Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 81 (2016) 144 – 151

Procedia
Computer Science

5th Workshop on Spoken Language Technology for Under-resourced Languages, SLTU 2016,
9-12 May 2016, Yogyakarta, Indonesia

Bottle-Neck Feature Extraction Structures for Multilingual Training and Porting

František Grézl*, Martin Karafiát

Brno University of Technology, Speech@FIT and IT4I Center of Excellence, Brno, Czech Republic

Abstract

Stacked-Bottle-Neck (SBN) feature extraction is a crucial part of modern automatic speech recognition (ASR) systems. The SBN network traditionally contains a hidden layer between the BN and output layers. Recently, we have observed that an SBN architecture without this hidden layer (i.e. direct BN-layer – output-layer connection) performs better for a single language but fails in scenarios where a network pre-trained in multilingual fashion is ported to a target language. In this paper, we describe two strategies allowing the direct-connection SBN network to indeed benefit from pre-training with a multilingual net: (1) pre-training multilingual net with the hidden layer which is discarded before porting to the target language and (2) using only the the direct-connection SBN with triphone targets both in multilingual pre-training and porting to the target language. The results are reported on IARPA-BABEL limited language pack (LLP) data.

Peer-review under responsibility of the Organizing Committee of SLTU 2016

Keywords: DNN topology; Stacked Bottle-Neck; feature extraction; multilingual training; system porting

1. Introduction

One of the recent challenges in speech recognition community is to build an ASR system with limited in-domain data. The data hungry algorithms for training ASR system components have to be modified to be effective with less data. This applies mainly to neural networks (NNs) which are part of essentially any state-of-the-art ASR system today and can be placed in any of the main ASR parts: feature extraction (e.g.[1]), acoustic model (e.g.[2]) and language model (e.g.[3]).

NNs usually have to be trained on a large amount of in-domain data in order to perform well. The need for large training data sets can be alleviated by layer-wise training[4] or unsupervised pre-training[5]. Another techniques such as dropout[6] and maxout[7] effectively reduce the number of parameters in the neural network during the training.

* Corresponding author. Tel.: +420-541-141-280 ; fax: +420-541-141-270.
E-mail address: grezl@fit.vutbr.cz

Peer-review under responsibility of the Organizing Committee of SLTU 2016
doi:10.1016/j.procs.2016.04.042

To improve the performance of a neural network, its size can be increased. The above mentioned dropout and maxout techniques are employed to prevent over-training. The over-training can be also prevented by introducing a regularization term into the objective function[8,9].

Another way to improve NN performance is to impose a certain structure on the NN or compose more NNs together. The typical example of the first method are Convolutive Neural Networks[10,11]. The NN compositions typically consist of two NNs, where the outputs of one NN form inputs to the other one. Those composed NNs are mostly used in place of feature extractors and the most typical compositions today are Stacked Bottle-Neck (SBN)[1], Hierarchical MRASTA[12] and Shifting Deep Bottle-Neck[13] which is very similar to[1] and its one-network version[14]. It became evident that there are two factors important for the success of these compositions:

- compression of the features through a Bottle-Neck (BN) layer[15]
- putting larger contexts of the first NN outputs into the input of the second NN

Another advantage of using a Bottle-Neck layer in a NN, at least to our experience, is, that it serves as some form of regularization and other regularization techniques are not necessary.

The IARPA BABEL program with its goal to quickly train a keyword spotting system for new language with minimum in-domain transcribed speech data encouraged research in training multilingual NN and porting such multilingual NN to new language[16,17,18]. Thus the effort to improve the NN performance has to be evaluated also in the context of multilingual training and porting of trained NN to target language.

2. Experimental setup

The setup is adopted from[16] and all results are directly comparable.

2.1. Data

The IARPA Babel Program requires the use of a limited amount of training data which simulates the case of what one could collect in limited time from a completely new language. It consists mainly of telephone conversation speech, but scripted recordings as well as far field recordings are present too. Two training scenarios are defined for each language – Full Language Pack (FLP), where all collected data are available for training – about 100 hours of speech; and Limited Language Pack (LLP) consisting only of one tenth of FLP. As training data, we consider only the transcribed speech. Vocabulary and language model (LM) training data are defined with respect to the Language Pack. They consist of speech word transcriptions of the given data pack.

The following data releases were used in this work: Cantonese IARPA-babel101-v0.4c (CA), Pashto IARPA-babel104b-v0.4aY (PA), Turkish IARPA-babel105-v0.6 (TU), Tagalog IARPA-babel106-v0.2g (TA), Vietnamese IARPA-babel107b-v0.7 (VI), Assamese IARPA-babel102b-v0.5a (AS), Bengali IARPA-babel103b-v0.4b (BE), Haitian Creole IARPA-babel201b-v0.2b (HA), Lao IARPA-babel203b-v3.1a (LA) and Zulu IARPA-babel206b-v0.1e (ZU).

The characteristics of the languages can be found in[19]. The FLP data of IARPA-babel10* (CA, PA, TU, TA, VI, AS, BE) languages are used for multilingual NN training. The rest of the languages (HA, LA, ZU) are considered as target languages. LLP data are used for NN porting and for training of GMM-HMM system. Statistics for LLP training set of target languages are given in Tab. 1 together with the development set used for system evaluation. The amounts of data refer to the speech segments after dropping the long portions of silence.

2.2. SBN DNN hierarchy for feature extraction

The NN input features are composed of logarithmized outputs of 24 Mel-scaled filters applied on squared FFT magnitudes (critical band energies, CRBE) and 10 F0-related coefficients. The filter-bank spans frequencies from 64Hz to 3800Hz. The F0-related coefficients consist of F0 and probability of voicing estimated according to[20] and