Improving the Performance of Low-resourced Speaker Identification with Data Preprocessing

Win Lai Lai Phyu[1,*], Hay Mar Soe Naing[2] & Win Pa Pa[3]

Natural Language and Speech Processing Lab.,
University of Computer Studies, Yangon.
No. (4) Main Road, Yangon, 11411, Myanmar
*E-mail: winlailaiphyu@ucsy.edu.mm

Abstract. Automatic speaker identification is done to tackle daily security problems. Speech data collection is an essential but very challenging task for under-resourced languages like Burmese. The speech quality is crucial to accurately recognize the speaker's identity. This work attempted to find the optimal speech quality appropriate for Burmese tone to enhance identification compared with other more richy resourced languages on Mel-frequency cepstral coefficients (MFCCs). A Burmese speech dataset was created as part of our work because no appropriate dataset available for use. In order to achieve better performance, we preprocessed the foremost recording quality proper for not only Burmese tone but also for nine other Asian languages to achieve multilingual speaker identification. The performance of the preprocessed data was evaluated by comparing with the original data, using a time delay neural network (TDNN) together with a subsampling technique that can reduce time complexity in model training. The experiments were investigated and analyzed on speech datasets of ten Asian languages to reveal the effectiveness of the data preprocessing. The dataset outperformed the original dataset with improvements in terms of equal error rate (EER). The evaluation pointed out that the performance of the system with the preprocessed dataset improved that of the original dataset.

Keywords: *Burmese speech dataset; data preprocessing; Mel-frequency cepstral coefficients (MFCCs); multilingual speaker identification; time delay neural network (TDNN).*

1 Introduction

Speech has acoustic properties that are unique to every individual such as the way of speaking characterized by accent, rhythm, intonation, pronunciation pattern, style, vocabulary and much more. In speech recognition tasks, researchers apply spoken languages to autonomous computerized systems. The quality of speech data is at the heart of every speech processing research because it plays a key role in obtaining reasonable outcomes. In speaker identification, the biometric

Received February 16th, 2023, Revised October 2nd, 2023, Accepted for publication 13 November 2023.
ISSN: 2337-5787, DOI: 10.5614/itbj.ict.res.appl.2023.17.3.1

recognition method is the task of deciding an unknown speaker to the corresponding speaker's identity based on a given speech recording. Using words like passwords, and digits has limitations in text dependent speaker identification. If the words contained in the training process are not used in testing phase, the identification will not be correct. However, in text- independent speaker identification, there are no constraints and limitations on the words that are uttered. The speakers can test the system by speaking freely. Therefore, this work implemented text-independent open-set speaker identification because it is more adaptable in real-world applications. In order to implement acoustic models, high quality speech data is needed as it is crucial to improve the performance of the models. For text-independent open-set identification, there are no publicly available speech datasets for under-resourced languages such as Burmese, even though well-resourced speech data are easily available. Therefore, we first collected data from online resources and we also recorded daily conversational data ourselves with microphones and telephones to create a Burmese speech dataset.

Moreover, it is also necessary to enhance the quality of the data for correct identification because the data collected from various available sources may not be completely clean. For example, disturbance by environmental noise may be present in some segmented speech utterances. Obtaining immunity from noisy data in building an acoustic model has been experimented in [1]. This work achieved a satisfying outcome for noisy conditions.

Collecting Burmese speech data and exploring how to get a high-quality speech dataset by data preprocessing were the two main goals of this study. The speaker models were built on TDNN with features of 39-dimensional Mel-frequency cepstral coefficients (MFCCs). The effectiveness of the proposed preprocessing methods was evaluated not only for Burmese speech but also for nine other Asian languages. The results showed that the speaker models based on the speech dataset with the proposed preprocessing methods achieved a more competitive performance in recognizing speaker identity than that based on the original dataset.

This paper is organized as follows. Section 2 describes the collection of the speech data and how to scrutinize the original data. The methodology used is explained in Section 3. The experimental setup, results and analysis are shown in Section 4, and the presented and future work are summarized and analyzed in Section 5.

2 Data Preprocessing

The main task at the front end is to prepare the data for well performing the next processing steps. Nowadays, researchers are trying to investigate and analyze their research's performance not only from a theoretical point of view but also from the point of view of the data quality in order to improve system performance. Doing data collection as well as possible yields more reasonable results and enhances system performance. If the data used in a research work has high quality, the next processes after data collection are made easier and the system becomes more robust. Therefore, well preparing and preprocessing the dataset was a main task in this work. This section discusses about collecting the speech data and how to prepare the collected data to get optimal results.

2.1 Collecting the Speech Data

Collecting speech data is the preceding step in any statistical based speaker identification task, particularly for under-resourced languages. A main problem when it comes to speaker recognition research is the lack of proper data when there are no existing speech datasets for low-resourced tonal languages like Burmese, even though datasets for resource-rich languages like English are easily available. Therefore, a speech dataset needs to build first. In the present study, the speech data were collected in two ways. The first was the collection of recorded data in the forms of news, delivered speech and talks taken from publicly available sources[1] with different audio formats. The collected data from internet sources has a clear, accurate and qualified tone because the speakers are well-educated and professional. Both local and foreign news about politics, health, sport, speech, education, crime, business news and weather, etc., are contained in the collected speech.

The second way of collecting speech data was by making recordings ourselves based on transcriptions of daily conversational dialogue text corpus spoken in restaurants, hotels, parks, and while traveling. The recordings used the same facilities as in our previous work [1]. The script for the recordings was collected from the dataset of the Uni-Trans project [2]. The recording was done by 24 internship students from three academic years and 25 lab members. The data were recorded in a quiet recording studio located in University of Computer Studies, Yangon. There were no external disturbances such as environmental noise or the room's echo. The recording was done with a Tascam DR-100MKIII[2] device

[1] Voice of America (VOA) Burmese, Democratic Voice of Burma (DVB), British Broadcasting Corporation (BBC) Burmese news, Radio Free Asia (RFA), Irrawaddy Burmese News, Mizzima News Myanmar, One News Myanmar Channel, 7days TV, Myanmar Radio and Television (MRTV), and Eleven broadcasting media

[2] https://tascam.com/us/product/dr-100mkiii/top

recommended to be used by audio engineers. The average length of daily conversational dialogue was 12 seconds which is shorter than the public data which had an average length of 30 seconds.

The quantity of public data exceeded that of our own recorded data, primarily due to inherent challenges in the acquisition of the latter. While obtaining a speech dataset from public sources poses no significant difficulties, self-collected datasets face various impediments. Notably, the internship students tasked with audio recording lacked prior experience in this domain, resulting in inconsistent tone and pace in their recordings. To mitigate this issue, multiple recording attempts were often necessary to attain a clear audio file. Furthermore, our own recorded data closely adhered to the Burmese accent, as it is intended for Burmese speaker identification. In terms of data quality, the self-recorded data were superior in clarity compared to the public data, some of which contained environmental background noise. Additionally, the linguistic diversity within our nation, characterized by 135 distinct ethnic groups, posed a significant challenge. Given the limited access to national websites, it was not feasible to collect data encompassing all these accents and speaking styles. Consequently, this limitation could potentially impede system performance and result in less control over the diversity of accents and speaking styles in the dataset.

2.2 Preprocessing the Speech Data

The quality of speech is important for more accurate recognition, while a sufficient amount of speech data is also needed in every speech processing task. There are many audio augmentations available related to tempo, speed, volume, etc., [3]. This section points out of the task of preprocessing the speech data in order to get the right speech quality to achieve an effective performance. Two preprocessing methods were first applied to the Burmese (mm) dataset on TDNN to investigate whether the recognizing rate increased or not. After that, they were applied to nine other languages: Arab (uz), Chinese (zhCN), English (eng), French (fr), Hindi (hi), Japanese (ja), Tamil (ta), Thai (th), and Vietnamese (vi).

2.2.1 Analysis of Preprocessing with Different SNRs

Changing the speech intensity can affect the performance of speech processing tasks [4]. Various intensity rates have been analyzed to see whether they improves speech recognition or not [5, 6]. Firstly, the intensity of each collected speech segment was preprocessed by calibrating with different SNRs levels (-10 dB, -5 dB, 5 dB, 10 dB) because the loudness of the sound waves collected from various sources can have different intensity levels. The data prepared with different dBs were experimented with by setting the intensity of all speech segments to the same SNR uniformly to know which dB scale is most appropriate [7]. According to the experiments, setting the intensity to 10dB on all speech segments gave the promising result of nearly 17% relative improvement on TDNN compared to the performance of the original collected data. There was no relative improvement when lowering the intensity to -10 dB and -5 dB. This caused a decrease of the system performance by nearly 5% for -10 dB and over 1% for -5 dB because decreasing the intensity level compared to that of a normal tone does not enable to distinguish spoken utterances better in tonal languages like Burmese. When raising the intensity to 5 dB and 10 dB, the relative improvement was over 2% and 16% respectively. Setting the intensity beyond 10 dB may change the tone color of speech utterances compared to the original utterance. Therefore, the experiments in this study were done by setting the maximum intensity level at 10 dB in this work.

Figure 1 shows a graphical representation of preprocessing the intensity on TDNN with layer-wise network content [t-8, t+8] in terms of equal error rate (EER). In Section 4.2, the reason why the layer-wise network content [t-8, t+8] was chosen is explained in detail.

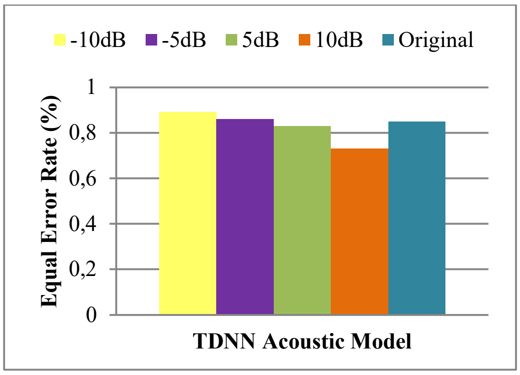

Figure 1 EER performance of speaker models on Burmese datasets with various SNRs.

2.2.2 Analysis of Preprocessing with Different Tempo

The second preprocessing method is based on a detailed analysis of changing the tempo of speech segments up or down. Text that is spoken too fast may not be recognized correctly, while at a slower pace it may be recognized better. In this work, changing the tempo factor rather than the speed factor was investigated because analyzing the tempo factor does not change the pitch of speech segments. Changing the speed factor influences both the tempo and the pitch, which causes the spectral shape of the speech segments to change as well. This may lead to losing the speaker's specific information in speech segments.

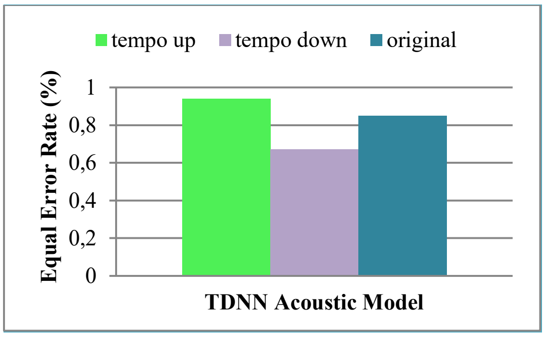

Figure 2 EER performance of speaker models on Burmese datasets with various tempos.

Figure 2 shows a performance comparison of analyzing the tempo factor only. Reducing the tempo by 0.2 times that of the original speech improved the results by nearly 27% relative to the normal or a higher pace. Increasing the tempo decreased the system performance by over 9%. The reason is that when the speech is slowed down, the system can catch more precisely what is being said, thus improving recognition.

Thus, according to the experiments, by playing the speech data more slowly, the system can recognize the speakers' specific information better and can better understand what is being said in terms of the clarity of the vocabulary. This can also improve automatic speech recognition.

3 Methodology

This section describes the detailed process for generating speaker models by a TDNN-based identification engine with parameter tuning, such as changing the layer-wise input network contexts. This is the most important part of every speaker identification system because the models built in this stage are employed to conduct feature matching in the identification phase.

3.1 Front-End Processing

This section explains the front-end processing, which is one of the major tasks in speaker identification. Front-end processing converts a high-quality speech signal to a low-quality speech signal subspace while keeping the distinctive features and vocal characteristics of the speaker [8].

3.1.1 Data Preprocessing

Data pre-processing is a crucial task in every speech-related system. If the data are prepared well, the results of the following processing steps will be more precise. As mentioned in Section 2.1, the raw Burmese language data were obtained from the Web and collected by ourselves. Being Web-based data, the collected videos had various formats (.mp4, .wma, .mp3) and frequency rates. These were converted to wave (.wav) file format uniformly. The converted wave files were formatted as 16-bit mono PCM with a frequency rate of 16 kHz. The frequency rate impacts the feature extraction process because it can increase the recognizing accuracy as it is most suitable for spoken tone and speaking rate. Each speech segment was split with Audacity3, an open source, cross-platform audio multi-track editing and recording software, to remove silent parts. For the nine other Asian datasets, the data were obtained from the Asian Multi-Speaker Verification (A-MSV) challenge of VLSP2022[4]. The data contained in these datasets came from YouTube and recorded common voices. The same preprocessing techniques as for the Burmese dataset were used for these datasets. All the speech data, including the Burmese dataset and the other nine languages' datasets, were combined to form the original training data.

As part of preparing the proposed preprocessed speech datasets, all speech data for the ten languages were preprocessed as mentioned in Section 2.2.

3.1.2 Feature Extraction

Extracting acoustic features from speech signals is a pivotal task and produces several different acoustic features of the speech signal, such as pace, pitch, and intensity. These speech features differentiate one speaker from others and its aim is to explore robust and discriminative phonetic features, providing an improved recognition rate on acoustic data. Short-duration low-level spectral features are more powerful and easy to extract than high-level features, which include more speaker-related information. In the latter case, however, the extraction process is more complicated and time-consuming when using low-level features. Therefore, 39-dimensional Mel-frequency cepstral coefficients (MFCCs) features were extracted every 10 ms in the frame size for short-time Fourier transform (STFT) of 25ms. A Hamming window [9] was used in the identification process to get better-quality speech features for the next processing step. This is a short-time cepstral representation of speech that is widely used as a feature in speech processing applications and human hearing system-based features [10]. Energy-based speech activity detection, VAD can also be used to remove features and posteriors corresponding to non-speech and silent frames order to increase the

[3] https://www.audacityteam.org/

[4] https://vlsp.org.vn/cocosda2022/a-msv

identification rate and save computation time. As VAD is language-independent, it can distinguish speech segments from background noise in audio streams.

3.2 Back-end Processing

The feature vectors extracted in the front-end processing stage are used to build the corresponding identification engine in the back-end processing stage with the aim of making decisions in the testing phase. In this work, a TDNN-based acoustic model was implemented as the identification engine.

3.2.1 Speaker Modeling

Time delay neural networks (TDNNs) have been found to be powerful in handling speech signals' context information and are designed to reveal a relation among inputs in real-time learning scale invariant feature transforms (SIFT). This is a well-performing architecture for DNN-based speaker identification systems because it models the phonetic content directly [11]. It is regarded as a precursor of convolution neural network because it is effective in capturing features from long-range temporal context dependencies [12] and improves the ability of x-vector learning by capturing more robust speaker characteristics [13]. The first layer process from narrow contexts is input into the speech signal. The deeper layers will process the input by splicing the output of the hidden activations from the previous layer to learn wider temporal dependencies [14]. In traditional TDNNs, splicing continuous windows of frames causes overlap and redundancy resulting in time-consuming model training. Therefore, a subsampling technique is applied to improve efficiency [13], allowing cracks between feature frames at each layer. This also decreases the number of parameters and increases the computational efficiency.

The TDNN architecture has three components: feature learning, statistical pooling, and speaker classification. In the present work, the feature learning involves five time-delay layers to learn frame-level speaker features designed to present the information to the model in a suitable form. Eight different slicing parameters were experimented with in the layer-wise context to investigate which network context impacts the model's efficiency. Among them, the slicing parameters for the five time-delay layers: {t-2, t-1, t, t+1, t+2}, {t-1, t, t+1}, {t-2, t, t+2}, {t-3, t, t+3}, {t} gives the optimal result in EER. Secondly, the statistical pooling evaluates the mean and standard deviation of the frame-level features from a speech segment. Activations from this pooling layer are then sent to the next nonlinear layers to discriminate the speaker at segment level. The third component, speaker classification distinguishes different speakers and has one fully connected layer corresponding to the number of speakers in the training data. Once trained, the 512-dimensional activations of the penultimate fully connected layer are extracted as an x-vector. These fixed-length feature vectors extracted per utterance can be used as input to the network.

These are the reasons why the TDNN architecture uses subsampling. If the system is trained by using a fully connected TDNN configuration without subsampling, the training time for the hidden layers of the network takes a long time [13]. To reduce the time complexity of a fully connected TDNN, optimization by excluding duplicate weights in the network was applied in this work. To reduce duplicate weights between nodes in the network, subsampling was used in TDNNs' implementation. It is expensive to compute hidden activations at all-time steps. Subsampling allows gaps between feature frames rather than splicing together contiguous temporal windows at each layer, which speeds up the training time and reduces the model size [13]. With subsampling, the overall necessary computation is reduced during the forward pass and back propagation due to the selective computation of time steps. Decreasing the number of parameters by minimizing the number of edges and nodes in the network and increasing the computational efficiency in training are the advantages of applying subsampling.

3.2.2 PLDA Backend

A simplified or Gaussian PLDA backend applicable to fixed-length feature vectors can be used for multiple input feature frames with subspace covariance modeling. This creates a hierarchical generative probability model, which puts the highly correlated feature vectors in subspaces. PLDA not only reduces the computation cost and feature dimension, but also improves the identification rate [15, 16]. It is used in feature vector-based speaker identification for scoring. The vector representations are centered and projected using linear discriminant analysis (LDA) to tune 512 dimensions. The representations are length-normalized and modeled by PLDA after dimensional reduction.

The log likelihood ratio is directly computed for the test case whether the two vectors (target and test) are or are not generated by the same speaker. The PLDA scoring method as implemented in [17, 18] is used. In two-vector scoring, the likelihood ratio is computed between a set of enrollment and test vectors. In the case of vectors v_1 for enrollment and v_t for testing, the PLDA scores $S(v_1, v_t)$ can be computed by determining the likelihood ratio given by the following equation:

$$S(v_1, v_t) = \log \frac{p(v_1, v_t \mid H_1)}{p(v_1 \mid H_0) p(v_t \mid H_0)} \tag{1}$$

Here, the hypothesis H_1 indicates that both vectors v_1 and v_t come from the same speaker while H_0 represents that both are independently drawn from different speakers.

For multisession scoring, the log likelihood ratio is computed for multiple test cases to detect the test which vectors belong to which speakers. This process is extended from two-vector scoring of multiple enrollment and test utterances. If multiple vectors $v_1, v_2, .., v_N$ are available for enrollment and $v_{t1}, v_{t2}, .., v_{tM}$ are available for testing, the log likelihood ratio can be computed with the following equation:

$$S(v_1,..,v_{t1},..)=\log\frac{p(v_1,..,v_N,v_{t1},..v_{tM}\,|\,H_1)}{p(v_1,..v_N\,|\,H_0)\,p(v_{t1},..,v_{tM}\,|\,H_0)} \tag{2}$$

4 Experiments

This section explains the experimental setup used in this work. The experiments were done to assess the quality of the preprocessed speech dataset by comparing it with the original collected speech dataset in performing speaker identification.

4.1 Experimental Setup

The original training dataset consisted of speech data from ten different Asian languages for multilingual speaker identification. Speech utterances in Burmese language were collected as mentioned in Section 2.1. The other nine languages' data were taken from the A-MSV challenge of OCOCOSDA 2022. The original training dataset comprising ten different languages was utilized with the two preprocessing methods mentioned in Section 2.2 to obtain the final dataset. After that, the original and preprocessed training datasets were obtained. There were 1,392 speakers in total involving 964 males and 428 females. The speakers' ages ranged from 20 to 70 years. Each speech segment contained in the dataset varied in length from 3~45 seconds. The speech utterances were formatted with a frequency rate of 16 kHz in 16-bit mono PCM.

Table 1 shows the detailed statistics of two speech datasets used in the experiments, with duration and total number of utterances. Each dataset was split into training, validation and testing data with proportions of 90%, 7%, and 3%, respectively, as shown in Table 1, according to the 224,212 total utterances. All speakers were included in training, validation and testing. This work aimed to achieve text-independent open-set identification. Although the natures of data training and testing are different, they come from the same speakers. Therefore, the system can accurately recognize the speaker's identity in various situations.

In comparing the duration of the two different datasets, the preprocessed dataset's length was longer than that of the original dataset because slowing down the speech tempo stretched the utterances' length in the original dataset. The Kaldi ASR open-source toolkit [19] was used to model the speaker models on a K80 GPU.

Table 1 Statistics of speech datasets.

Dataset	Original (hh:mm:ss)	Preprocessed (hh:mm:ss)	No. of Utterances
Training	279:33:34	323:06:33	201,790
Validation	10:52:51	13:10:24	15,695
Test set	05:12:17	06:29:07	6,727
Total	295:38:42	342:46:04	224,212

4.2 Experiment Results and Analysis

Firstly, the impacts of the two training datasets in different network contexts of TDNN with subsampling on the Burmese speech dataset were investigated. The results of the recognition rate of the speaker models are reported in terms of equal error eate (EER) of the the total test-set in percentage [20]. Table 2 shows eight different network input contexts of the layer-wise parameter tuning used in TDNN.

Table 2 Layer-wise context parameter tuning settings.

Network Context	Layer-wise Context				
	1	2	3	4	5
[-7,7]	[-2,2]	{-2,2}	{-3,3}	{0}	{0}
[-8,8]	[-2,2]	{-1,1}	{-2,2}	{-3,3}	{0}
[-9,7]	[-2,2]	{-2,2}	{-5,3}	{0}	{0}
[-10,6]	[-2,2]	{-2,2}	{-6,2}	{0}	{0}
[-10,8]	[-2,2]	{-1,1}	{-2,2}	{-5,3}	{0}
[-11,6]	[-2,2]	{-2,2}	{-7,2}	{0}	{0}
[-11,7]	[-2,2]	{-1,1}	{-2,2}	{-6,2}	{0}
[-12,7]	[-2,2]	{-1,1}	{-2,2}	{-7,2}	{0}

The preprocessed dataset for Burmese speech was first prepared by setting the SNR level to 10 dB and setting the tempo factor to 0.2 times (slowing 20%

compared to the normal pace) according to the comparable experimental results in Section 2 and implemented to investigate whether the performance improves or not on every layer-wise context of the network.

According to the aspects of the network contents defined in Table 2, the experiments were done on two types of training datasets comprising ten different languages. According to the experiments in Figure 3, the network contexts of [t-7, t+7], [t-9, t+7], [t-10, t+6] and [t-11, t+6] containing two fully connected layers in upper layers 4 and 5 obviously decreased the error rate value, with a relative improvement of over 15% when comparing the results of the original data with the preprocessed data. The network contexts of [t-8, t+8], [t-10, t+8], [t-11, t+7] and [t-12, t+7], which contain only one fully connected layer before producing the output decreased the error rate with a relative improvement of nearly 18% in [t-10, t+8], [t-11, t+7] and only over 10% on [t-12, t+7]. This is because the hidden layer 4 of [t-11, t+7] takes the time step of the upper and lower bound very far causing loss of specific information of the speakers. Among them, the reason why the network context of [t-8, t+8] was chosen as the optimal network context was that it obviously reduced the EER not only in the preprocessed dataset but also in the original dataset, with a relative improvement of over 25%. This is because the network takes the adjacent frames on narrow temporal contexts in the lower layers.

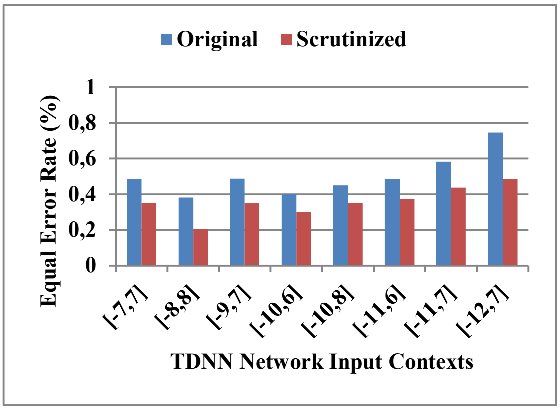

Figure 3 EER performance of Burmese dataset on different layer-wise network contexts of TDNN.

Figure 3 depicts the EER performance comparison that was assessed on the speaker models of the original and preprocessed Burmese datasets implemented on eight diversified layer-wise contexts of TDNN-based identification engines. In Figure 3, it can clearly be seen that the error rate of the preprocessed dataset decreased on every network context but the error rate of [-8, 8] context sharply decreased compared with the other network contexts. It can be seen that the proposed data preprocessing technique yielded comparable results on all of the speaker models, with a reduced error rate on each respective model.

Therefore, the original speeches of ten languages were combined into one to show the performance of multilingual speaker identification according to the effective results on the Burmese speech dataset depicted in Figure 3. Moreover, both preprocessing methods were employed to this combined training dataset to obtain the multilingual preprocessed dataset.

The recognition of ten Asian languages was implemented on the network context of [t-8, t+8] because it can be seen that the network context of [t-8, t+8] was the best temporal context compared with the other network input contexts. It splices the frames from wider temporal contexts in the upper layer of the network by incrementally processing the adjacent frames on narrow temporal contexts in the lower layer. For multilingual speaker identification, TDNN-based speaker models were built by using the two multilingual datasets (original and preprocessed) in order to see it the preprocessed dataset outperformed the original dataset. Moreover, the performance of the Gaussian Mixture Model-Universal Background Model (GMM-UBM) used in [21] was also evaluated, to see if the recognition rate of TDNN was more accurate than that of GMM-UBM.

According to the results in Figure 4, it is clear that the error rate of the preprocessed dataset in every language obviously decreased on both acoustic models (GMM-UBM and TDNN) compared to the original dataset. The experimental results of GMM-UBM are represented by red bars (original dataset) and blue bars (preprocessed dataset). The performance of the preprocessed dataset (blue bars) in GMM-UBM gave the satisfactory results with a relative improvement of over 22% for Arabic, 11% for Burmese, 15% for Hindi, 11% for Japanese, 9% for Thai, 7% for Vietnamese, nearly 5% for Chinese, 10% for English, 9% for Tamil, and exactly 50% for French compared with the error rates of the original dataset (red bars).

Moreover, the performance results on the TDNN depicted with purple bars (original dataset) and light green bars (preprocessed dataset) in Figure 4 are shown together to point out that the preprocessed dataset gave a higher recognition rate than the original dataset in every acoustic model, proving that TDNN outperformed GMM-UBM. In the aspect of TDNN, the performance of the preprocessed dataset (light green bars) was enhanced with a relative improvement of up to 27% for Arabic, over 46% for Burmese, 20% for Chinese, 24% for English, nearly 38% for French, 13% for Hindi, 18% for Japanese, 28% for Tamil, 9% for Thai and 28% for Vietnamese compared with the error rates of the original dataset (purple bars). It is obvious that TDNN outperformed GMM-

UBM. Therefore, the proposed data preprocessing technique had a satisfactory outcome with all speaker models and can greatly reduce the error rate of each respective model, especially on the preprocessed dataset with TDNN.

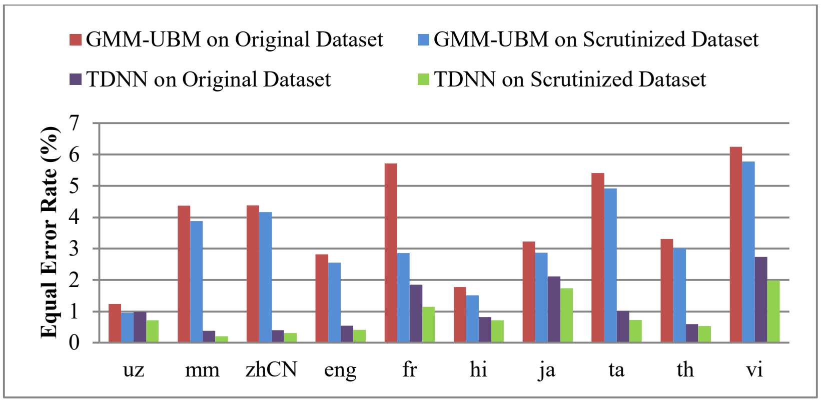

Figure 4 Performances of GMM-UBM and TDNN on the original and the preprocessed test sets.

5 Conclusion and Future Work

This paper investigated the effectiveness of preprocessed data in conducting the speaker identification and the impacts of reducing the error rate on various speaker models built with two different training data sets that comprised ten Asian languages. A preprocessed speech dataset was built for better performance compared to using the original collected speech data for multilingual speaker identification because getting and building high-quality speech datasets is important for speaker recognition research, especially for low-resourced tonal languages like Burmese. The preprocessed speech dataset was created by using two data preprocessing methods: increasing the speech intensity in SNRs to 10 dB and lowering the speech tempo by 0.2 times without affecting the pitch of the utterances. The performance of the dataset was compared using TDNN and GMM-UBM recognition models with MFCC features to evaluate the robustness of the preprocessed training dataset compared with the original training dataset. Moreover, the effectiveness of TDNN in processing various input contexts was also shown. An input temporal context of [t-8, t+8] was found to be the optimal network context because it splices the frames from wider temporal contexts in the upper layer of the network by incrementally processing the adjacent frames on narrow temporal contexts in the lower layer. Exploration of TDNN was first done before applying speaker identification on the Burmese language speech data. According to the outcomes, the performance of the models using the preprocessed speech dataset outperformed that using the original collected speech dataset in TDNN-based acoustic speaker modeling, with relative improvements of over 25% in terms of EER. Moreover, these preprocessing methods will be applied in other research, such as the study of speaker recognition through further optimization. This work has some limitations.

System performance is degraded in testing with wave files containing environmental background noise. There are still challenges related to robustness against noise. The speaker models can also have difficulty recognizing some speech because of the speaker's emotional condition. This work cannot assist in recognizing speech containing emotions. Getting better recognition also depends on the duration, and frequency range of the speech, the recording environment, the speaker's accent, and the physical conditions of the speakers because the stability of the identification process is not sufficient. Data preprocessing helps in getting not only high-quality data but also abundant data for low-resourced languages like Burmese.

Preprocessing speech datasets gives many benefits for future speaker identification research and can be further extended by more speech data. End-to-end learning will be pursued in speech recognition as a future work for further improving the performance of speaker identification. To reduce the background environmental noise, an additional voice activity detection algorithm will be applied in the future.

In this study, we found that the performance of speaker identification is influenced by two kinds of processing: raw audio signal processing in the speech recognition task and the end-to-end system. An end-to-end speaker identification system that directly uses raw audio signals (commonly used in speech recognition tasks) is proposed as a future work. This research will entail deliberate integration of a custom-designed embedded pre-processing layer into a TDNN, with the specific objective of diminishing the time complexity associated with network training and enhancing overall performance.

References

[1] Phyu, W.L.L. & Pa, W.P., *Building Speaker Identification Dataset for Noisy Conditions*, The 18th International Conference on Computer Applications, ICCA IEEE 2020, Yangon, Myanmar , pp. 182-188, 27-28 Feb 2020.

[2] Haizhou, L. & Ti, A.A., *ASEANN Language Speech Translation through U-STAR*, https://www.nict.go.jp/en/asean_ivo/lde9n2000000selb-att/lde9n2000000sesr.pdf, ASEAN IVO Forum 2019, Manila, Philippines, 21 Nov 2019.

[3] Ko, T., Peddinti, V., Povey, D. & Khudanpur, S., *Audio Augmentation for Speech Recognition,* Proceedings of INTERSPEECH2015, pp. 3586-3589, 2015.

[4] Beigi, H., *Speaker Recognition: Advancements and Challenges*, in New Trends and Developments in Biometrics, 28 Nov 2012. https://www.intechopen.com/books/3120.

[5] Li, A., Zheng C. & Li, X., *Glance and Gaze: A Collaborative Learning Framework for Single-Channel Speech Enhancement*, Applied Acousitcs, **187**, 108.499, 1 Feb 2022.

[6] Lemercier, J.M., Julius, R., Simon, W. & Timo, G., *StoRM: A Diffusion-Based Stochastic Regeneration Model for Speech Enhancement and Dereverberation,* IEEE ACM TASLP 2023, **31**, pp. 2724-2737, 2023.

[7] Naing, H.M.S., Hidayat, R., Hartanto, R. & Miyanaga, Y., *A Front-End Technique for Automatic Noisy Speech Recognition,* The 23th International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques, Oriental COCOSDA 2020, Yangon, Myanmar, Nov, 2020.

[8] Imam, S.A., Bansal, P. & Singh, V., *Review: Speaker Recognition Using Automated Systems,* AGU International Journal of Engineering & Technology, AGUIJET 2017*,* **5**, pp. 31-39, Jul-Dec, 2017.

[9] Mezghani, E., Charfeddine, M., Nicolas, H., Ben Amar, C., *Speaker Gender Identification Based on Majority Vote Classifiers*, Proceedings of SPIE: International Conference on Machine Vision (ICMV2016), Nice, France, 17 Mar 2017, pp. 47-51, 2017.

[10] Ali, Y.M., Emilia, N., Nor Fadzilah, M., Siti Zubaidah Md, S., Mohd Hanapiah, A. & Chee Chin, L., *Speech-based Gender Recognition Using Linear Prediction and Mel-Frequency Ceptral Coefficients*, IJEECS, vol. 28(2), pp. 753-761, 1 Nov 2022.

[11] Synder, D., Garcia-Romero, D. & Povey, D., *Time Delay Deep Neural Network-Based Universal Background Models for Speaker Recognition*, IEEE Workshop on Automatic Speech Recognition and Understanding, IEEE ASRU, Scottsdale, AZ, pp. 92-97, 1 December 2015.

[12] Waibel, A., Hanazawa, T, Hinton, G., Shikano, K. & Lang, K., *Phoneme Recognition Using Time Delay Neural Networks*, IEEE Transactions on Acoustics, Speech, and Signal Processing, **37**(3), pp. 328-339, Mar, 1989.

[13] Peddiniti, V., Povey, D. & Khudanpur, S., *A Time Delay Neural Network Architecture for Efficient Modeling of Long Temporal Contexts*, Proceedings of Interspeech, pp. 3214-3218, 2015.

[14] Park, H., Lee, D., Lim, M., Kang, Y., Oh, J. & Kim, J.H., *A Fast-Converged Acoustic Modeling for Korean Speech Recognition: A Preliminary Study on Time Delay Neural Network*, 11 July 2018, Retrieved from https://arxiv.org/abs/1807.05855.

[15] Ge, Z., Sharma, S.R. & Smith, M.J.T.,*PCA/LDA Approach for Text-Independent Speaker Recognition,* Proceedings of Society of Photo-Optical Instrumentation Engineers, SPIE 8401, Independent Component Analyses, Compressive Sampling, Wavelets, Neural Net, Biosystems, and Nanoengineering X, 23-27 April 2012.

[16] Chakroun, R. & Frikha, M., *A Deep Learning Approach for Text-Independent Speaker Recognition with Short Utterances,* Multimedia Tools and Applications, pp. 1-23, Mar, 2023.

[17] Rajan, P., Afanasyev, A., Hautamaki, V. & Kinnunen, T., *From Single to Multiple Enrollment I-Vectors: Practical PLDA Scoring Variants for Spaeaker Verification,* Journal of Digital Signal Processing, **31**, pp. 93-101, 1 Aug 2014.

[18] Ahilan, K., Vogt, R., Dean, D. & Sridharan, S., *PLDA Based Speaker Recognition on Shor Utterances,* The Speaker and Language Recognition Workshop, Odyssey 2012, Singapore, pp. 28-33, 25-28 June 2012.

[19] Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hennemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G. & Vesely, K., *The Kaldi Speech Recognition Toolkit*, IEEE 2011 Workshop on Automatic Speech Recognition and Understanding (ASRU2011), 2011.

[20] Cheng, J.M. & Wang, H.C., *A Method of Estimating the Equal Error Rate for Automatic Speaker Identification,* International Symposium on Chinese Spoken Language Processing (ISCSLP 2004), Hong Kong, pp. 285-288, Symposium conducted at The Chinese University of Hong Kong, 15-18 Dec 2004.

[21] Phyu, W.L.L. & Pa, W.P., *Text Independent Speaker Identificaiton for Myanmar Speech,* The 11th International Conference on Future Computer and Commnications, ICFCC 2019, Yangon, Myanmar , pp. 86-89, 27-28 Feb 2019.