Neural Networks 64 (2015) 49–58

Contents lists available at ScienceDirect

Neural Networks

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

2015 Special Issue

Frame-by-frame language identification in short utterances using deep neural networks

Javier Gonzalez-Dominguez [a,b,*], Ignacio Lopez-Moreno [a], Pedro J. Moreno [a], Joaquin Gonzalez-Rodriguez [b]

[a] Google Inc., NY, USA

[b] ATVS-Biometric Recognition Group, Universidad Autonoma de Madrid, Madrid, Spain

ARTICLE INFO

Article history:
Available online 3 September 2014

Keywords:
DNNs
Real-time LID
i-vectors

ABSTRACT

This work addresses the use of deep neural networks (DNNs) in automatic language identification (LID) focused on short test utterances. Motivated by their recent success in acoustic modelling for speech recognition, we adapt DNNs to the problem of identifying the language in a given utterance from the short-term acoustic features. We show how DNNs are particularly suitable to perform LID in real-time applications, due to their capacity to emit a language identification posterior at each new frame of the test utterance. We then analyse different aspects of the system, such as the amount of required training data, the number of hidden layers, the relevance of contextual information and the effect of the test utterance duration. Finally, we propose several methods to combine frame-by-frame posteriors. Experiments are conducted on two different datasets: the public NIST Language Recognition Evaluation 2009 (3 s task) and a much larger corpus (of 5 million utterances) known as Google 5M LID, obtained from different Google Services. Reported results show relative improvements of DNNs versus the *i*-vector system of 40% in LRE09 3 second task and 76% in Google 5M LID.

1. Introduction

Automatic language identification (LID) refers to the process of automatically determining the language in a given speech sample (Muthusamy, Barnard, & Cole, 1994). The need for reliable LID is continuously growing due to several factors. Among them, the technological trend towards increased human interaction using hands-free, voice-operated devices and the need to facilitate the coexistence of a multiplicity of different languages in an increasingly globalized world.

In general, language discriminant information is spread across different structures or levels of the speech signal, ranging from low-level, short-term acoustic and spectral features to high-level, long-term features (i.e. phonotactic, prosodic). However, even though several high-level approaches are used as meaningful complementary sources of information (Ferrer, Scheffer, & Shriberg, 2010; Martinez, Lleida, Ortega, & Miguel, 2013; Zissman, 1996), most LID systems still include or rely on acoustic modelling (Gonzalez-Dominguez et al., 2010; Torres-Carrasquillo et al., 2010), mainly due to their better scalability and computational efficiency.

Indeed, computational cost plays an important role, as LID systems commonly act as a pre-processing stage for either machine systems (i.e. multilingual speech processing systems) or human listeners (i.e. call routing to a proper human operator) (Li, Ma, & Lee, 2013). Therefore, accurate and efficient behaviour in real-time applications is often essential, for example, when used for emergency call routing, where the response time of a fluent native operator is critical (Ambikairajah, Li, Wang, Yin, & Sethu, 2011; Muthusamy et al., 1994). In such situations, the use of high-level speech information may be prohibitive, as it often requires running one speech/phonetic recognizer per target language (Zissman & Berkling, 2001). Lightweight LID systems are especially necessary in cases where the application requires an implementation embedded in a portable device.

Driven by recent developments in speaker verification, the current state of the art in acoustic LID systems involves using *i*-vector front-end features followed by diverse classification mechanisms that compensate speaker and session variabilities (Brummer et al., 2012; Li et al., 2013; Sturim et al., 2011). The *i*-vector is a compact representation (typically from 400 to 600 dimensions) of a

* Corresponding author at: ATVS-Biometric Recognition Group, Universidad Autonoma de Madrid, Madrid, Spain. Tel.: +34 914977558.
E-mail address: javier.gonzalez@uam.es (J. Gonzalez-Dominguez).

http://dx.doi.org/10.1016/j.neunet.2014.08.006
0893-6080/

whole utterance, derived as a point estimate of the latent variables in a factor analysis model (Dehak, Torres-Carrasquillo, Reynolds, & Dehak, 2011; Kenny, Oullet, Dehak, Gupta, & Dumouchel, 2008). However, while proven to be successful in a variety of scenarios, *i*-vector-based approaches suffer from two major drawbacks when coping with real-time applications. First, the *i*-vector is a point estimate and its robustness quickly degrades as the amount of data used to derive it decreases. Note that the smaller the amount of data, the larger the variance of the posterior probability distribution of the latent variables, and thus, the larger the *i*-vector *uncertainty*. Second, in real-time applications, most of the costs associated with *i*-vector computation occur after completion of the utterance, which introduces an undesirable latency.

Motivated by the prominence of deep neural networks (DNNs), which surpass the performance of the previous dominant paradigm, Gaussian mixture models (GMMs), in diverse and challenging machine learning applications – including acoustic modelling (Hinton et al., 2012; Mohamed, Dahl, & Hinton, 2012), visual object recognition (Ciresan, Meier, Gambardella, & Schmidhuber, 2010), and many others (Yu & Deng, 2011) – we previously introduced a successful LID system based on DNNs in Lopez-Moreno et al. (2014). Unlike previous works on using shallow or convolutional neural networks for small LID tasks (Cole, Inouye, Muthusamy, & Gopalakrishnan, 1989; Leena, Srinivasa Rao, & Yegnanarayana, 2005; Montavon, 2009), this was, to the best of our knowledge, the first time that a DNN scheme was applied at a large scale for LID and benchmarked against alternative state-of-the-art approaches. Evaluated using two different datasets—the NIST LRE 2009 (3 s task) and Google 5M LID—this scheme significantly outperformed several *i*-vector-based state-of-the-art systems (Lopez-Moreno et al., 2014).

In the current study, we explore different aspects that affect DNN performance, with a special focus on very short utterances and real-time applications. We believe that the DNN-based system is a suitable candidate for this kind of application, as it could potentially generate decisions at each processed frame of the test speech segment, typically every 10 ms. Through this study, we assess the influence of several factors on the performance, namely: (a) the amount of required training data, (b) the topology of the network, (c) the importance of including the temporal context, and (d) the test utterance duration. We also propose several blind techniques to combine frame-by-frame posteriors obtained from the DNN to get identification decisions.

We conduct the experiments using the following LID datasets: a dataset built from Google data, hereafter, Google 5M LID corpus and the NIST Language Recognition Evaluation 2009 (LRE'09). First, by means of the Google 5M LID corpus, we evaluate the performance in a real application scenario. Second, we check if the same behaviour is observed in a familiar and standard evaluation framework for the LID community. In both cases, we focus on short test utterances (up to 3 s).

The rest of this paper is organized into the following sections. Section 2 defines a reference system based on *i*-vectors. The proposed DNN system is presented in Section 3. The experimental protocol and datasets are described in Section 4. Next, we examine the behaviour of our scheme over a range of configuration parameters in both the task and the neural network topology. Finally, Sections 6 and 7 are devoted to presenting the conclusions of the study and potential future work.

2. Baseline system: *i*-vector

Currently, most acoustic approaches to perform LID rely on *i*-vector technology (Dehak, Kenny, Dehak, Dumouchel, & Ouellet, 2011). All such approaches, while sharing *i*-vectors as a feature representation, differ in the type of classifier used to perform the final language identification (Martinez, Plchot, Burget, Glembek, & Matejka, 2011). In the rest of this section we describe: (a) the *i*-vector extraction procedure, (b) the *i*-vector classifier used in this study, and (c) the configuration details of our baseline *i*-vector system. This system will serve us as the baseline system.

2.1. I-vector extraction

Based on the MAP adaptation approach in a GMM framework (Reynolds, 1995), utterances in language or speaker recognition are typically represented by the accumulated zero- and centred first-order Baum–Welch statistics, N and F, respectively, computed from a Universal Background Model (UBM) λ. For the UBM mixture $m \in 1, \ldots, C$, with mean, μ_m, the corresponding zero- and centred first-order statistics are aggregated over all frames of the utterance as

$$N_m = \sum_t p(m|o_t, \lambda) \tag{1}$$

$$F_m = \sum_t p(m|o_t, \lambda)(o_t - \mu_m), \tag{2}$$

where $p(m|o_t, \lambda)$ is the Gaussian occupation probability for the mixture m given the spectral feature observation $o_t \in \Re^D$ at time t.

The total variability model, hereafter TV, can be seen as a classical FA generative model (Bishop, 2007), with observed variables given by the *supervector* (CD $\times$ 1) of stacked statistics $F = \{F_1, F_2, \ldots, F_C\}$. In the TV model, the vector of hidden variables $w \in \Re^L$ is known as the utterance *i*-vector. Observed and hidden variables are related by the rectangular low rank matrix $T \in \Re^{CD \times L}$

$$N^{-1}F = Tw, \tag{3}$$

where the zero-order statistics N are represented by a block diagonal matrix $\in \Re^{CD \times CD}$, with C diagonal $D \times D$ blocks. The mth component block is the matrix $N_m I_{(D \times D)}$. Given the imposed Gaussian distributions of $p(w)$ and $p(F|w)$, it can be seen that the mean of the posterior $p(w|F)$ is given by

$$w = (I + T^t \Sigma^{-1} N T)^{-1} T^t \Sigma^{-1} F, \tag{4}$$

where $\Sigma \in \Re^{CD \times CD}$ is the diagonal covariance matrix of F. The TV model is thus a data driven model with parameters $\{\lambda, T, \Sigma\}$. Kenny et al. (2008) provides a more detailed explanation of the derivation of these parameters, using the EM algorithm.

2.2. Classification

Since T constrains all the variabilities (i.e. language, speaker, session), and it is shared for all the language models/excerpts, the *i*-vectors, w, can be seen as a new input feature to classify. Further, several classifiers—either discriminative (i.e. Logistic Regression) or generative (i.e. the Gaussian classifier and linear discriminant analysis)—can be used to perform classification (Martinez et al., 2011). In this study, we utilized LDA, followed by cosine distance (LDA_CS), as the classifier.

Even though using a more sophisticated classifier (Lopez-Moreno et al., 2014) would have resulted in slightly increased performance, we chose the LDA_CS considering the trade-off between performance and computational time efficiency. In this framework, the similarity measure (score) of the two given *i*-vectors, w_1 and w_2, is obtained as

$$S_{w_1,w_2} = \frac{(A^t w_1)(A^t w_2)}{\sqrt{(A^t w_1)(A^t w_1)}\sqrt{(A^t w_2)(A^t w_2)}} \tag{5}$$

where A is the LDA matrix.