Robust Language Classification on Short Utterances

Ms. Snehal V.Gite¹, Prof. J.V.Shinde²

¹Department of Computer Engineering, L.G.N. Sapkal COE, Nasik, gitesnehal7@gmail.com
²Department of Computer Engineering, L.G.N. Sapkal COE, Nasik, jv.shinde@rediffmail.com

Abstract—Automatic Language Identification is the process of classifying spoken words as belonging to one of a number of previously encountered languages. Achieving accurate performance with the shortest possible speech segment in a robust fashion is the main challenge in language identification. The proposed system works on robust language identification that involves rapid learning of new language identities and reduce the computational complexity. The proposed approach that transforms the spoken words to a low dimensional i-vector representation on which classification methods are applied. Universal background model (UBM) and i-vector extraction is used in proposed system in order to meet the challenges involved in rapidly making reliable decisions about the spoken language. By the deployment of a robust feature extraction scheme that capture the relevant language under acoustic conditions.

Keywords-Phonotactic Language, Automatic identification, i-vector, feature extraction, UBM.

I. INTRODUCTION

Natural Language Processing is a theoretically inspired range of computational techniques for analyzing and representing texts at one or more levels of linguistic analysis for achieving human like language processing for a range of applications. The goal of NLP is to accomplish human like language processing. The choice of the word, and processing is very consciously, and should not be recovered with understanding. The field of NLP was basically referred to as Natural Language Understanding (NLU) of AI, it is agreed today and the goal of NLP is true Natural Language Understanding, that goal has not yet been achieved. In NLP Language identification is the challenging task. Automatic language identification is the process of identifying the language i.e Spoken from a sample of speech by an unknown speaker. For speech recognition humans are the most accurate language identification systems. Within seconds of hearing speech, people are able to determine whether it is a language they know or not. If it is a language with they are not familiar, they often can make subjective judgments as its similarity to a language they know, e.g., “sounds like Farsi.” Language identification (LID) is the process of identifying the spoken language from the recording. Number of methods are introduces for automated language identification that are alternative to human audition and processing. When the language is unfamiliar to human the subjective decision are made, but it become less reliable. So this paper aims to specifically contribute to the domain of Robust Language identification in which reliable decisions about the spoken language need to be made quickly and accurate with as few seconds of speech as possible problem. For the task of text-independent speaker verification, likelihood ratio detection using Maximum A-Posteriori (MAP) adapted Gaussian Mixture Models (GMMs) [11] from a Universal Background model (UBM) has become the standard approach. While this approach provides very good performance, a continuing challenge for robust speaker verification is dealing with channel or session variability. JFA involves UBM to reduce the variability from non-language related effect. JAF [12] is mainly used for the speaker verification. So the aim of robust language identification on short utterances to identify language with accurate accuracy in short duration. For that we use extracted I-vector model and UBM for training and testing phase.
Paper formulates the problem on how to minimize the computational complexity of identified language. It also proposed LID for short utterances in for short duration to improve the accuracy and performance.

II. RELATED WORK

Language identification using phoneme recognition and phonotactic language modeling follows by n-gram language models and uses PRLM. It introduce gender-dependent acoustic model. This technique is used for improving speech recognition performance. But due to gender dependent accuracy is low so our system can improve accuracy for the gender-dependent.[4]

Language Identification (LID) based on language-dependent phone recognition uses a number of features and their combinations that are extracted by language-dependent recognizers were evaluation is based on the same database. Two methods are used that are [5]

- Forward and backward bigram based language models
- Context-dependent duration models.

Both the methods are used for language identification, backward bigram is used to capture backward phonetic constrains only.

2.1 GMM –Based Classification

SVM based speaker verification using GMM Model, Gaussian mixture models with universal backgrounds (UBMs) have become the standard method for speaker recognition. A speaker model is formed by MAP adaptation of the means of the UBM. A GMM supervector is constructed by the means of adopted mixture components. A recent research is that factor analysis of this GMM supervector is an effective method for variability compensation. We consider this GMM supervector in the of support vector machines. We construct a support vector machine using method called as GMM supervector.[6]

Acoustic, phonetic and discriminative approaches to automatic language identification that introduced 3 methods GMM, phone recognition and support vector machine classification but correct accuracy is not achieved.[7]

2.2 Total Variability Model

This method uses i-vector approach based on Joint Factor Analysis for speaker verification. JFA modeling based on speaker and channel factors consists of two distinct spaces: the speaker space factor defined by the eigenvoice matrix V and the channel space factor represented by the eigenchannel matrix U. Only single space is used instead of two, which refer to as the “total variability space. The total variability matrix contains the eigenvectors that have the largest eigenvalues of the total variability covariance matrix. Given an utterance, the new speaker- and channel-dependent GMM supervector defined as follows:

\[ M = m + Tw \] 2.1

Joint factor analysis versus Eigen channels in speaker recognition this technique presented two approaches to the problem of session variability in Gaussian mixture model (GMM)-based speaker verification, Eigenchannels, and joint factor analysis that factor analysis was far more effective than Eigenchannel modeling.[8] In proposed system we use JFA methods for accuracy in speaker verification task.

Front-end factor analysis for speaker verification this technique presented a new speaker verification system where further analysis is used to define new low-dimensional space that models both speaker and channel variability. [9]
I-vectors in the context of phonetically-constrained short utterances for speaker verification future scope of this technique is to exploring the impact of phonetic information on i-vector normalization by considering the correlation between the existing speaker discrimination scoring and difference between utterances phonetic distance for very short duration[10]

III. PROPOSED SYSTEM

In above mentioned techniques identifying language with the help of varies methods. Gaussian mixture models with universal backgrounds (UBMs) have become the standard method for speaker recognition. In proposed system, obtaining a high accuracy at a low computational cost is essential for making rapid and reliable decisions about the spoken language on utterances with short duration. For that first we introduce our system module, together with the proposal of an acoustic feature set capturing multiple unequal speech characteristics. Next, we restate the modification to the i-vector modeling [8] that was proposed in previous section to improve LID performance in terms of computational load. This simplified i-vector[13][14], system is further extended to the framework of UBM-fused total variability modeling. It will be shown that significant improvements in accuracy, while maintaining the system’s complexity, are achieved when the i-vector space is estimated in this framework and by training on utterances with long duration.

So proposed system mainly works in 4 steps:
1. Audio recording and preprocessing
2. Voice Activity Detection
3. Feature Extraction
4. Normalization

![Figure 1. Proposed System](image)

In proposed system first input audio is taken then preprocessing is done i.e. speech enhancement, audio should capture the acoustic properties. In language identification as the change in background noise can change in acoustic condition. Voice Activity Detection (VAD) to prevent non-speech audio segments from interfering with the classification decision, a speech enhancement method to compensate for noise distortions and a robust feature extraction module. Normalization step is done to further reduce the sensitivity of the features to the acoustic variability. After feature extraction we use simplified i-vector model for the LID on both training and testing phase. Achieving a high accuracy and performance on a small amount of speech data is the main aim of this paper. In GMM-based Language identification, the statistical probabilities derived from the acoustical representation
of the spoken words will be accumulated over time and propagated until the final classification stage. Although their performance also increases when more statistics can be accumulated, these systems tend to be more robust on short utterances as they do not based on rule-based approaches applied on phonetic transcription. Various attempts of accurate language identification on short-duration sentences using the i-vector frame work have been proposed. To reduce the computational complexity RLID adopts the simplified i-vector framework and UBM –fused total variability model. These simplified work of i-vector system slightly reduces the complexity of the conventional i-vector baseline.

3.1 Simplified i-vector Model

For the complexity reduction in proposed system we use simplified i-vector model. This simplified and supervised i-vector modeling is applied in the task of robust and efficient speaker verification. In that first we concatenating the mean supervector and the i-vector factor loading matrix with the label vector and the linear classifier matrix, the i-vectors are then extended to label-regularized supervised i-vectors. These supervised i-vectors are optimized to reconstruct the mean supervectors as well as minimize the mean squared error between the original and the reconstructed label vectors, such that they become more discriminative. Second, factor analysis (FA) can be performed on the pre-normalized centered GMM first order statistics supervector is used to ensure that the Gaussian statistics sub-vector of each Gaussian component is treated equally in the FA, which reduces the computational cost significantly.

IV. RESULT

Partial result of our module is extracted features from the input audio. Features that are extracted from the input are the acoustic feature of the spoken words. so the first step of the system is to take a input audio and extract the feature as shown in fig.2

![Figure 2. Feature Extraction](image_url)

CONCLUSION

This paper presents our recent technological advances in the domain of robust language identification on short utterances. The proposed RLID system extracts relevant acoustic features of the spoken languages and deploys an i-vector based framework. Proposed system includes a novel...
feature representation set, was proposed that lowers the LID error rates when compared to standard features. To make fast and almost instant decision about the spoken language, a simplified i-vector modeling framework is used within the system to increase the efficiency of the i-vector extraction process. Techniques for accurate i-vector space modeling are introduced to further improve the identification performance on short duration speech utterances.

ACKNOWLEDGMENT

With all respect and gratitude, I would like to thank all people who have helped us directly or indirectly for the paper presentation. I would like to thank my guide, Prof. J.V. Shinde, for her guidance and support. I will forever remain grateful for the constant support and guidance extended by guide, in making this paper. Through our many discussions, she helped me to form and solidify ideas. The invaluable discussions I had with her, the penetrating questions she has put to me and the constant motivation, has all led to the development of this paper.

REFERENCES
