Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 81 (2016) 15 – 22

Procedia
Computer Science

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

Study of Large Data Resources for Multilingual Training and System Porting

František Grézl*, Ekaterina Egorova, Martin Karafiát

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

Abstract

This study investigates the behavior of a feature extraction neural network model trained on a large amount of single language data ("source language") on a set of under-resourced target languages. The coverage of the source language acoustic space was changed in two ways: (1) by changing the amount of training data and (2) by altering the level of detail of acoustic units (by changing the triphone clustering). We observe the effect of these changes on the performance on target language in two scenarios: (1) the source-language NNs were used directly, (2) NNs were first ported to target language.

The results show that increasing coverage as well as level of detail on the source language improves the target language system performance in both scenarios. For the first one, both source language characteristic have about the same effect. For the second scenario, the amount of data in source language is more important than the level of detail.

The possibility to include large data into multilingual training set was also investigated. Our experiments point out possible risk of over-weighting the NNs towards the source language with large data. This degrades the performance on part of the target languages, compared to the setting where the amounts of data per language are balanced.

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

Keywords: Stacked Bottle-Neck; feature extraction; multilingual training; large data; Fisher database

1. Introduction

Multilingual resources are of great help in case the data from the target language are not sufficient to train good acoustic model. In such case, the multilingual model, which is usually trained beforehand, is ported to the target language. Such multilingual models outperform the model trained only on limited target data[1,2,3]. The same holds for neural network model used for feature extraction[4,5,6].

It has been shown that increasing the number of languages used to train the multilingual model decrease the WER after porting to target language[5]. However, in multilingual training, it is desired to use the languages which are potentially close to the target one. A study has been carried out to show that careful selection of languages used to

* 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.2016.04.024

train the multilingual NN leads to an improvement on target language[7]. Improvements can carry past the porting stage, when the NN is retrained on the target language data. The selection does not have to stop on the level of the language as an atomic unit. It is possible to select only appropriate sentences or even frames.

The disadvantage of such language selection is the necessity to know the target language a-priori and subsequent training of the multilingual NN on potentially large amount of data. Moreover, the optimal thresholding for data selection may differ depending on the target language. Also, very distinct languages may not benefit from this technique at all.

Although there is a lot of studies comparing different strategies of multilingual NN training and theirs effect on the target language e.g.[8,9,10,11,12] a detailed analysis of the important issues of training language handling is missing. By a handling we mean properties, which can be altered. For example, the acoustic characteristic of the language cannot be changed, but we can change the modeling or labeling granularity of the given acoustic space.

Such study can be better made on a single training language, thus eliminating interaction between training languages during multilingual NN training. The advantage of multilingual training is a rich phoneme set seen over several languages, but variety in used recording device, which can be also language dependent as certain locations may tend to use specific handsets is a clear drawback. There is a danger of conditioning certain phonemes rather on the audio channel than on the underlying acoustic information. A big collection of one language should provide rather homogeneous recording conditions.

In this work we study the behavior of NN used for feature extraction trained on a large database corpus – English Fisher. Although the performance of ported monolingual system would be worse in comparison with the multilingual one (due to the limited acoustic space coverage and phonemic variability), it still should reveal the trends.

The focus of this study is to find out, how the coverage and partitioning of acoustic space bounded by a single language phonology will affect the performance in the target language. The limitation to a single language phoneme set makes it possible to alter the phonetic resolution of the acoustic space by means of triphone clustering. Such change should reveal if finer resolution of otherwise the same acoustic space will lead to better performance in the target language. An alternative to the phonetic resolution is acoustic coverage of the units. By changing the amount of training data – by changing the number of speakers as well as the number of utterances per speaker, the acoustic variability of given unit will also change.

It would be also interesting to see how the language with large data can be used together with a multilingual set. The databases used for multilingual model training are usually more or less balanced. If the aim of multilingual processing is to use any transcribed data, large differences in the amounts of data per language may appear. A case study can reveal if this might be a problem or if the multilingual training procedure can deal with it.

2. Experimental setup

In this study, we observe the WER obtained from a tandem[13] system where the features for the final GMM-HMM classifier are the Bottle-Neck (BN)[14] features obtained from Stacked Bottle-Neck (SBN) Neural Network (NN) hierarchy[15]. A simple maximum-likelihood trained model without any speaker adaptation is used.

The GMM-HMM model is trained on the target language which is represented by the limited language pack of the following data sets release:

Telugu – TE – IARPA-babel303b-v1.0a – is a Dravidian language spoken in the south-eastern part of India. Telugu phoneme set used for the experiments contains 39 phonemes, vowels showing long/short dichotomy and containing two diphthongs. Consonant set contains quite a few retroflex phonemes.

Lithuanian – LI – IARPA-babel304b-v1.0b – language belongs to the family of Baltic languages, and the phoneme set used for the experiments consists of 110 phonemes. On vowels and voiced consonants, it contains markings of stress and of falling or rising tone where applicable. Apart from that, vowels have long and short versions. Nearly every consonant in the Lithuanian consonant set has two versions: palatalized and non-palatalized

Haitian Creole – HA – IARPA-babel201b-v0.2b – a French Creole language spoken in Haiti. It is based mainly on French, but is also influenced by other European languages, such as Spanish and Portuguese, and West African languages. The phoneme set is relatively simple, with just 32 phonemes, all of them typical to the aforementioned European languages.