A Review on Robust Language Identification

Snehal V. Gite¹, Prof. J.V. Shinde²

gitesnehal7@gmail.com¹, jv.shinde@rediffmail.com²

M. E. Student, Dept. of Computer Engg. Late G.N. Sapkal COE, Nashik, Maharashtra, India¹
Assistant Professor, Dept. of Computer Engg. Late G.N. Sapkal COE, Nashik, Maharashtra, India²

Abstract

Automatic language identification of audio data has become an important pre-processing step for speech/speaker recognition. In order to identify the language on short utterances is critical challenge for achieving accurate performance. In this paper, we review language identification methods. First we discuss Phonotactic systems use phone recognition and language modelling (PRLM). The second category of LID techniques attempts to classify languages by using Gaussian mixture models (GMMs) to capture the acoustic properties of speech. For the accuracy of language identification on short utterances the proposed system works on robust language classification that transforms the spoken words to a low dimensional i-vector representation on which language classification methods are applied. Language classification model use the approach of universal background model for the better performance.

Keywords: Automatic language identification, PRLM, GMM, i-vector

1. Introduction

Over the last years, number of methods have been proposed and investigated for the use in biometric recognition. Such as most popular fields like fingerprint recognition and face recognition, speaker recognition, for the purpose of automatically recognize the speaker identities through the voice of persons, has made its debut in several commercial applications. It can be used in a wide range of applications including access control, forensic evidence providing, and user authentication. Current speaker recognition methods provide a satisfying performance when data is sufficient. However, in some situations, only a short utterance such as one or two words is available to recognize the speaker, and in other situations short utterances can provide a better user experience.

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. If it is a language with which they are not familiar, they often can make subjective judgments as its similarity to a language they know, e.g., “sounds like Spanish.”

Language identification (LID) is the task of identifying the spoken language from speech recordings. 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 Classification in which reliable decisions about the spoken language need to be made quickly with as few seconds of speech as possible.

Research in the field of automated LID system development has along two main directions related to the language information cues they are as follows:

1. Phonotactic systems use phone recognition and language modelling (PRLM).
2. The second category of LID methods attempts to classify languages by using Gaussian mixture models (GMMs) to capture the acoustic properties of speech.

In PRLM phone recognizer first converts speech signals into a sequence of phone symbols or tokens. The tokenization of speech is followed by training a language model to extract language-specific information. From the token string specific phonotactic information is derived. GMM-based language identification and significant progress in LID performance has been made by employing super vector modeling and the introduction of Joint Factor Analysis (JFA). JFA involves the adaptation of a large general speech GMM, named the Universal Background Model (UBM), is used to reduce the variability from non-language related effects, such as differences in recording channels, session aspects and background noises in the data.

In this paper, we review the issue of Language identification -duplication check.
2. Literature survey

I) Phone Recognition and Language Modelling:
M.A.Zissman[4] proposed language identification using phoneme recognition and phonotactic language modeling follows by n-gram language models and uses PRLM. It also introduce gender-dependent acoustic model. The use of gender-dependent acoustic models is a well-known technique for improving speech recognition performance. For LID, we hoped that gender-dependent phoneme recognizers would produce a more reliable tokenization of the input speech relative to their gender-independent counterparts; therefore, n-gram analysis might prove more effective. Future scope of this technique for availability of larger, standardized multi-language speech corpora. Phonotactic Language model is,

\[ P(x|l) = WMP(x|\lambda l) + WFP(x|\lambda l) + WGIP(x|\lambda l) \]

Yonghong Yan [5] An approach to Language Identification (LID) based on language-dependent phone recognition is presented in this paper. A number of features and their combinations extracted by language-dependent recognizers were evaluated based on the same database. Two main information sources for LID are as follows:
(1) Forward and backward bigram based language models
(2) Context-dependent duration models.
An LID system is developed by Hidden Markov Models and neural network. The system was trained and evaluated using the OGI-TS database. Three model are used in this paper

* Acoustic model: The acoustic model is used for phone recognition as well as for acoustic likelihood calculation for LID. A three-state left-to-right HMM is used for each phone in each language. Four Gaussian mixtures are used to model the probability density for each state in a model.

* Language Model: This language model is based on the interpolated n-gram language model. It exploits left context information, thus only the forward information is captured. Although a trigram-based language model can capture both the right and left context information, a larger database is needed in order to get a well estimated model.

* Duration Model: Duration modeling has been used in to capture a certain class of prosodic information in the different languages.

II) GMM Model
W. M. Campbell [6] introduce the SVM based speaker verification using GMM Model, Gaussian mixture models with universal backgrounds (UBMs) have become the standard method for speaker recognition. Typically, a speaker model is constructed by MAP adaptation of the means of the UBM. A GMM supervector is constructed by stacking the means of the adapted mixture components. Recent discovery is that latent 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 kernel using the GMM supervector.

Fig. 1 GMM Supervector Concept

E. Singer, P. A. Torres-Carrasquillo [7] proposed 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.


Patrick Kenny, Gilles Boulianne, Pierre Ouellet, and Pierre Dumouchel [9] proposed Joint factor analysis versus eigenchannels 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.

Najim Dehak, Patrick J. Kenny, Réda Dehak, Pierre Dumouchel, and Pierre [10] Ouellet 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.

D. Martinez, O. Plchot, L. Burget, O. Glembek, and P. Matjeka[11]- proposed Language recognition in iVectors space The concept of so called iVectors, where each utterance is represented by fixed-length low-dimensional feature vector, novel approach for language recognition that model provides excellent performance over all conditions future scope is try to obtain iVectors from the utterances and the corresponding sufficient statistics in a more direct way.

A. Kanagasundaram, R. Vogt, D. B. Dean, S. Sridhar an, and M. W. Mason[12]-proposed I-vector based speaker recognition on short utterances This paper has presented a study investigating how the current selection of factor analysis techniques perform when utterance lengths are significantly reduced. Problems of very short utterance with factor analysis approaches will be investigated in future.

A. Larcher, P. Bousquet, K. A. Lee, D. Matrouf, H. Li, and J.-F. Bonastre [13]- proposed 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.

3. 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 [11] that was proposed in previous section to improve LID performance in terms of computational load. This simplified i-vector 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.

Block Diagram Of Proposed System:

Figure 2: Proposed architecture

first feature that are extracted from the audio data should capture the acoustic properties that are relevant to discriminate between the languages. Furthermore, deployment of LID in real-life scenario requires the features to be relatively invariant for a wide range of adverse acoustic conditions, such as variations in the background noise environment and changes in the audio transmission channels or recording devices. Next 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 followed by a normalization step to further reduce the sensitivity of the features to the acoustic variability

4. Conclusion

The focus of this survey paper is to study Language classification technique with different methods. The existing methodologies help us to understand the GMM technique. The previous work and methodologies motivate us to develop our proposed system which will be the robust language Classification technique. We include GMM model with modification and Extraction methods in proposed system.

6. Acknowledgement

I am thankful to Prof J .V. Shinde Assistant professor in the Department of Computer Engineering in Late G.N. Sapkal College of Engineering, Nasik. For providing constant guidance and encouragement for this research work.
7. References


