IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 2 Issue 3, March 2015.
www.ijiset.com
ISSN 2348 – 7968

A Survey On Language Identification System

Sunil Kumar Gupta, Om Prakash Singh, Prashanta Chandra Pradhan
Department of Electronics and communication Engineering, Sikkim ManipalUniversity,SikkimManipal Institute of Technology, Majitar, East Sikkim, PIN 73713,Sikkim , India

Abstract

In the recent years researchers are taking great interest in automatic language identification systems to facilitate people who are unknown to the language coming to their way. This paper will give basic idea of how to deal with spoken language identification (LID) systems and how we can obtain good efficiency with better identification rate(I.R.). Spoken language is not linear function, it's a logarithmic function so to make successful language identification systems we need to look for acoustic ,phonotactic , prosodic, lexical and syntax level of speech information. The speech information will tell us what state-of- art to use for better efficiency.

Keywords: *Language Identification System(LID), Identification Rate (I.R.), Acoustic, syntax.*

1. Introduction

Language Identification system is basically a system that helps you accurately recognize an unknown speakers or the unknown utterances of a language. Today the accuracy to recognize a language is highest for homo sapiensapien. From approximately 6900[1] living languages we are intelligent enough to recognize the language, if not the exact language we are able to tell about the accent that this is from a particular region or place etc. we say that the humans are most accurate identifier because for a short period of training we give the fastest result. Till date no electronic or electrical system have been developed to recognize1/4th of what human can recognize of unknown language or utterances. Today we need LID system as it has got many applications to the places where dealing with many languages take place . We cannot afford to put so much cost in training huge number of people for the same few languages and then deploy them to various centre or places. Advantage of developing best LID is to reduce cost in deploying the system, but the big challenge is developing of accurate LID with minimum dictionary or database size. There are different approaches to build dictionary or database for the language identification system.

2. Types of Speech Information

For automatic language identification system we need to extract features and that features are very important to distinguish between different languages so that the system can identify the different unknown speech information. The features goes from low level to high level and different level have got different techniques to extract the features.

- **Acoustic**

ISSN 2348 – 7968

This speech information is the simplest ,which is obtained by amplitude and frequency components of the speech wave[2]. Acoustic information is basic form of information because it can be obtained during the feature extraction or say speech parameterization process by taking the raw data directly.

- **Phonotactic**

In this, speech information is done through collecting the sound patterns of a particular language which is not involved or found in any other languages. When we compare to acoustic model ,then we can say that this carries more information than it.

- **Prosodic**

There are some phonemes which are shared across different other languages but their duration characteristics will depend on the accent of the language or say its speaking time . In this ,speech information is done through the accent of the speech like the tone, stress ,pitch ,duration and even the rhythm.

- **Lexical**

In this ,speech information we get the internal structure of words[3]. The word roots are usually different for different languages and so, this can also be used for language identification system, here the task are performed at word level.

- **Syntax**

In the language studied throughout the world, syntax is the study of the principles and rules that govern the way that words in a sentence come together [4].

3. Functionality of Language Identification System(LID)

The LID system involves two phases firstly, the Training phase and secondly, the testing phase.

ISSN 2348 – 7968

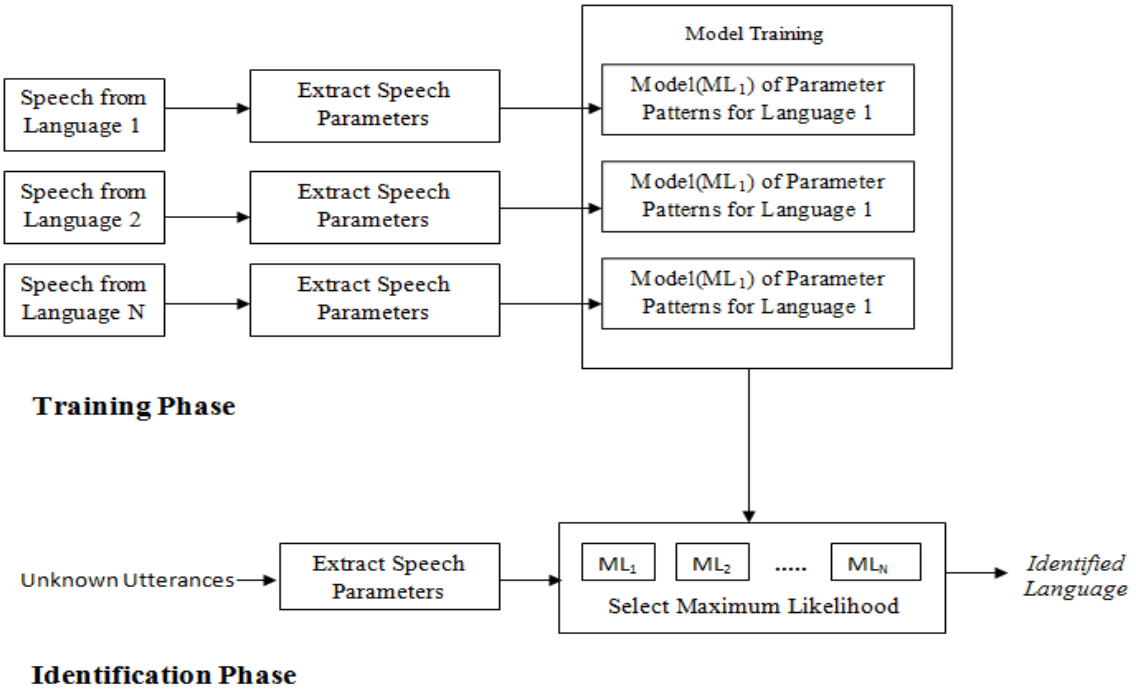

Fig.1.: Block Diagram of LID

The LID system involves different back-end and front-end , various back-end identifiers are used, depending on the front-end features. Recent works involve these different identification systems like Artificial Neural Network , Gaussian Mixture Model , Hidden Markov Model , Support Vector Machine , etc., have been widely used in the LID systems [5-8]. The most recent State-of-art is Identity vector or commonly known as I-vectors [9]. By using different state of art techniques we built our backend model considering the speech information. Now our model for each language is completed and now we have our training model. For the testing phase we take the unknown speech utterances and through open software we extract the speech parameters and then we compare the extracted parameters to the backend model or the training model. At this point we use select maximum Likelihood tool or better knowas probability to reduce the processing time. Now , finally we get the language identified in terms of percentage.

4. Experimental Setup

In language identification we first take the samples from different speakers for different languages, then, using some open source software we take out the features and then do the parameterization for dictionary that we will be making for our back-end of the system. Use of Dynamic Speaker is best because of its low sensitivity. The utterances to be stored in the external hard disk as the data will be voluminous. While training the data we need high configuration personal computer.

ISSN 2348 – 7968

The Identification Rate will match the language to each other in terms of percentage and display the result in matrix tabular form.

5. Conclusions

Language identification system provides a standard for measuring the identification rate(IR) of language in the database. Use of maximum likelihood probability reduces the processing or overall time of the system.The different level of LID features will give different efficiency. Suppose, if we consider only 5 languages on acoustic information with 40 minutes of each speech then its efficiency will be lesser than for same 5 languages with 2 hours of each speech.

References

[1.] R. Gordon and B. Grimes, *Ethnologue: Languages of the World.* Dallas:SIL International, 2005, vol. 1272.

[2.] J. Laver, *Principles of Phonetics*. Cambridge, U.K.: Cambridge Univ. Press, 1994.

[3.] L. Bauer, *Introducing Linguistic Morphology*.Georgetown Univ. Press, 2003.

[4.] A. Carnie, *Syntax: A Generative Introduction*, 2nd ed. New York: Wiley-Blackwell, 2006.

[5] R. A. Cole, J. W. T. Inouye, Y. K. Muthusamy, and M. Gopalakrishnan, "Language identification with neural networks: A feasibility study," in Proc. IEEE Pacific Rim Conf. Communications, Computers and Signal Processing,1989, pp. 525–529.

[6.] E. Wong and S. Sridharan, "Methods to improve Gaussian mixture model based language identification system," in Proc. Int. Conf. Spoken Language Processing (ICSLP-2002), 2002, pp. 93–96.

ISSN 2348 – 7968

[7.] S. Nakagawa, Y. Ueda, and T. Seino, “Speaker-independent, textindependent language identification by HMM,” in Proc. Int. Conf. Spoken Language Processing (ICSLP-1992), 1992, pp. 1011–1014.

[8.] Z. Lu-Feng, S. Man-hung, Y. Xi, and H. Gish, “Discriminatively trained language models using support vector machines for language identification,” in Proc. Speaker and Language Recognition Workshop, 2006. IEEE Odyssey 2006, pp. 1–6.

[9.] O.P.Singh , B.C. Haris, and R. Sinha ,” Sparse Representation based Language Identification using Prosodic Features for Indian Languages” in Proc. Annual IEEE India Conference (INDICON-2013).