Speech Communication 100 (2018) 30–40

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

Using language cluster models in hierarchical language identification

Saad Irtza[a,b,*], Vidhyasaharan Sethu[a], Eliathamby Ambikairajah[a,b], Haizhou Li[c]

[a] School of Electrical Engineering and Telecommunications, UNSW, Australia
[b] DATA61, CSIRO, Australia, Sydney NSW 2015, Australia
[c] Department of Electrical and Computer Engineering, National University of Singapore, Singapore

ARTICLE INFO

Keywords:
Language identification
Decision strategy
Language cluster specific features

ABSTRACT

Hierarchical language identification systems can be employed to take advantage of similarities and disparities between languages to organize them into clusters and decompose the language identification problem into a tree of potentially simpler sub-problems of language group identifications. In this paper, a novel approach is proposed to incorporate knowledge of the language clusters into the front-ends of the classification systems employed in each node of a hierarchical language identification system. This approach investigates the use of feature representations tuned to the particular language cluster identification sub-problem at each node. In addition, we explore a novel decision strategy that incorporates information about language cluster model memberships into the front-ends at each node. Experimental results included in this paper demonstrate that both approaches lead to improved language identification performance of the overall hierarchical system on the NIST LRE 2015 database.

1. Introduction

The aim of automatic language recognition is to identify the language being spoken from a group of possible languages (Ambikairajah et al., 2011; Li et al., 2013). It is an enabling technology for a wide range of multilingual speech processing applications, such as call-routing by language, multilingual speech recognition, spoken language translation and language profiling of speech archives. Humans and machines use a variety of information found in speech to distinguish one language from another. The most effective information for automatic Language Identification (LID) includes acoustic and phonotactic features (Ambikairajah et al., 2011; Dehak et al., 2011; Soufifar et al., 2011). Using phonotactic information, we tokenize a speech signal into acoustic-phonetic units, from which we derive statistics, such as phone log-likelihood ratios, to make a decision (Díez et al., 2012, 2013). To make use of acoustic information, we represent speech signals as sequences of short-term spectral and/or prosodic feature vectors. Longer term information is then typically captured through the use of supervector representations of utterances or through a total variability factor analysis in the i-vector framework. More recently, Deep Neural Networks (DNNs) have been employed either in the front-end using bottleneck features (Richardson et al., 2015) or in end-to-end architectures for language identification (Ma et al., 2016; Geng et al., 2016).

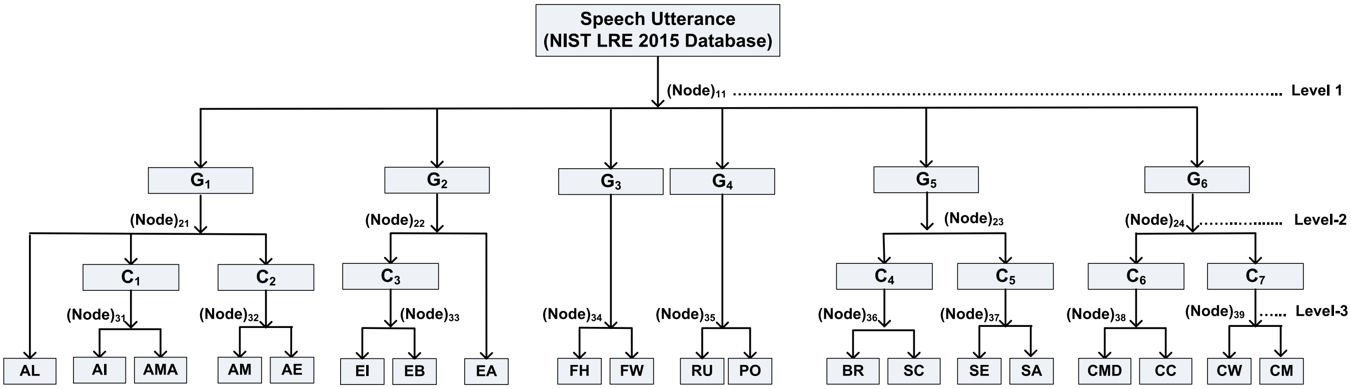
In the literature, a LID task is often formulated as a hypothesis test where we decide if a language identity claim of a speech sample is true or false. As such, LID is also called language verification/recognition, especially when the hypothesis test involves languages that are outside of the training set. The National Institute of Standards and Technology (NIST) Language Recognition Evaluation (LRE) campaigns provide a common protocol for reporting the system performance, which we will follow in this paper.

In most LID studies, all language hypotheses are treated independently without exploring any information about similarities between languages. This is called a single-level approach. We have also seen successful use of inter-language information in LID tasks, where similar languages are grouped to improve language models (Li et al., 2013; Bekker et al., 2016; Bing et al., 2012). In this approach, any available data from additional languages that are similar to but different from the target languages are effectively combined with training data (Li et al., 2013). This approach has also been successfully employed in pairwise LID tasks (Bing et al., 2012).

The hierarchical LID framework has been previously proposed as an alternative approach that makes use of language similarity information that identifies languages through a multi-level decision. In this way, we solve a language identification problem through a top-down hierarchy of smaller sub-problems, with initial high level decisions pertaining to identification of language groups followed by identification of specific languages (Irtza et al., 2016a,b; Jothilakshmi et al., 2012; Yin et al., 2008, 2007). The framework shown in Fig. 1 uses a tree structure where

[*] Corresponding author at: School of Electrical Engineering and Telecommunications, UNSW, Kensington 2033, Australia.
E-mail address: s.irtza@unsw.edu.au (S. Irtza).

https://doi.org/10.1016/j.specom.2018.04.004
Received 21 April 2017; Received in revised form 17 January 2018; Accepted 9 April 2018
Available online 22 April 2018



* 1-AL=Arabic Levantine, 2-AI=Arabic Iraqi, 3-AMA=Arabic Maghrebi , 4-AM=Arabic Modern, 5-AE=Arabic Egyptian, 6-EI=English Indian, 7-EB=English British, 8-EA=English American, 9-FH=French Haitian, 10-FW=French West African, 11-RU=Russian, 12-PO=Polish, 13-BR=Brazilian, 14-SC=Spanish Caribbean, 15-SE=Spanish European, 16-SA=Spanish American, 17-CMD=Chinese Ming Dong, 18-CC=Chinese Cantonese, 19-CW=Chinese Wu, 20-CM=Chinese Mandarin

Fig. 1. A hierarchical LID structure using NIST LRE 2015 dataset as an example (Irtza et al., 2016b).

the root node distinguishes between broad language clusters/groups and the nodes in each subsequent level model the differences between language sub-clusters, which are sub-sets of the language cluster corresponding to the branch leading to that node. The nodes at the final level are the only ones that model individual languages. For instance, in the hierarchical LID structure shown in Fig. 1, Node$_{11}$ models the differences between broad language groups (denoted G$_1$ to G$_6$) while the subsequent node, Node$_{21}$, models the differences between sub-clusters of languages within G$_1$ (denoted C$_1$ and C$_2$). The final level node, Node$_{31}$, models the differences between languages AI (Arabic Iraqi) and AMA (Arabic Maghrebi) which together constitute language cluster C$_3$. It should be noted that the number of nodes in each path from the root to the final language (leaf) need not be equal. In the structure depicted in Fig. 1, language groups G$_3$ and G$_4$ are comprised of only two languages each and consequently only one more node is required to distinguish between these language pairs (Node$_{34}$ for G$_3$ and Node$_{35}$ for G$_4$). Finally, nodes can also model differences between language clusters and individual languages, e.g. Node$_{21}$ models the three-way differences between language AL (Leventine Arabic) and language sub-clusters C$_1$ and C$_2$.

The specific structure of the hierarchical framework could be constructed according to a range of different speech cues or prior linguistic knowledge. For example, automatic language clustering algorithms can be used to form the hierarchical structure using similarity measures based on acoustic models (Yin et al., 2008). Alternatively, language families (Lewis et al., 2009) such as those identified from Ethnologue can also form the basis for a hierarchical structure (Irtza et al., 2016b).

While the hierarchical LID system makes multi-level decisions, all nodes of current systems make use of identical factor analysis models to represent speech utterances in these sub-problems. The gains obtained so far have come from training the back-end to only model the differences between small sets of languages and/or language clusters, and from choosing the most suitable frame-level features for each sub-problem from a set of established features. This paper proposes a novel approach to incorporating knowledge about language clusters into the utterance-level representations employed in each node. The underlying motivation is that the representations that best capture the differences between the branches arising from one node may not be the same as those for another node. For instance, the most suitable factor analysis model (i-vector model) to distinguish between EI (Indian English) and EB (British English) in Node$_{33}$ is likely to be different to the best model for RU (Russian) and PO (Polish) in Node$_{35}$.

In this approach, the utterance-level representation of features in the front-ends of each node is designed to capture only the differences between the languages that fall within the language cluster corresponding to the branch leading to that node. Three ways to accomplish this are explored, two within the well-established i-vector framework and one using DNN-based bottleneck features. In addition to the proposed approach, we also explore a novel decision strategy whereby knowledge of all language clusters corresponding to the hierarchical structure is incorporated in the form of scores produced by all nodes in the tree for a given input speech utterance. We will present the working framework in Section 2, the novel approach in Section 3, and the proposed decision strategy in Section 4.

2. Hierarchical language identification

Given an input speech utterance s, the LID task can be expressed as:

$$\mathscr{L}(s) = \arg\max_{\ell\in\Lambda} P(\ell|s) \tag{1}$$

where $\mathscr{L}(s)$ denotes the language identified by the system given the input speech utterance s, and ℓ denotes elements of the set of all target languages, Λ.

In a single-level language identification system, the posterior probabilities of each language given the input speech, $\{P(\ell|s); \forall\ell\in\Lambda\}$, are all estimated by a single back-end that models the set of all target languages. However, within the hierarchical framework, the language posteriors are computed as a product of conditional probabilities of language clusters given a larger cluster based on the path from the root node of the hierarchical tree to the leaf nodes representing the target languages as:

$$P(\ell|s) = P(\ell|c_N, s)\left(\prod_{i=1}^{N} P(c_i|c_{i-1}, s)\right) \tag{2}$$

where each term of the product $P(c_i|c_{i-1}, s)$ is computed by a different node on the path from the root node to the leaf node representing the target language ℓ; c_i denotes the cluster of languages corresponding to a branch leading out of that node and c_{i-1} denotes the cluster of languages corresponding to the branch leading into the node, with $c_i \subset c_{i-1}$ and $c_0 = \Lambda$; and $P(\ell|c_N)$ is computed by the last node on the path.

For instance, in the structure shown in Fig. 1, the posterior probability of ‘English British’ (EB) given an input speech utterance s is computed as follows:

$$P(EB|s) = P(EB|C_3, s)P(C_3|G2, s)P(G2|\Lambda, s) \tag{3}$$

where, $P(EB|C_3, s)$ is estimated by Node$_{33}$, $P(C_3|G2, s)$ is estimated by Node$_{22}$, $P(G2|\Lambda, s)$ is estimated by Node$_{11}$, and Λ denotes the set of all target languages at the third level in Fig. 1 i.e. $\Lambda = \{\text{AL, AI, AMA, ... CW, CM}\}$.

If we consider the world's spoken languages as a finite set, the hierarchical LID architecture represents a system approach towards the LID problem where inter-language knowledge is applied to differentiate one language from another. It emulates a human perceptual process in