ISBN 978-981-14-1455-8
Proceedings of 2019 The 11th International Conference on Future Computer and Communication
WCSE_2019_SPRING
Yangon, Myanmar, February 27-March 1, 2019, pp. 86-89

Text Independent Speaker Identification for Myanmar Speech

Win Lai Lai Phyu [1+] and Win Pa Pa [2]

[1, 2] Natural Language Processing Lab., University of Computer Studies, Yangon, Myanmar

Abstract. Nowadays, speech signal processing is one of the emerging application areas of digital processing. There are many research areas related to speech processing such as speaker recognition, speech recognition, and speech synthesis. Speaker identification is the task of analyzing the speakers' characteristics in speech to exactly identify individuals. The identification task performs better when there is enough background training data. Mel Frequency Cepstral Coefficient (MFCC), Perceptual Linear Prediction (PLP) and Filter Bank features are extracted as front-end processing. Constructing Universal Background Model (UBM) is the main component of i-vector system as it is essential for collecting statistics from speech utterances and for clustering the speaker models. This paper indicates that the impacts of unlimited speech data in speaker identification by using i-vector method with probabilistic linear discriminative analysis (PLDA) approach and the import role of speaker models in identification process.

Keywords: Speaker Identification (SI), Universal Background Model (UBM), MFCC, Filter Bank, PLP, i-vectors, Speaker Model, PLDA, Myanmar Speech

1. Introduction

Speaker Identification is an innate description of human-computer interaction (HCI) that automatically recognized the identity of persons from their voiceprints. The speech of living things especially for human involves numerous discriminative acoustic features that can be discerned who he/she is. The structural formation of vocal tract is unique for everyone. There are two types of speaker identification: text dependent and text independent. Although text dependent speaker identification needs to utter exactly the same utterance to determine who he/she is, the 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. A universal background model (UBM) is constructed using sound speech samples from all of the different speakers and the adaptable Maximum A Posteriori (MAP) is used for getting individual speaker models. In recent years, Gaussian Mixture Models (GMMs) have been used in speaker identification. But, GMMs have not tackled the channel distortions happened in the case of its guesses for the corresponding speaker are not exactly identical when different speech signals in varying recording conditions are used. To handle this problem, Joint Factor Analysis (JFA) modelled speaker variability and channel variability as two distinct subspaces. Nevertheless, the recognition process cannot well perform because the useful speaker related information may include in session variability subspace. To deal with these problems, i-vector method evolved from GMM super vectors is used. The i-vectors are ones that exist in low dimensional spaces that are smaller in size to reduce the recognizing time [1]. Probabilistic Linear Discriminant Analysis (PLDA), Support Vector Machine (SVM), and Cosine Distance Scoring (CDS) have been used as classifiers. In order to improve the performance, it has been used the batch-likelihood ratio between a target and test i-vector as the scoring method and appraised the accuracy with equal error rate (EER) and minimum decision cost function (DCF). Therefore, this paper is aimed the studying effects of human spoken speech data on i-vector with PLDA based speaker identification. This paper is comprised as follows. In section 2, the detail of speaker

[+] Corresponding author. Tel.: +95-9-259686621; fax: +95-1-610633
E-mail address: winlailaiphyu@ucsy.edu.mm

identification system is presented and experimental setup is described in section 3. Experimental results are expressed in section 4. Conclusion is addressed at section 5 and references are described in section 6.

2. Speaker Identification System

Speaker Identification is one of challenging topics in digital speech signal processing and the validating task of claimed identity by machine. It is intended to match the input sound signal with the sound signals that have already existed [2]. Three basic processes of speaker identification system are feature extraction, speaker modeling and PLDA based speaker identification in this paper. The architecture of speaker identification system overview is presented in figure1.

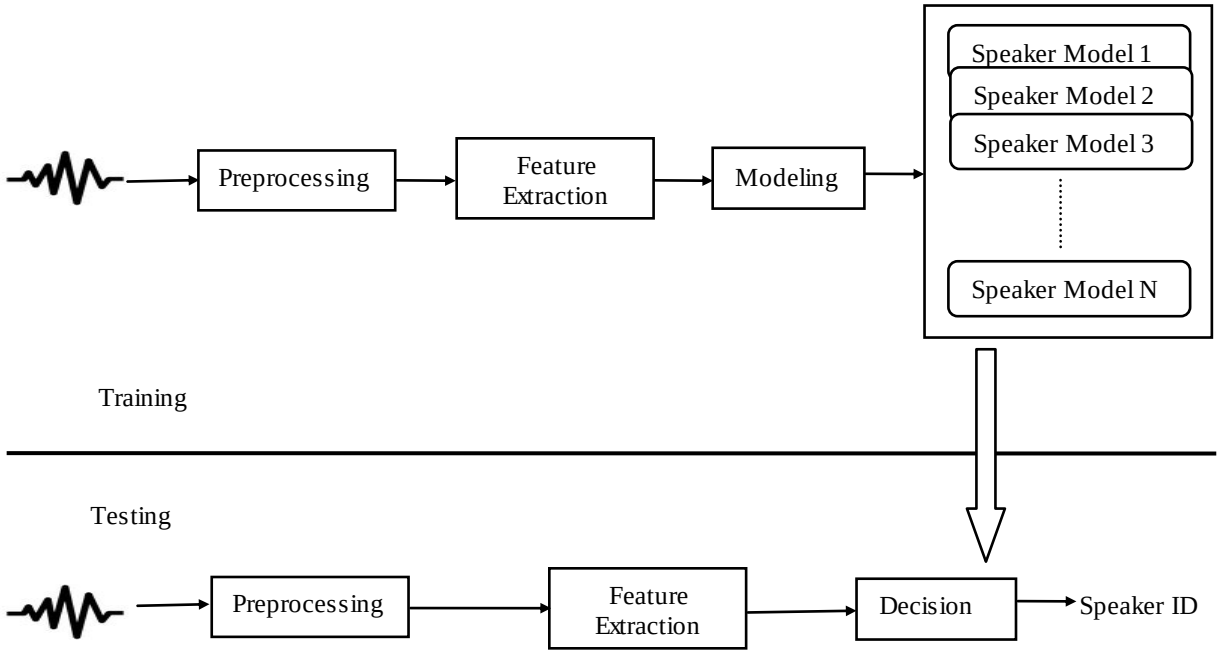

Fig. 1: Architecture of Speaker Identification System Overview

2.1. Feature Extraction

In this step, the raw spoken speech signals are converted into feature vectors as front end processing. Many distinct kinds of features are MFCC (Mel Frequency Cepstral Coefficient), PLP (Perceptual Linear Prediction), Filter Bank, LPC (Linear Predictive Code), RASTA PLP (Relative Spectral PLP), PNCC (Power Normalized Cepstral Coefficient). Depending on your choice, your system's recognition accuracy varies. Low level features are more powerful to build the best speaker identification system than high level features because it can easily extract the features. Although high level features involve more speaker related information, the extraction process is more complicated than low level features [3]. Therefore, MFCC, PLP and Filter Bank are extracted as features to build i-vector speaker models in this paper. All of the speech utterances were recorded at 16 kHz, 16 bits in mono PCM duration of ranging from 10 to 27 seconds each. This frequency rate quality affects the feature extraction and type of model because we use the frequency rate suitable for tone of Myanmar's spoken speech.

2.2. Speaker Modeling

Speaker modeling takes the features extracted from the speech samples in the feature extraction stage to develop the speaker models. This stage is the main part of speaker identification system because the models established in this stage are used for comparing in speaker classification/identification. Different types of modeling methods are GMM (Gaussian Mixture Model), DNN (Deep Neural Network), i-vector method and HMM (Hidden Markov Model). In this, i-vector method is used for speaker modeling.

2.3. PLDA Based Speaker Identification

This stage involves the identification of input speech signal. Different scoring methods for identification are SVM (Support Vector Machine), CDS (Cosine Distance Scoring), and PLDA (Probabilistic Linear Discriminant Analysis). In this paper, i-vector based speaker identification with PLDA was used for implementation. PLDA, a hierarchical generative probability model takes the correlations of feature vector in subspaces. The fixed length i-vectors extracted per utterance can be used as input to the pattern recognition algorithm. Therefore, we use PLDA applicable to fixed length input vectors. By using PLDA model, we can directly compute the log likelihood ratio for the test case corresponding to whether the two i-vectors are or aren't generated by the same speaker. To point out the amount of speech data required for the development of robust identification system, the effects of limited development data on PLDA based identification were investigated. To experiment the reduction of training and/or test data, the consequences of using short utterances for PLDA i-vector speaker identification were explored. As the i-vectors of long utterances vary, short utterances i-vectors also vary on their linguistic content of the utterances. This paper highlights the flexibility of PLDA that can use the highly correlated feature vectors. PLDA can be able to use multiple input feature frames with the use of subspace covariance modeling [4].

3. Experimental Setup

3.1. Data Preparation

In this experiment, Test Case 1 is about 18 minutes and 17 seconds including 111 utterances in known speech data and Test Case 2 is about 19 minutes and 17 seconds in unknown speech data with 111 utterances. The development data set size is about 54 minutes and 36 seconds of 321 utterances. The training data is totally 7 hours, 9 minutes and 46 seconds. There are 2517 utterances in the training data. There are totally 25 females and 12 males of speakers in all of cases. The experiments are done on the data from [5]. Table 1 shows how to prepare the data for speaker identification.

Table. 1: Data Preparation for PLDA Based Speaker Identification System

Data	Size	
	Number of Utterances	Time Duration
Training	2517	7 hours, 9 minutes, 46 seconds
Development	321	54 minutes, 36 seconds
Test Case 1 (Known Speech Data)	111	18minutes, 17seconds
Test Case 2 (Unknown Speech Data)	111	19 minutes, 17 seconds

3.2. Constructing i-Vectors-based Speaker Model

To build the speaker model, we use the Kaldi ASR open source toolkit [6]. Adopting the Kaldi scripts that have already exists, we used MFCC, PLP and Filter Bank features to build i-vectors based speaker models. I-vectors based system aims at modelling the total variability of the training data, compressing the information to a low dimensional vector and can be viewed as a front-end for modelling [7]. In our experiment, we initialized Universal Background Model (UBM) for speaker models with 200 components and 100 dimensions for extracting i-vectors. We have done experiments on 7 hours training data with 2517 utterances on totally 25 female and 12 male speakers.

3.3. Evaluation

False acceptance rate (FAR) and false rejection rate (FRR) are the two types of errors in speaker identification system. FAR is a type of error that allows the impostor speaker is falsely identified as the known speaker and FRR, an opposite of FAR, which incorrectly denied the actual speaker known by the system as impostor. Equal Error Rate (EER) is one where FAR equals to FRR, the compromise between FAR and FRR and also the point where FAR and FRR are optimal and minimal. EER is also known as the crossover rate or Crossover Error Rate (CER) [8]. Therefore, we employ the evaluation of EER to assess the performance of speaker models in this paper. The equations of FAR and FRR are described in equation (1) and (2). EER is one where FAR equals FRR. The lower the EER value, the higher the accuracy of speaker models.

$$FRR = \frac{TotalFalse\,Rejection}{TotalTrueAttempts} \quad (1)$$

$$FAR = \frac{TotalFalse\,Acceptance}{TotalFalse\,Attempts} \quad (2)$$

4. Experimental Results

In this section, we are going to present the evaluation results on building speaker models with PLDA approach. We built the speaker models on about 7 hours training data. We evaluated the equal error rate on training data by building the speaker models with three extracted features. Table 2 will present the performance of speaker models based on MFCC, PLP and Filter Bank with PLDA.

Table. 2: EER % based on MFCC, PLP and Filter Bank with PLDA

Types of Features	Equal Error Rate (%)		
	Training	Test Case 1	Test Case 2
Mel Frequency Cepstral Coefficient (MFCC)	3.738	35.51	35.83
Perceptual Linear Prediction (PLP)	2.804	29.6	31.15
Filter Bank	3.125	32.19	30.63

5. Conclusion

Speaker identification 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. It is the authentication of a person from characteristics of voices (voice biometrics) and is different from speech recognition. Speaker identification decides the identity of a person who is speaking although speech recognition recognizes what is being said. We have been presented how to construct the speaker models and how to implement the speaker identification on i-vectors based PLDA backend in this paper. We evaluated the identification process by constructing the speaker models as possible as the best from the size of training data together with using MFCC, PLP and Filter Bank features on building the i-vector based speaker models. Moreover, we analyzed the rate of change of EER on using various features. In this experiment, the EER of speaker model is mainly based on the amount of training data. A large amount of training data can construct the best speaker models. Moreover, the longer the duration of utterances, the more identify the speakers. As a result, the EER is decreased when the amount of training data is increased.

6. References

[1] Mohammed Senoussaoui, Patrick Kenny, Najian Dehak, Pierre Dumouchel, An i-vector Extractor Suitable for Speaker Recognition with both Microphone and Telephone Speech, *Odyssey 2010*, Brno, 28 June 2010.

[2] Ms. Arundhati S. Mehendale and Mrs. M. R. Dixit, Speaker Identification, *SIPIJ*, Vol.2, No.2, June 2011.

[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, 22-24 August 2017.

[4] Ahilan Kanagasundaram, Robbie Vogt, David Dean, Sridha Sridharan, PLDA based Speaker Recognition on Short Utterances, *Odyssey 2012*, Singapore , 25-28 June 2012.

[5] Aye Nyein Mon, Win Pa Pa, Ye Kyaw Thu, Building HMM-SGMM Continuous Autimatic Speech Recognition on Myanmar Web News, *International Conference on Computer Applications (ICCA 2017)*, Yangon, Myanmar, February 16-17, 2017.

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

[7] Ahilan Kanagasundaram, Robbie Vogt, David Dean, Sridha Sridharan, Michael Mason, i-vector Based Speaker Recognition on Short Utterances, *INTERSPEECH 2011*, Florence, Italy, 28-31 August 2011.

[8] Jyh-Min CHENG and Hsiao-Chuan WANG, A Method of Estimating the Equal Error Rate for Automatic Speaker Identfication, *International Symposium on Chinese Spoken Language Processing (ISCSLP 2004)*, Hong Kong, December 15-18, 2004, 0-7803-8678-7/04/$20.00 © 2004 IEEE.