# Building Speaker Identification Dataset for Noisy Conditions

Win Lai Lai Phyu
Natural Language Processing Lab.,
University of Computer Studies
Yangon, Myanmar
*winlailaiphyu@ucsy.edu.mm*

Win Pa Pa
Natural Language Processing Lab.,
University of Computer Studies
Yangon, Myanmar
*winpapa@ucsy.edu.mm*

## Abstract

*Speech signal processing plays a crucial role in any speech-related system whether Automatic Speech Recognition or Speaker Recognition or Speech Synthesis or something else. Burmese language can be considered as an under resourced language due to its linguistic resource availability. For building Burmese speaker identification system, the sufficient amount of speech data collection is a very challenging task in a short time. In order to get higher data size, this paper analyzes that the getting higher duration of speech data actually combining with various noises encountering in our surroundings. For increased noisy state speech data, we also used the voice activity detection (VAD) technique to acquire only the speaker specific information. For feature extraction, we used MFCC, Filter Banks and PLP techniques. The experiments were developed with i-vector methods on GMM-UBM together with PLDA and presented the performance of different data set in the form of EER with two models trained on clean and noisy data to prove that the developed speaker identification system is noise robust.*

***Keywords***— *Burmese Speaker Identification, noise robustness, VAD, MFCC, Filter Banks, PLP, GMM-UBM, PLDA*

## I. INTRODUCTION

The speech of living things especially for human involves numerous discriminative acoustic features that can be discerned who they are because of the structural formation of vocal tract is unique for everyone. Speaker identification is the process by which the acoustic speech signals to its corresponding speaker and is applied in many applicable areas. Speech corpora collection is the very first step in building speaker identification system. In order to develop the speaker identification system, the sufficient speech data are needed to train and test the spoken speech data. The performance of the system is also depended on the amount of speech data. There are many variations in speaker identification system. The first one is the duration of utterances. The longer the utterances, the better recognizes the corresponding speaker. A second variation is noise as any kinds of noise make the identification process harder. The third variation is accent or speaker specific facts. The speaker is easier to identify if he/she speaks a standard dialect or the ones that matches the speech data the system trained on. The final variation is the speech recorded conditions. Therefore, we will propose Burmese speaker identification how to construct with noisy data. The paper is organized as follows. Related works will be presented in section II. In section III, speaker identification process will be introduced. Section IV will be described the types of speaker recognition and the proposed architecture of speaker identification system with noisy data will be expressed in section V. Experimental setup will be addressed in section VI and experimental results will be discussed in section VII. Finally, conclusion will be expressed in section VIII.

## II. RELATED WORKS

There are many speaker recognitions with various approaches found in publications.

Arnab Poddar, Md Sahidullah and Goutam Saha [2] presented the comparison of two different speaker recognition systems, i-vector based and GMM_UBM in utterance duration variability. It revealed that GMM_UBM system outperforms i-vector system for very short test utterances if the speaker are enrolled with sufficient amount of training data whereas total variability (i-vector) based system degrades with the reduction in test utterance length and also require huge computational resource development data for identifying the speaker although GMM_UBM don't require the huge amount of development data.

Comparison of text independent speaker identification systems using GMM and i-vector

methods are done by Nayana P.K., Dominic Mathew, and Abraham Thomas [3]. It was observed that appending formants and pitch high level features to basic features: PNCC (Power Normalized Cepstral Coefficients) and RASTA PLP (Relative Spectral PLP) obtain the better accuracy for speaker identification. It was also showed that the accuracy of i-vector method with PLDA classifier is better than that of Cosine Distance Scoring (CDS) classifier. Moreover, it revealed that the system performance enhances when longer utterances are used.

Analysis of various feature extraction techniques for robust speaker recognition was presented by Qin Jin and Thomas Fang Zheng [4] to help the researchers for catching the current front end features classified as low level and high level features. They surveyed the speaker recognition system on different feature extraction techniques: MFCC, MVDR, FDLP, MHEC, SCF/SCM, FFV, HSCC and Multitaper MFCC and presented the strength and weakness of these techniques.

R.ARUL JOTHI M.E [5] presented the analysis of suitable extraction methods and classifiers for speaker identification since 2017. It identified the speaker's voice whether original or disguised voice based on MFCC, Delta MFCC and Delta-Delta MFCC with SVM classifier. MFCC with SVM classifier improves the performance of system and accuracy rates up.

An improved approach for text independent speaker recognition was proposed by Rania Chakroun [6]. It proposed that the new feature extraction method combining MFCC and Short Time Zero Crossing Rate (ZCR) of the signal. ZCR is the number of times the zero axes crossed by the signal per frame. By comparing the performance of two speaker recognition systems with the use of MFCC and combination of MFCC and STZCR, it showed the new proposed feature extraction yields better outcome and reduced in EER.

## III. SPEAKER IDENTIFICATION

Speaker identification determines the speaker identity from which of the registered speakers a given utterance comes. It is a very challenging task because human speech signals are highly variable due to various speaker characteristics, different speaking styles, environmental noises, and so on. There exist various feature extraction methods and approaches for speaker identification system. There are three main steps in speaker identification.

As part of feature extraction, a set of feature vectors are obtained from the raw speech signal to more emphasize the speaker related information because the speech signal can contain many features which are not required to claim the speaker. Therefore, feature extraction process is advantageous when you need to diminish the resources required for processing without losing relevant information. There are many different types of features that can be extracted. The recognition accuracy rate varies according to chosen extraction methods. Various types of available features are high level features, spectra-temporal features, short term (low level) spectral features, and prosodic features [3]. In these, low level features are easy to extract and very effective to recognize the speaker. Although high level features contain more speaker specific information, the extraction process is more complicated.

To generate the speaker models representing to each speaker, the features attained from the feature extraction stage are used and stored these speaker models into a database as UBM for performing the comparison during testing. It is the main session of speaker identification system as the models created in this stage are applied to perform comparison in the identification stage. Different modeling methods are HMM (Hidden Markov Models), GMM (Gaussian Mixture Models), DNN (Deep Neural Network), and i-vector method.

Identifying the test speech signal is the final stage of every speaker identification system. Relative scores corresponding to each of the speaker models are computed and then the one which has the highest score is identified as the target speaker. Different scoring methods used for identification are CDS (Cosine Distance Scoring), PLDA (Probabilistic Linear Discriminant Analysis), LLR (Log Likelihood Ratio), and SVM (Support Vector Machine) and so on.

## IV. TYPES OF SPEAKER RECOGNITION

There are two types of speaker recognition: speaker verification and speaker identification. If the speaker claims to be of a certain identity and the voice is used to verify this claim, this is called verification or authentication. Speaker verification is a 1:1 match where one speaker's voice is matched to one template (also called a "voice print" or "voice model"). Speaker identification is the task of determining an unknown speaker's identity. Therefore, it is a 1: N match where the voice is compared against N templates. It involves

two phases: enrollment and testing. During enrollment, the speaker's voice is recorded and typically a number of features are extracted to form a voice print, template, or model. In testing phase, a speech sample or "utterance" is compared against a previously created voice print. Moreover, there exist two types of speaker identification: text dependent and text independent. Text dependent speaker identification needs to utter exactly the same utterance to determine who they are. Text independent speaker identification has no limits and constraints on the spoken words that are uttered. It is more flexible and usable in real world applications. Verification is faster than identification because of the processing time of matching. This paper proposes text independent speaker dependent identification because text independent systems are most applicable in real world.

## V. Proposed Architecture Of Speaker Identification System

Feature extraction, speaker modeling and identification process are the three crucial stages of any speaker recognition system. This section presents the detail description of the whole speaker identification process. The proposed architecture of speaker identification system with noisy data exhibits in Fig. 1.

### *A. Data Preprocessing*

Before feature extraction, we need to firstly preprocess the data. In this stage, the whole recorded audio speech data is chopped into utterance level speech segments with ***Audacity*** which is open source and cross-platform audio multi-track audio editor and recorder software. And then, the preprocessed speech data are the utterance level speech segmented data with 16 bits mono PCM in 16 kHz ranging from 10 to 27 seconds. After that, we randomly added the utterance level segmented speech data with various noises found in our surrounding prepared by ourselves. By contaminating, our data set size increases the duration than that of original clean data set size. And then, the original clean data and noise-combined data are combined to use in feature extraction. Surrounding noises include car's horn, fly buzz, the dog's bark, fire alarm, ringtone, cat meow, whistle, roar, shouting, wind blowing, birds chirping, banging of hammer, school bell, beating a drum and so on.

### *B. Feature Extraction and Voice Activity Detection*

There is no standard rule for choosing among these features for the question 'Which feature extraction technique should one use?' It depends on our needs like intended application, robustness, computing resources and amount of data available. Because short-term spectral (low level) features are easy to compute and provide good results, exploring with these types of feature enhances the system performance. Feature extraction stage is one of the most important components in any SR systems and its objective is to find robust and discriminative features in acoustic data because better features give the more improved recognition rate. In our proposed system, clean and noisy data are combined to extract the features. And then, Voice activity detection (VAD) is applied for noisy data. It is a technique used to detect the speech or non-speech section in recorded speech data with the aim of removing the silence frames in segmented speech data, saving the computing time and enhancing the recognition accuracy rate. It also refers to the problem of distinguishing speech segments from background noise in an audio stream and is also language independent. Moreover, we exploited with three kinds of low level feature extraction techniques for system performance: Mel Frequency Cepstral Coefficient (MFCC), Filter Bank and Perceptual Linear Prediction (PLP). The system's recognition rate diverges depending on our choice because there are no standard rules for choosing among these features. It depends on our destined needs in related applied areas.

### *C. Building Speaker Models*

The feature vectors extracted in the feature extraction stage take to build the speaker models. UBM is the key element of an i-vector (existing in low dimensional spaces that are smaller in size to reduce the recognizing time) system as it is necessary for collecting statistics from speech utterances. It is constructed using feature values of sound speech samples from the different speakers and Maximum A Posteriori (MAP) is used to get the speaker models each [1]. It is the central part of this system because it is used in comparison with the test speech segment's feature vector for describing who the speaker is. In this paper, we implement the speaker identification system with i-vector method by using Kaldi ASR open source toolkit to build the speaker models:

Model in Clean Data (Model_1) and Model in Noisy Data (Model_2) [8].

### *D. Identification Process*

Different scoring methods: support vector machine (SVM), Probabilistic Linear Discriminative Analysis (PLDA), and Cosine Distance Scoring (CDS) for identifying the speaker are applied in any speaker identification system. In this paper, i-vector based speaker identification with PLDA is put to work for recognizing the noisy test speech input signal in the sense of three feature extraction techniques on two speaker models: Model_1 and Model_2 for verifying our proposed system, Model_2 is noise robust. PLDA method used in this paper is the simplified or Gaussian PLDA with 200 Gaussian components of 100 dimensions in i-vectors. It is computed the similarity scores as the ratio of the probability that both test and reference i-vector belong to the same speaker to the probability that they both belong to different speakers.

## VI. EXPERIMENTAL SETUP

### *A. Data Preparation*

The experiment is implemented with total number of 37 speakers including 12 male and 25 female speakers. There are about 7 hours of clean data comprising of 2516 utterances in clean data and 16 hours of noisy training data set comprising of 5032 utterances in noisy data. It has double the size of data than the clean data. The development data are about 54 minutes in the clean data with 321 utterances and nearly 2 hours of noisy data with 642 utterances. In this paper, we will do the experiments with two test sets: TestSet1 and TestSet2. For TestSet1 and TestSet2, there exist 111 utterances of clean and noisy test sets with the length of 18 seconds and 23 seconds each. These are evaluated based on Model_1 and Model_2 for approving the noise robustness of the model trained with noisy data. Speech data utterances were recorded at 16 bits mono PCM in 16 kHz with the duration of ranging from 10 to 27 seconds each. This frequency rate affects in the feature extraction process and building the speaker models because this rate is suitable for Myanmar's spoken speech tone. The total number of speakers included in this experiment is shown in table I. Data preparation for Model_1 and Model_2 is shown in table II. The experiment using clean data for Model_1 is taken from [7]. Moreover, for Model_2, we randomly reckon additional noise to the original clean data. Test case preparation for TestSet1 and TestSet2 is shown in table III. TestSet1 is one which contains original clean test data and TestSet2 is the test data that randomly combines additional noise to the original clean data.

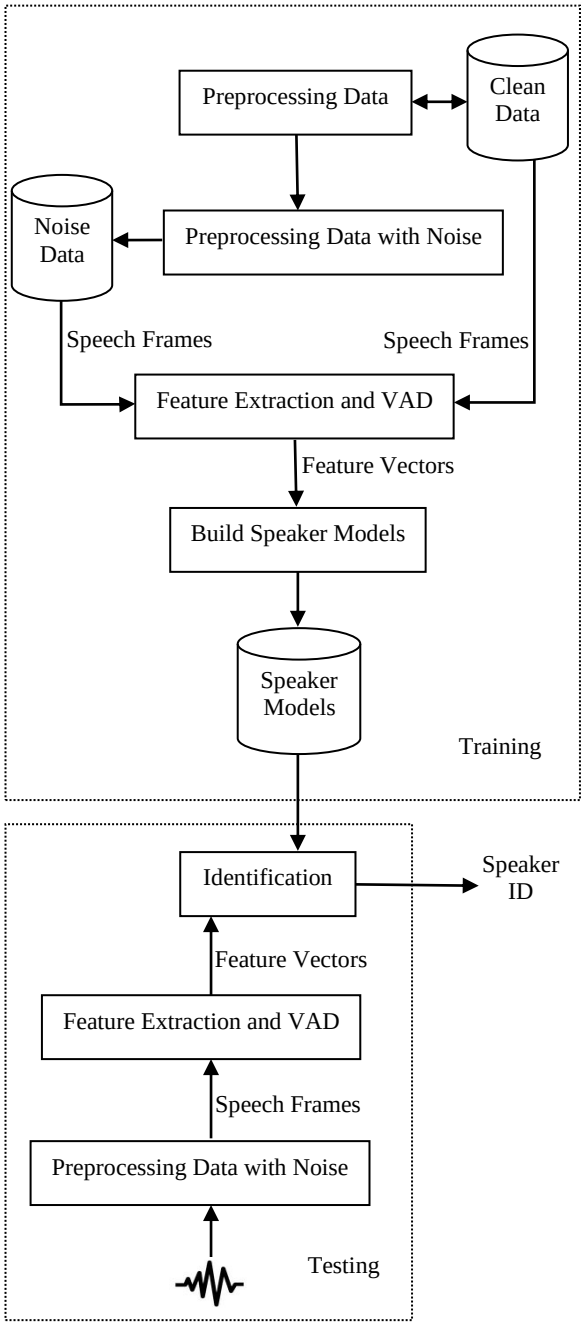


**Figure 1. Proposed architecture of noise robust speaker identification system**

### *B. Evaluation: Equal Error Rate(EER)*

We appraised automatic evaluation of equal error rate (EER) for assessing the performance of speaker identification models in both conditions. False

acceptance rate (FAR) (1), is a type of error allowing the impostor speaker is improperly identified as the known speaker and false rejection rate (FRR) (2) is incorrectly denied the actual speaker known by the system as impostor. Equal Error Rate (EER) is one where FAR equals to FRR and also the point where FAR and FRR are optimal and minimal. EER of speaker model is mainly based on the amount of training data. Therefore, we are going to collect the speech data more and more in the future. The lower the EER value, the higher the recognizing rate of speaker models.

**TABLE I. TOTAL NUMBER OF SPEAKERS**

| Male | Female | Total Number of Speakers |
|---|---|---|
| 12 | 25 | 37 |

**TABLE II. DATA PREPARATION FOR THE EXPERIMENTS**

| Data | Number of Utterances | | Duration (hr: min: sec) | |
|---|---|---|---|---|
| | *Model_1* | *Model _2* | *Model _1* | *Model _2* |
| Train | 2516 | 5032 | 07:10:39 | 16:05:04 |
| Dev | 321 | 642 | 00:54:36 | 01:58:35 |

**TABLE III. STATISTICS OF THE TEST SETS**

| Test Sets | Number of Utterances | Duration (hr: min: sec) |
|---|---|---|
| TestSet1 | 111 | 00:18:26 |
| TestSet2 | 111 | 00:23:12 |

$$FAR = \frac{Total\ False\ Acceptance}{Total\ False\ Attempts} \quad (1)$$

$$FAR = \frac{Total\ False\ Rejection}{Total\ True\ Attempts} \quad (2)$$

### C. *Evaluation:Test Samples' Accuracy*

To evaluate the performance of every test speech samples, we also applied the automatic evaluation shown in (3). This automatic evaluation is based on how many test speech samples recognized by the speaker models differs from the correct test speech samples.

*Accuracy (%) = ((TTSs – WDSs) / TTSs) * 100* (3)

where, *Accuracy*=Test Case's Accuracy in Percentage

*TTSs*=Total Test Speech Samples in Test Case

*WDSs* = Wrong Detected Samples .

## VII. EXPERIMENTAL RESULTS

We will describe the experimental results based on the models built in Model_1 and Model_2 using PLDA identification. Table IV shows the performance of speaker models, Model_1 and Model_2 on the development data, and two test sets. To evaluate the performance of models, TestSet1 and TestSet2 is used for assessments in terms of equal error rate (EER%). TestSet1 is the original clean test data and TestSet2 is the data prepared with additional noise to the clean data. The performance of speaker models in varying surrounding conditions on the development data sets depicts with a chart in Fig. 2.

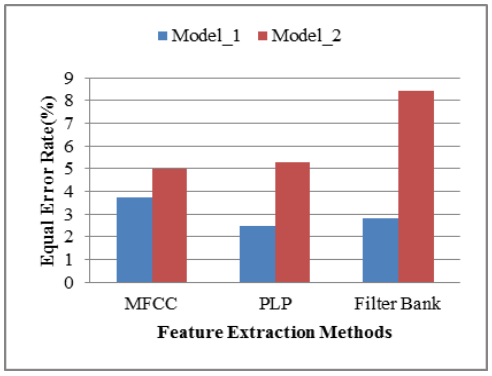


**Figure 2. Performance of EER on building speaker models**

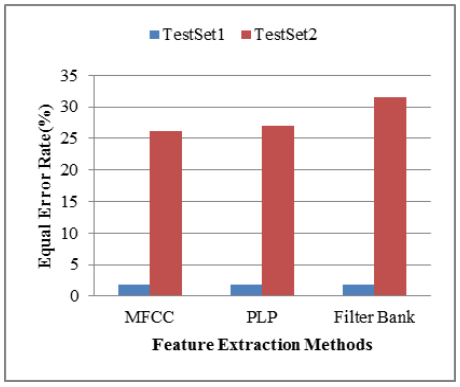


**Figure 3. EER of TestSet1 and TestSet2 on Model_1**

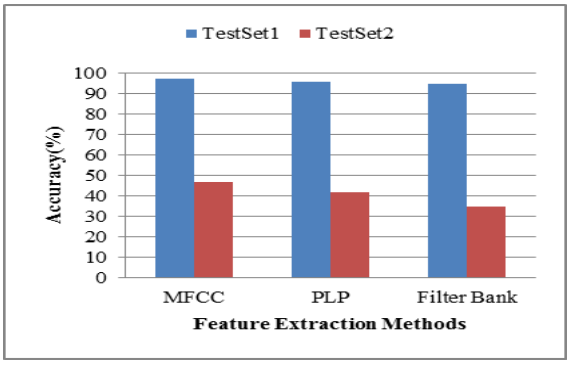


**Figure 4. Performance of Model_1 with three feature extraction methods**

Figure 3 and Figure 4 reveals that the evaluation results of TestSet1 and TestSet2 on speaker models built in clean data. Although the EERs and accuracies of models show good results on clean test data, the performance degrades on noisy test data. This experiment approves the noisy data needs to be trained on the noisy data.

Table IV shows the performance of two models in terms of EERs on two test sets: TestSet1 and TestSet2. It can be seen that the EER of clean data, TestSet1 has got comparable results in both clean and noisy models. It means, the noisy model can give the similar performance on clean and noisy test data. EER of noisy model, Model_2, was decreased significantly on TestSet2, noisy data, by all feature extraction techniques at most 20.72% than clean model, showing the noisy training data are important for noisy test data. From the experiments, we found MFCC give better results among three feature extraction techniques for clean and noisy conditions of our data.

**TABLE IV. PERFORMANCE OF TESTSET1 AND TESTSET2 ON MODEL_1 AND MODEL_2**

| Feature Extraction Methods | Equal Error Rate (%) | | | | | |
|---|---|---|---|---|---|---|
| | *Dev* | | *TestSet1* | | *TestSet2* | |
| | *Model_1* | *Model_2* | *Model_1* | *Model_2* | *Model_1* | *Model_2* |
| MFCC | 3.738 | 4.984 | 1.802 | 1.802 | 26.13 | 8.108 |
| PLP | 2.492 | 5.296 | 1.802 | 2.703 | 27.03 | 10.81 |
| Filter Bank | 2.812 | 8.424 | 1.802 | 2.703 | 31.53 | 10.81 |

**TABLE V. ACCURACIES ON TESTSETS (%)**

| Feature Extraction Methods | Accuracy Rate (%) | | | |
|---|---|---|---|---|
| | *Model_1* | | *Model_2* | |
| | *TestSet1* | *TestSet2* | *TestSet1* | *TestSet2* |
| MFCC | 97.27 | 46.36 | 93.64 | 76.36 |
| PLP | 95.45 | 41.82 | 92.73 | 75.45 |
| Filter Bank | 94.55 | 34.55 | 94.55 | 76.36 |

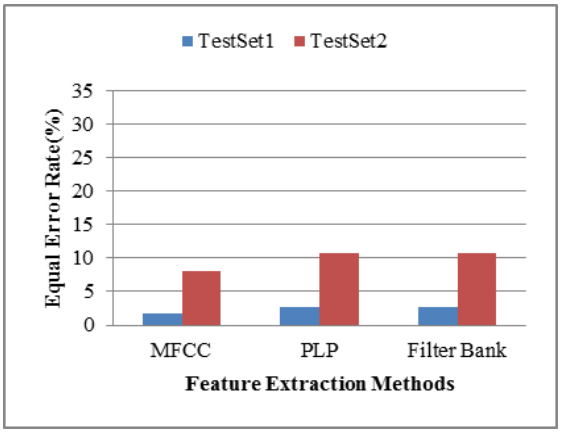


**Figure 5. EER of TestSet1 and TestSet2 on Model_2**

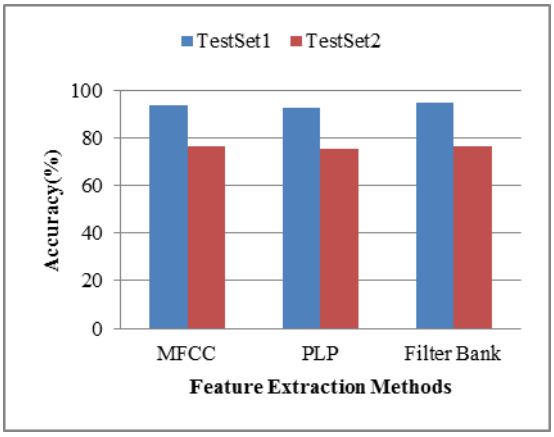


**Figure 6. Performance of Model_2 with three feature extraction methods**

Fig. 5 and Fig. 6 show that the results of TestSet1 and TestSet2 on noisy speaker models. This paper approaches to the point of view of every noise aggravates the recognition rate on every speech-related processing. The error rates of clean data are sharply higher than that of testing on noisy data.

As Fig. 3, Fig. 4, Fig. 5, Fig. 6, Fig. 7 and Fig. 8 show, it can be seen clearly that the model with noisy data is better than Model_1 in noisy condition and noisy data helps to improve the performance of

speaker identification in both clean and noisy condition. According to the Table IV and V, TestSet2 of on Model_1 and Model_2, show the improved equal error rate on our model, Model_2. The system performance degrades in clean data when testing with noisy test speech data but Model_2 yields satisfiable results on both conditions clean and noisy. In this analysis, we showed the error rates are obviously decreased almost one-third by Model_2 compared to Model_1.

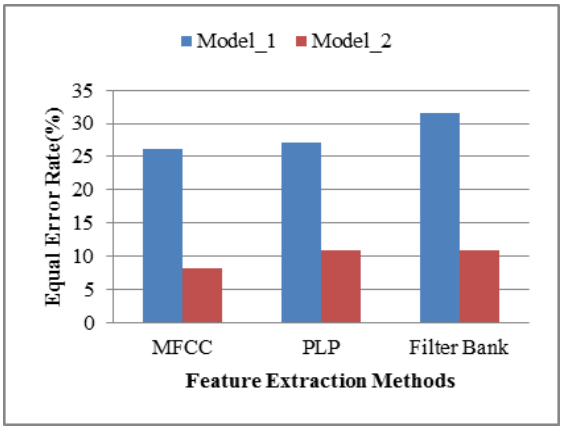


**Figure 7. EER on two models with TestSet2**

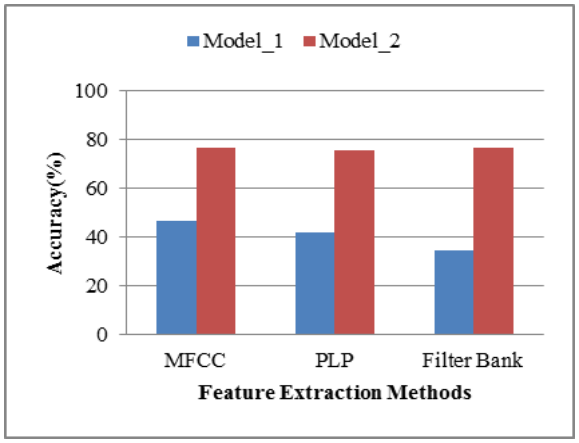


**Figure 8. Performance of TestSet2 on two models with three feature extraction methods**

## VIII. CONCLUSION

Noise delivers detriments in speech-related processing. This paper shows the importance of noisy data preparation for speaker identification. Experiments were done on speaker identification training on clean and noisy data. We also analyzed the rate of change of EER with three feature extraction methods in the experiments. From the experiments, applying MFCC features gave the best results among three different feature extraction techniques. We also found that, integrating additional noise to the original clean data improves the recognizing rate in every feature extraction method with increasing the size of data. From the experiments, it is clear that the preparation of noisy data is effective for noise robust speaker identification system and the results are acceptable.

## REFERENCES


**[1]** Win Lai Lai Phyu, and Win Pa Pa, "Text Independent Speaker Identification for Myanmar Speech", ICFCC 2019, Yangon, Myanmar, 27-28 February 2019.

**[2]** Arnab Poddar, Md Sahidullah, Goutam Saha, "Performance Comparison of Duration Variability", IEEE 2015.

**[3]** Nayana P.K., Dominic Mathew, Abraham Thomas, "Comparison of Text Independent Speaker Identification Systems using GMM and i-vector Methods", ICACC 2017, Cochin, India, pp.22-24 August 2017.

**[4]** Qin Jin, Thomas Fang Zheng, "Overview of Front-end Features for Robust Speaker Recognition", APSIPA ASC 2011, Xi'an, China.

**[5]** R.ARUL JOTHI M.E, "Analysis of Suitable Extraction Methods and Classifiers for Speaker Identification", IRJET 2017, Volume: 04 Issue: 03, Mar 2017.

**[6]** Rania Chakroun, Leila Beltaifa Zouari, Mondher Frikha, "An Improved Approach for Text-Independent Speaker Recognition", IJACSA 2016, Vol. 7, No. 8, 2016.

**[7]** Aye Nyein Mon, Win Pa Pa, Ye Kyaw Thu, "Building HMM-SGMM Continuous Automatic Speech Recognition on Myanmar Web News", ICCA 2017, Yangon, Myanmar, 16-17 February, 2017.

**[8]** Daniel Povey, Arnab Ghoshal, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, Karel Vesely, "The Kaldi Spech Recognition Toolkit", ASRU, 2011.