican.com/article/voice-analysis-should-be-used-with-caution-in-court/.

[3] Yucesoy, E. and Nabiyev, V. V. (2013). Gender Identification of a Speaker using MFCC and GMM. *ELECO 2013 - 8th International Conference on Electrical and Electronics Engineering*. IEEE Computer Society, 626-629.

[4] Zeng, Y. M., Wu, Z. Y., Falk, T. and Chan, W. Y. (2006). Robust GMM based Gender Classification using Pitch and RASTA-PLP Parameters of Speech.*Proceedings of the 2006 International Conference on Machine Learning and Cybernetics*, 3376-3379.

[5] Sarhan, S. and Elsoud, M. A. (2015). Text Independent Speaker Identification based on MFCC and Deep Neural Networks Text Independent Speaker Identification based on MFCC and Deep Neural Networks.

[6] Khojah, B. Z. and Alhalabi, W. S. (2017). Text-Independent Speaker Identification System Using Different Pattern Matching Algorithms, 4(1), 22-29.

[7] GitHub - SuperKogito/Voice-based-gender-recognition: Voice based Gender Recognition using Mel-frequency Cepstrum Coefficients (MFCC) and Gaussian Mixture Models (GMM).

[8] Sukhwal, A. and Kumar, M. (2016). Comparative Study of Different Classifiers based Speaker Recognition System using Modified MFCC for Noisy Environment. *Proceedings of the 2015 International Conference on Green Computing and Internet of Things, ICGCIoT 2015*. Institute of Electrical and Electronics Engineers Inc. 976-980.

[9] Voice Gender Detection using GMMs : A Python Primer - Machine Learning in Action.

[10] Bagul, S. G. and Shastri, R. K. (2013). Text Independent Speaker Recognition System using GMM. *2013 International Conference on Human Computer Interactions, ICHCI 2013*. IEEE Computer Society.

[11] Machine Learning with Python: Expectation Maximization and Gaussian Mixture Models in Python.

[12] Sinith, M. S., Salim, A., Sankar K, G., Narayanan K V, S. and Soman, V. (2010). A Novel Method for Text-independent Speaker Identification using MFCC and GMM', *ICALIP 2010 - 2010 International Conference on Audio, Language and Image Processing, Proceedings*. IEEE, 292-296.