Accepted Manuscript

Cross-lingual Adaptation of a CTC-based multilingual Acoustic Model

Sibo Tong, Philip N. Garner, Hervé Bourlard

 PII:
 S0167-6393(18)30030-X

 DOI:
 https://doi.org/10.1016/j.specom.2018.09.001

 Reference:
 SPECOM 2586

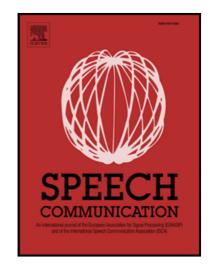
To appear in:

Speech Communication

Received date:26 January 2018Revised date:8 August 2018Accepted date:2 September 2018

Please cite this article as: Sibo Tong, Philip N. Garner, Hervé Bourlard, Cross-lingual Adaptation of a CTC-based multilingual Acoustic Model, *Speech Communication* (2018), doi: https://doi.org/10.1016/j.specom.2018.09.001

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.



Cross-lingual Adaptation of a CTC-based multilingual Acoustic Model

Sibo Tong^{1,2}, Philip N. Garner¹, Hervé Bourlard^{1,2}

¹Idiap Research Institute, Martigny, Switzerland ²Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland

Abstract

Multilingual models for Automatic Speech Recognition (ASR) are attractive as they have been shown to benefit from more training data, and better lend themselves to adaptation to under-resourced languages. However, initialisation from monolingual *context-dependent* models leads to an explosion of context-dependent states. Connectionist Temporal Classification (CTC) is a potential solution to this as it performs well with monophone labels.

We investigate multilingual CTC training in the context of adaptation and regularisation techniques that have been shown to be beneficial in more conventional contexts. The multilingual model is trained to model a universal International Phonetic Alphabet (IPA)-based phone set using the CTC loss function. Learning Hidden Unit Contribution (LHUC) is investigated to perform language adaptive training. During cross-lingual adaptation, the idea of extending the multilingual output layer to new phonemes is introduced and investigated. In addition, dropout during multilingual training and cross-lingual adaptation is also studied and tested in order to mitigate the overfitting problem.

Experiments show that the performance of the universal phoneme-based CTC system can be improved by applying dropout and LHUC and it is extensible to new phonemes during cross-lingual adaptation. Updating all acoustic model parameters shows consistent improvement on limited data. Applying dropout during adaptation can further improve the system and achieve competitive performance with Deep Neural Network / Hidden Markov Model (DNN/HMM) systems on limited data.

Keywords: multilingual Automatic Speech Recognition (ASR), Connectionist Temporal Classification (CTC), cross-lingual adaptation, Learning Hidden Unit Contribution (LHUC), dropout

1. Introduction

Automatic speech recognition (ASR) systems have improved dramatically in recent years. Although it has been shown that recognition accuracy can reach human parity on certain tasks (Xiong et al., 2017), building ASR systems with good performance requires a lot of training data. While sufficient data is available for languages like English, issues with data scarcity arise for under-resourced languages. Recently, there is increased interest in rapidly developing high performance ASR systems with limited data.

A common solution is to explore universal phonetic structures among different languages by sharing the hidden layers in deep neural networks (DNNs). In DNN, the hidden layers can be considered as a universal feature extractor. Therefore, the hidden layers can be trained jointly using data from multiple languages to benefit each other. The target of the multilingual DNN can be either the universal International Phonetic Alphabet (IPA) based multilingual senones (e.g., Dupont et al., 2005; Lin et al., 2009; Vu et al., 2014) or a layer consisting of separate activations for each language (e.g., Scanzio et al., 2008; Huang et al., 2013; Ghoshal et al., 2013; Heigold et al., 2013). The latter architecture has been shown to outperform the monolingual DNN but Lin et al. (2009) and our previous work (Tong et al., 2017) reported the performance of IPA-based multilingual DNN sometimes degrades. Although the universal model may share data among various language, mixture of data creates more variation especially for those identical IPA symbols shared among different languages.

Another common approach for creating models for lowresourced languages is to transfer the knowledge learned from other well-resourced languages to the target language. The bottleneck approach extracts features from a bottleneck layer of a multilingual model and uses bottleneck features as additional input to train the acoustic model of a target language (e.g., Thomas et al., 2012; Knill et al., 2013; Grézl et al., 2014). Bottleneck features are believed to contain a minimal multilingual subspace, they generalize well even on new languages. Knowledge can also be transferred by replacing the output layer of a well trained model and re-training the model to predict the targets of a low-resourced language (e.g., Huang et al., 2013; Ghoshal et al., 2013). The hidden layers are shared and transferred from rich-resourced languages to the target lowresourced language.

All of these models are based on a conventional DNN/HMM framework (Morgan and Bourlard, 1990, 1995; Hinton et al., 2012). In order to perform well, DNNs model context-dependent states to mitigate the error associated with the

Email address: {sibo.tong,phil.garner,bourlard}@idiap.ch (Sibo Tong^{1,2}, Philip N. Garner¹, Hervé Bourlard^{1,2})

Preprint submitted to Speech Communication

Download English Version:

https://daneshyari.com/en/article/10151553

Download Persian Version:

https://daneshyari.com/article/10151553

Daneshyari.com