Available online at www.sciencedirect.com
ScienceDirect
Computer Speech & Language 00 (2017) 1–16
www.elsevier.com/locate/csl

Multilingually trained bottleneck features in spoken language recognition☆

Radek Fér*, Pavel Matějka, František Grézl, Oldřich Plchot, Karel Veselý, Jan Honza Černocký

Brno University of Technology, Speech@FIT and IT4I Center of Excellence, Božetěchova 2, Brno 612 66, Czech Republic

Received 5 October 2016; received in revised form 21 June 2017; accepted 28 June 2017

Abstract

Multilingual training of neural networks has proven to be simple yet effective way to deal with multilingual training corpora. It allows to use several resources to jointly train a language independent representation of features, which can be encoded into low-dimensional feature set by embedding narrow bottleneck layer to the network. In this paper, we analyze such features on the task of spoken language recognition (SLR), focusing on practical aspects of training bottleneck networks and analyzing their integration in SLR. By comparing properties of mono and multilingual features we show the suitability of multilingual training for SLR. The state-of-the-art performance of these features is demonstrated on the NIST LRE09 database.

Keywords: Multilingual training; Bottleneck features; Spoken language recognition

1. Introduction

Neural networks (NN) have become a widely used technique for state-of-the-art Large Vocabulary Continuous Speech Recognition (LVCSR) systems and are rapidly expanding to other fields of speech recognition. Notably, bottleneck (BN) features (Kramer, 1991), extracted from a narrow layer of NN, have brought speech signal parametrization to a quantitatively different level (Grézl et al., 2007). These features convey information about phonetic content in a nonlinearly compressed form which can be directly used for the task of spoken language recognition (SLR), where they have demonstrated state of the art performance (Matějka et al., 2014; Jiang et al., 2014a; Ferrer et al., 2016).

Despite the excellent results, these features exhibit strong coupling to a language used during the NN training. This can be circumvented by means of multilingual training (Schultz and Waibel, 2001; Scanzio et al., 2008) and that is also the main focus of this paper. The term multilingual means that the NN is trained on several languages simultaneously. The NN thus learns (to some extent) a language independent representation of speech that gets

☆ This paper has been recommended for acceptance by Roger Moore.

* Corresponding author.
E-mail addresses: ifer@fit.vutbr.cz, radomilec@gmail.com (R. Fér).

http://dx.doi.org/10.1016/j.csl.2017.06.008
0885-2308/

encoded into bottleneck features. Such features were used for the LVCSR task and they were found to be superior to the ones trained on a single language (Scanzio et al., 2008; Veselý et al., 2012; Grézl et al., 2014).

In this paper, we extend our previous work (Fér et al., 2015) by showing more detailed results and further analysis of multilingual bottleneck features. Specifically, we focus on differences in mono- and multi-lingual features that should be addressed in context of SLR. We add experiments with different NN architectures and output layer setup. The SLR metrics are reported together with ASR related measures to see the level of their correlation. Note that all experiments from the original paper were re-scored using different SLR backend, so the corresponding values will not be the same.

The approach is tested on (clean) NIST LRE 2009 database (NIST, 2009), comparing the performance of monolingual (i.e., trained on single language) and multilingual systems. The results obtained with widely used Mel-frequency Cepstrum Coefficients (MFCC) Shifted Delta Cepstra (SDC) features (Torres-Carrasquillo et al., 2002) are included for reference.

1.1. Related work

Several different approaches to allow use of multilingual training corpora in SLR has been proposed. Zissman and Singer (1994) used six phoneme recognizers running in parallel, each producing a language-dependent likelihood based on an N-gram phonotactic model. Final score was obtained by averaging corresponding log-likelihoods. This is known as Parallel Phone Recognition followed by Language Modeling (PPRLM). Corredor-Ardoy et al. (1997) reported similar error rates to PPRLM approach when language dependent phonotactic models were trained on a merged phoneme set of four languages. The merging was done using Agglomerative Hierarchical Clustering with phoneme similarity based on Hidden Markov Model (HMM) phone likelihoods.

Big effort has been carried out for multilingual resource collection. Namely GlobalPhone, a high-quality multilingual database, was developed by the team from University of Karlsruhe (Schultz, 2002). In Schultz and Waibel (2001), International Phonetic Alphabet (IPA) was used to create a cross-lingual phoneme set by unifying the phoneme sets of different languages from this database.

The need for explicit phoneme set unification was mitigated in Scanzio et al. (2008) by dividing the output softmax layer of a NN into a set of independent softmax output layers, one for each training language. The authors show that for ASR, despite a lower word accuracy of multilingually trained features over baseline language-specific features, the multilingual features are more robust in conditions with non-native speakers.

The idea of language independent features based on universal speech attributes was investigated in Siniscalchi et al. (2013). They used manner and place of articulation to fully describe parts of speech in any language. In their SLR system, the sequences of these attributes were then modeled using a vector space modeling techniques. By using such articulatory features, there is no need for (language-dependent) phonetic transcriptions and the features can be also considered as language-independent.

The use of bottleneck features for SLR was investigated in Matějka et al. (2014); Jiang et al. (2014a). The authors of Matějka et al. (2014) report a 45% relative improvement to acoustic features baseline on DARPA RATS database. The authors of Jiang et al. (2014a) trained two deep bottleneck neural networks on English and Mandarin. The resulting features are then fused either on feature or score level.

Another approach to use neural networks in SLR was proposed by Lopez-Moreno et al. (2014, 2016): the NN was trained for frame-by-frame language classification. The final decision is based on language log-posteriors averaged over frames. This approach works great for short utterances. However, for long utterances the conventional i-vector approach (Lopez-Moreno et al., 2016) is still superior.

The use of Long Short Term Memory (LSTM) cells to directly classify languages has been investigated by Gonzalez-Dominguez et al. (2014); Zazo et al. (2016). The advantage of recurrent architectures is in natural handling of time context by memorizing internal state over time and also very small number of parameters compared to standard i-vector system. As it happens with other DNN approaches, this technique outperforms conventional i-vectors only in short durations.

A nice summary of neural network approaches for SLR can be found in Ferrer et al. (2016). The paper compares three types of features (SDC, bottleneck and probabilistic/posterior ones) that are modeled using two different approaches: standard GMM/UBM i-vector system and i-vector system using statistics collected using DNN alignment. The results show that for SLR, standard GMM/UBM i-vector system using bottleneck features performs the best.

Please cite this article as: R. Fér et al., Multilingually trained bottleneck features in spoken language recognition, Computer Speech & Language (2017), http://dx.doi.org/10.1016/j.csl.2017.06.008