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

Sibo Tong[1,2], Philip N. Garner[1], Hervé Bourlard[1,2]

[1] *Idiap Research Institute, Martigny, Switzerland*
[2] *Ecole 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 low-resourced 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 low-resourced 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. Garner[1], Hervé Bourlard[1,2])

Markov assumption. However, it creates more challenges for multilingual and cross-lingual ASR because of the large increase in context dependent labels arising from the phone set mismatch. According to Schultz and Waibel (2000), for example, 85% monophones in Portuguese can be covered by German, but the triphones coverage drops to 57%. Although approaches to adapt cluster trees have been proposed (Schultz and Waibel, 2000), the simple and effective way is to build a language-specific decision tree for the target language and replace the whole output layer of a DNN with the new targets, or to train a completely new network using bottleneck features.

Recently, the Connectionist Temporal Classification (CTC) framework has been successful in ASR (Graves et al., 2006). In CTC training, the neural network is trained to convert a sequence of acoustic features into a sequence of phones or graphemes. CTC-based systems learn to model context implicitly by the use of a recurrent neural network (RNN). Even monophone-based CTC systems can achieve equal or better performance than DNN/HMM hybrid systems when a large amount of data is available (Sak et al., 2015; Miao et al., 2016). Thus, a phoneme-based CTC system gets around the problem of context-dependent state mismatch, and does not require prior alignments between the input and output, potentially making the multilingual and cross-lingual modeling simpler and more straightforward.

CTC-based models, however, are more sensitive to the amount of training data. The advantage of CTC training over DNN/HMM can be exploited when adequate data is available. Therefore, we hypothesize that multilingual CTC training can further exploit the network by sharing data from multiple languages and that language adaptive training can also boost the performance as in DNN/HMM (Tong et al., 2017). To this end, we discuss the universal phoneme-based multilingual CTC-based model and language adaptive training in Section 3. Given the fact that the multilingual CTC model outputs monophone targets, we hypothesize that the universal phoneme-based multilingual CTC model can serve as a strong prior model when cross-lingual adaptation to a target language is required. Instead of removing the entire output layer and discarding all the information, the output layer of multilingual CTC model can be retained and easily extended to the unseen phonemes in the target languages. Different cross-lingual adaptation approaches based on the CTC framework are discussed in Section 4. In many of our preliminary experiments with CTC, consistent overfitting was observed on limited data. In preparatory work (Tong et al., 2018), we showed that dropout improves CTC-based cross-lingual adaptation. In order to further minimize the overfitting problem, we propose to apply dropout in also in multilingual training. We hypothesize that dropout can not only help avoid overfitting on limited data, but it can also prevent the multilingual model being overfitted in language-specific optimum during multilingual training, thus making the model more language-independent. Dropout is introduced in Section 5. Experimental results and analysis are provided in Section 6. Finally, Section 7 concludes the paper.

The contribution of this paper is threefold: First, we demonstrate that Learning Hidden Unit Contribution (LHUC) is an effective language adaptive training approach to improve multilingual CTC model. Second, we show phoneme-based multilingual CTC model is extensible to unseen phonemes during cross-lingual adaptation. Knowledge in the output layer can also be transferred to other languages. Third, dropout can help avoid overfitting in both multilingual CTC training and cross-lingual adaptation.

2. Related Work

When applied to acoustic modeling, CTC training allows the model to automatically learn the alignments between acoustic features and labels. Thus, CTC removes the need for building the initial Gaussian Mixture Model (GMM) to generate frame-level labels. Used together with Recurrent Neural Networks (RNNs), CTC has been shown to achieve state-of-the-art performance on large-scale acoustic modeling tasks (e.g., Graves and Jaitly, 2014; Hannun et al., 2014; Zweig et al., 2017).

Since the success of CTC training in ASR, there have been a few attempts to apply CTC also in multi-accent and multilingual ASR. Yi et al. (2016) used phoneme labels for training a multi-accent CTC-based ASR system in a multitask setting. Rao and Sak (2017) trained grapheme-based acoustic models for multi-accent speech recognition using a hierarchical recurrent neural network architecture with CTC loss. Different from multi-accent ASR, phoneme set or grapheme set is not the same across languages in multilingual problem. Some pre-published work (e.g., Kim and Seltzer, 2017; Müller et al., 2017b) investigated the use of a universal grapheme set by merging identical graphemes shared among languages and train the model using CTC loss. However, learning the spelling directly from acoustic features still requires large amount of data and graphemes can differ a lot from language to language. Müller et al. (2017a) and their recent work (Müller et al., 2017b) investigated phoneme-based multilingual CTC training with respect to label error rate. In this paper, we add to this knowledge base by also reporting word error rate (WER).

In phoneme-based multilingual training, identical phonemes can be shared among languages and can also be trained in language adaptive ways to further improve the performance. In previous work (Tong et al., 2017), we investigated several language adaptive training approaches originating from speaker adaptive training in DNN. Müller et al. (2016) additionally input a language code extracted from the bottleneck layer of a language classification DNN to enable language adaptive training in DNN. More recently, the same authors extended this approach to multilingual CTC training in their pre-published work (Müller et al., 2017b,c). However, it is demonstrated that the proposed approach cannot outperform the corresponding monolingual CTC models, although it yields improvement over the multilingual model. In this work, Learning Hidden Unit Contribution (LHUC) is shown to improve the multilingual CTC model and outperform the corresponding monolingual CTC models.

The cross-lingual ability of the CTC model has not been well studied. Kunze et al. (2017) shows a low-resource grapheme-based system can be initialized with a well-trained high-resourced model. In another pre-published work (Scharenborg et al., 2017), an iterative method is proposed to build a CTC-based ASR system for low-resourced languages, where the high-resourced model is iteratively adapted to the target language using the phoneme transcription generated from the adapted model. After independently investigating the CTC-based cross-lingual adaptation, we found that similar ideas had been very recently studied by Dalmia et al. (2018). However, The author used a multi-task multilingual CTC; the output consists of separate activations for each language. By contrast, our multilingual CTC system models the IPA-based universal phoneme set, and therefore it has the unique property that the output layer can be easily extended to new languages. Furthermore, this paper discusses dropout in the CTC-based cross-lingual adaptation and provides comparisons with DNN-based framework.

3. Multilingual CTC Model

3.1. CTC-based Acoustic Model

The Connectionist Temporal Classification (CTC) approach (Graves et al., 2006) is an objective function for sequence labeling problems without requiring any frame-level alignment between the input and target labels. CTC allows repetitions of output labels and extends the set of target labels with an additional *blank* symbol, which represents the probability of not emitting any labels at a particular time step. It introduces an intermediate representation called the CTC *path*. A CTC path is a sequence of labels at the frame level, allowing repetitions and the blank to be inserted between labels. The label sequence can be represented by a set of all the possible CTC paths that are mapped to it.

For an input sequence $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_T)$, the conditional probability $P(\mathbf{y}|\mathbf{X}, \theta)$ is then obtained by summing over all the probabilities of all the paths that correspond to the target label sequence $\mathbf{y}$ after inserting the repetitions of labels and the blank tokens, i.e.,

$$p(\mathbf{y}|\mathbf{X}, \theta) = \sum_{\hat{\mathbf{y}} \in \Omega(\mathbf{y})} p(\hat{\mathbf{y}}|\mathbf{X}, \theta) = \sum_{\hat{\mathbf{y}} \in \Omega(\mathbf{y})} \prod_{t=1}^{T} p(\hat{y}_t|\mathbf{x}_t, \theta) \tag{1}$$

where $\Omega(\mathbf{y})$ denotes the set of all possible paths that correspond to $\mathbf{y}$ after repetitions of labels and insertions of the blank token and θ represents the model parameters. The conditional probability of the labels at each time step, $P(\hat{y}_t|\mathbf{x}_t, \theta)$, is estimated using a neural network. The model can be trained to maximize (1) by using gradient descent, where the required gradients can be computed using the forward-backward algorithm (Graves et al., 2006).

As formulated by Zeyer et al. (2017), CTC can be identified as a special case of the generalized hybrid HMM/NN training procedure using the full-sum over the hidden state sequence. The generalized HMM training optimizes the likelihood of observing $\mathbf{X}$ given a target sequence $\mathbf{y}$ with state sequences $\mathbf{s}$ as hidden variable and model parameters θ, given by:

$$p(\mathbf{x}|\mathbf{y}, \theta) = \sum_{\mathbf{s}:\mathbf{y}} \prod_{t=1}^{T} p(s_t|s_{t-1}, \mathbf{y}) \cdot p(\mathbf{x}_t|s_t, \theta) \tag{2}$$

In HMM/NN models, $p(\mathbf{x}_t|s_t, \theta)$ is modeled as

$$p(\mathbf{x}_t|s_t, \theta) \sim \frac{p(s_t|\mathbf{x}_t, \theta)}{p(s_t)} \tag{3}$$

In this context, CTC can be considered as a special reduced HMM topology which has no transition probabilities, no state prior probability model but a special blank state and is trained with Baum-Welch soft alignments.

3.2. Universal Phone Set Multilingual CTC Model

The main goal of multilingual acoustic modelling is to share the acoustic data across multiple languages in order to learn the common properties shared among languages. Many present-day languages evolved from common ancestors. It is therefore natural that they share some common graphemes and phonemes. Very recently, building multilingual speech recognition systems using a universal grapheme set as output has been investigated (Kim and Seltzer, 2017; Toshniwal et al., 2017). However, modelling graphemes includes implicit modelling of spelling, which requires large amount of data. Moreover, graphemes can differ a lot from language to language. Languages that have nothing in common in terms of graphemes also share some common phonemes. With this motivation, and following Imseng et al. (2011), we propose a multilingual architecture that uses a universal output label set consisting of the union of all phonemes from the multiple languages. This universal phone set can be either derived in a data-driven way, or obtained from the International Phonetic Alphabet (IPA). In this study, the monolingual phones are merged if they share the same symbol in the IPA table. The network is trained to model the universal phoneme targets using the CTC loss function on data from multiple languages.

3.3. Learning Hidden Unit Contribution for Language Adaptive Training

Since the multilingual CTC model produces IPA targets, it may suffer the same problem as the IPA-DNN. Learning Hidden Unit Contribution (LHUC) was first proposed as a method for speaker adaptation (Swietojanski and Renals, 2014, 2016). It linearly re-combines hidden units in a speaker- or environment-dependent manner. Further investigation of LHUC in language adaptive training is provided in our previous work (Tong et al., 2017). Given language-specific data, LHUC re-scales the contributions (amplitudes) of the hidden units in the model without actually modifying their feature receptors. A language-dependent amplitude function is introduced to modify $\mathbf{o}_i^{sl}$, the hidden unit output of unit i in layer l for language s:

$$\mathbf{o}_i^{sl} = \xi(r_i^{sl}) \cdot \psi_i(\mathbf{o}^{l-1})$$

$r_i^{sl} \in \mathbb{R}$ is an adaptable language-dependent parameter, which is re-parametrised by a function $\xi : \mathbb{R} \to \mathbb{R}^+$. A sigmoid function with range $(0, 2)$ is usually used. ψ is the transformation function in a hidden layer. It can be, for instance, a feedforward or recurrent connection with non-linear activation or a Long Short-Term Memory (LSTM) block. ψ_i is the i^{th} row of the corresponding activations.

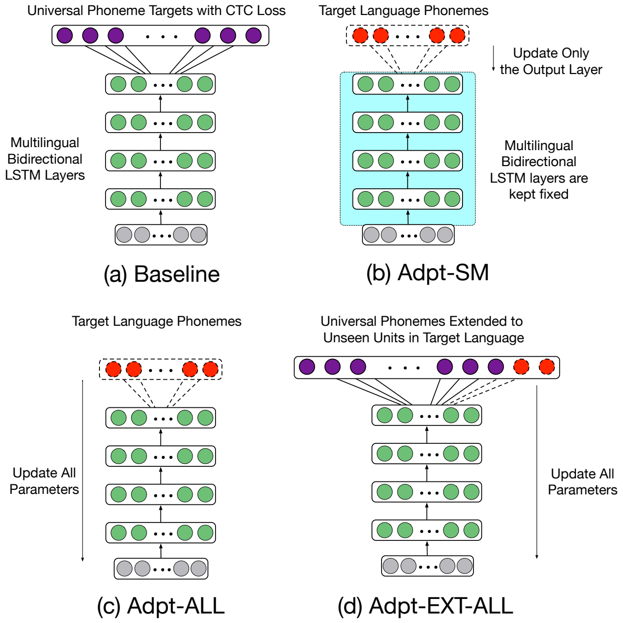
Universal Phoneme Targets with CTC Loss
Multilingual Bidirectional LSTM Layers
(a) Baseline
Target Language Phonemes
Update Only the Output Layer
Multilingual Bidirectional LSTM layers are kept fixed
(b) Adpt-SM
Target Language Phonemes
Update All Parameters
(c) Adpt-ALL
Universal Phonemes Extended to Unseen Units in Target Language
Update All Parameters
(d) Adpt-EXT-ALL

Figure 1: Approaches to adapt multilingual CTC model to the target language. (a) shows the baseline multilingual CTC model. In (b), a new Softmax (SM) output layer replaces the multilingual targets. The hidden layers are fixed and only the output layer is re-estimated. We can also update all the parameters as shown in (c). In (d), the multilingual CTC model is extended to new phonemes by adding new connections. Adaptation is performed by updating all the parameters.

The hidden units are trained to capture both good average representations and language-specific representations by estimating language-specific hidden unit amplitudes for each training language. In this paper, LHUC is applied as an approach of language adaptive training in the context of CTC training. The purpose is to improve the multilingual acoustic model for the given languages. Therefore, the language-specific parameters will not be re-estimated for existing languages after adaptive training.

4. CTC-based Cross-lingual Adaptation

In the DNN framework, the shared hidden layers extracted from the multilingual DNN can be considered to be an intelligent feature extractor and are transferable across languages (Huang et al., 2013). It is therefore interesting to investigate if the hidden layers in a CTC-based model can be carried over to distinguish phonemes in new languages.

The basic procedure of cross-lingual model adaptation on a CTC model is simple. As first proposed for DNN models by Huang et al. (2013), the output layer is removed and a new randomly initialized Softmax (SM) layer, corresponding to the target language phone set, is added on top of the hidden layers. Usually the hidden layers are fixed and only the softmax layer will be re-estimated using training data from the target language. If enough data is available, further tuning of the entire network can be considered.

One major advantage of the universal phoneme-based multilingual CTC model over the multilingual DNN is that monophone modeling gets around the problem of mismatch of context-dependent states. It therefore becomes straightforward to extend the existing multilingual model to extra phonemes when a new target language arrives. Therefore, we propose to extend the multilingual output layer by adding connections to the unseen mono phones of the target language, rather than discarding all the information already learned in the output layer. As is shown in Figure 1, those weights connecting to the unseen phones are randomly initialized and trained from scratch. The others can be quickly adapted from the multilingual model with little adaptation data.

5. Dropout

In many of our preliminary experiments with CTC, consistent overfitting was observed on limited data. Although adapting from multilingual model mitigates overfitting to some extent, the problem still exists. Dropout has been well established for feed forward networks by Srivastava et al. (2014), and it has been also proved to significantly improve the performance of LSTM networks for sequence labelling tasks (Reimers and Gurevych, 2017). More recently, various approaches of dropout on feedforward and recurrent connections were explored in the context of CTC (Billa, 2017). Inspired by this work, we propose to combine dropout with both multilingual training and cross-lingual adaptation to minimize overfitting on limited data. Moreover, we hypothesize applying dropout in multilingual training has an additional advantage: It can help the model avoid being overfitted in an optimum specific to any languages, thus making the model more language-independent. The dropout approach applied in this work is a combination of dropout on two different levels, as described by Billa (2017).

- **Dropout on feed forward connections** Dropout is applied on the feed forward connections at sequence level where the composite LSTM cell is the unit to be dropped. The dropout mask is retained across a complete utterance to eliminate cross-sampling noise.

- **Dropout on recurrent connections** Recurrent dropout without memory loss (Semeniuta et al., 2016) is applied to the incremental LSTM cell memory update at sequence level following

$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{m}_t \odot \mathbf{i}_t \odot \phi(\mathbf{W}_c \mathbf{x}_t + \mathbf{R}_c \mathbf{h}_{t-1} + \mathbf{b}_c) \quad (4)$$

  where $\mathbf{c}_t$ is the LSTM cell state at time t, $\mathbf{f}_t$ and $\mathbf{i}_t$ respectively denote the forget gate and input gate, $\mathbf{x}_t$ is the input vector at time t, $\mathbf{h}_{t-1}$ represents the LSTM output at time $t-1$, $\mathbf{W}_c$, $\mathbf{R}_c$ and $\mathbf{b}_c$ are the corresponding weights and bias, $\mathbf{m}_t$ represents the dropout mask at time t. The mask is again retained across a complete sequence.

For each minibatch, either a forward or recurrent dropout is picked randomly with equal probability. The combination was observed to outperform single dropout training.

Table 1: Statistics of the dataset of each language used in this work: the amounts of speech data are in hours.

Application	Language	Dataset	Train	Dev	Test
Multilingual Training	EN	WSJ	81h	1.1h	1.1h
	FR	BREF/GP	120h	10.3h	8.8h
	GE	BCN	136h	1.1h	5.7h
	Total Amount		337h		
Cross-lingual Adaptation	PO	GP	21h	1.6h	1.8h

6. Experiments

6.1. Experimental Database

We investigated the performance of the proposed universal phoneme-based CTC model on English (EN), French (FR), and German (GE). The English data was obtained from the Wall Street Journal (WSJ) corpus (Paul and Baker, 1992). Data preparation gave us 81 hours of transcribed training speech. WSJ dev93 and the union of eval92 and eval93 were used as the development set and the evaluation set, respectively. The French data was extracted from the BREF (Lamel et al., 1991) and GlobalPhone (GP) corpora (Schultz et al., 2013), which consist of 120 hours of data. From the German Broadcast News (BCN) corpus (Weninger et al., 2014), we used 136 hours of data for training. In total, 337 hours of multilingual data was used for multilingual CTC training. All the training data is quite clean read speech from similar acoustic conditions. In cross-lingual adaptation experiments, Portuguese (PO) from GlobalPhone was considered as the target low-resourced language, which has only 21 hours data. The detailed statistics for each of the languages are shown in Table 1. The development sets were used to tune the hyper-parameters for training.

6.2. Setup

We used 40-dimensional log-mel filterbank coefficients as acoustic features together with their first and second-order derivatives, derived from 25 ms frames with a 10 ms frame shift. The features were normalized via mean subtraction and variance normalization on a speaker basis. All the monolingual phones were mapped to IPA symbols and we merged the phonemes from EN, FR and GE to create the universal phone set for multilingual training. Note that we removed the stress makers in EN phone set in order to map the phonemes to IPA symbols.

The multilingual CTC model has 4 layers of Bidirectional Long Short-Term Memory (BLSTM), with 320 cells in each layer and direction. All the weights in the models were randomly initialized and were trained using stochastic gradient descent with momentum. A learning rate of 0.00004 was used and early stopping on the validation set was applied to select the best model. For decoding, individual weighted finite-state transducer (WFST) decoding graphs were built using language-specific lexicons and language models. All the DNNs compared in this work have 6 hidden layers, each consisting of 1024 units. Thus, it contains slightly more parameters (8.8 vs 8.5 million) than the CTC models. All CTC models were trained based on the EESEN implementation (Miao et al., 2015) and DNN/HMM systems were built using the Kaldi (Povey et al., 2011).

Table 2: Comparison between monolingual CTC baseline systems and multilingual CTC training in WER(%). Dropout is not applied. Notice that the English test set is much smaller than those in French and German. However, we only use it to indicate trends, drawing more concrete conclusions from the French and German results.

	system	EN	FR	GE
	ML-DNN-LHUC	8.8	7.3	8.6
	monolingual CTC	9.5	8.5	8.9
sys 1	universal ML-CTC	9.6	8.1	9.0
sys 2	+LHUC	9.2	7.7	8.4

Table 3: Comparison between monolingual CTC baseline systems and multilingual CTC training in WER(%). Dropout is applied.

	system trained w/ dropout	EN	FR	GE
	monolingual CTC	9.2	7.7	8.7
sys 3	universal ML-CTC	9.4	7.8	8.3
sys 4	+LHUC	**8.9**	**7.4**	**7.8**

6.3. Results

6.3.1. Multilingual Training

Previous research has shown that an adequate amount of data is the key to training a good CTC-based system. We first evaluated if a better model can be trained using data from multiple languages. The comparison between multilingual CTC and baseline monolingual CTC systems is listed in Table 2. The table shows that multilingual CTC system sometimes fails to outperform monolingual models, even though it was trained on more data. We observed a similar result in our previous work on an IPA-based universal DNN system. Although the universal multilingual modelling enjoys richer data resources, the mixture of data creates more variation, especially for those identical IPA symbols shared among different languages. Similar degradation was also reported in another recent independent study (Müller et al., 2017b). This motivates us to apply language adaptive training in the multilingual CTC model. LHUC was applied on top of each bidirectional LSTM layer. Each language has its own corresponding LHUC parameters. As shown in Table 2, applying LHUC improves the multilingual performance and yields better word error rate (WER) than the monolingual CTC models in all languages.

It has been reported that dropout can help overcome the overfitting problem in monolingual CTC training (Billa, 2017). Dropout was further tested in multilingual conditions as described in Section 5 and the dropout rate was set to 0.2. Comparing Table 3 and Table 2, we can find that overfitting problem still exists in multilingual CTC training and dropout can help improve the generalization of the multilingual model. The systems trained with dropout consistently outperform the corresponding non-dropout systems in all languages. Combining LHUC and dropout yields the best performance.

Table 2 also lists the performance of the DNN-based multilingual training. Both models were trained on the same multilingual data with IPA labels. The IPA-based labels for the CTC training were obtained from the context-dependent state alignments of the multilingual GMM/HMM model. LHUC was also applied on top of each layer. Our experiment shows that dropout cannot improve DNN-based acoustic modeling. Therefore, dropout was not applied. The comparison shows multilingual CTC training can achieve competitive performance with DNN-based multilingual training.

6.3.2. Dropout in Cross-lingual Adaptation

While the first goal of this work was to create a universal phoneme-based multilingual model, we were interested in its transfer ability to other languages when the training data is limited. Previous experiments show that dropout is helpful in CTC training. We hypothesize that dropout can also improve cross-lingual adaptation where the available data is even more limited. In the present experiment, the multilingual model **sys 1** in Table 2 was used as the seed model, and cross-lingual adaptation was performed on limited amounts of Portuguese training data. The adaptation was done simply by replacing the multilingual output layer with a new output layer corresponding to the Portuguese phonemes and updating all the parameters. The same dropout strategy was tested on different amounts of adaptation data. As shown in Figure 2, although the improvement becomes smaller when more data is available, dropout consistently improves the adaptation performance. Similar improvements were also observed in the adaptation from other multilingual models and using different adaptation approaches in our experiments. Therefore, we keep applying dropout in the remaining cross-lingual adaptation experiments.

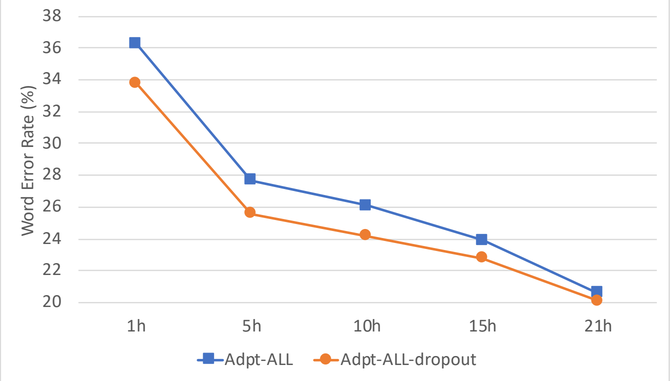

Figure 2: WERs (%) after cross-lingual adaptation with or without dropout.

6.3.3. Which Is the Best Seed Model for Cross-lingual Adaptation

The next problem is to choose the best multilingual model to initialize cross-lingual adaptation. In this work, the multilingual models **sys 1**, **sys 3** and **sys 4** were tested. We omitted **sys 2** as we have no a-priori belief that it will outperform **sys 4**. The adaptation was done simply by replacing the multilingual output layer with a new output layer corresponding to the Portuguese phonemes and updating all the parameters. When adapting an LHUC multilingual model, two approaches were compared: 1) updating the whole network after removing the LHUC layers and, 2) re-estimating Portuguese-specific LHUC parameters and the Softmax (SM) output layer while keeping the rest fixed. In comparison with the latter one, adapting only the output layer from **sys 3** was also tested.

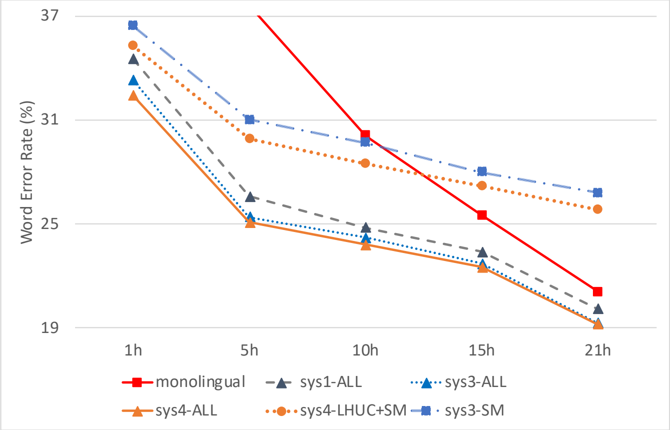

Figure 3: WERs (%) after cross-lingual adaptation of different multilingual models on various amounts of data. Dropout is applied in all systems. sys1-ALL denotes adapting all the parameters from **sys1**. sys4-ALL is updating the whole network after removing the LHUC layers. sys4-LHUC+SM represents adapting only the Softmax output layer and the LHUC parameters from **sys4**. sys3-SM is adapting only the output layer from **sys3**.

Comparing the **sys1-ALL** and **sys3-ALL**, we can clearly find that adaptation from the dropout multilingual model performs better. One conjecture is that dropout can help the multilingual model avoid being overfitted in a language-specific optimum and captures language-independent information better. Comparing **sys3-ALL** and **sys4-ALL**, we observed that the multilingual model trained with LHUC yields slightly better WER than the non-LHUC multilingual training when adapted to a new language, although the improvement is not significant. We did not report the performance of updating the LHUC parameters in addition to the whole network from **sys 4** because we found it is not helpful since the LHUC layers are merely additional adaptation parameters and may lead to overfitting.

Ideally, re-estimating only the LHUC parameters for Portuguese while keeping the rest fixed allows the adapted model to keep the performance on EN, FR and GE. However, adapting LHUC parameters as well as the output layer (**sys4-LHUC+SM**) performs already much worse than updating the whole network on the target language. Nevertheless, it yields improvement over updating only the output layer (**sys3-SM**), which still demonstrates the benefit of using an LHUC-based seed model. Given the above observation, **sys 4**, trained on 3 languages using LHUC and dropout, was used as the seed model for the following cross-lingual experiments.

6.3.4. Output Layer Extension in CTC-based Cross-lingual Adaptation

Although Figure 3 shows that updating all the parameters performs better than updating only the output layer, it is still worth investigating their performance after output layer extension. Therefore, four approaches were investigated in this section: re-training a new output layer and the LHUC parameters while keeping the others fixed (Adpt-LHUC+SM); extending the multilingual model by concatenating parameters corresponding to the new phonemes to the output layer and then updating the extended output layer and the LHUC layers (Adpt-EXT-LHUC+SM); updating the whole network with a randomly initialized new output layer (Adpt-ALL in Figure 1c); updating the whole network after extending the multilingual output layer to the target language (Adpt-EXT-ALL in Figure 1d). Experiments on different amounts of data were conducted using these approaches. Figure 4 shows all the comparisons.

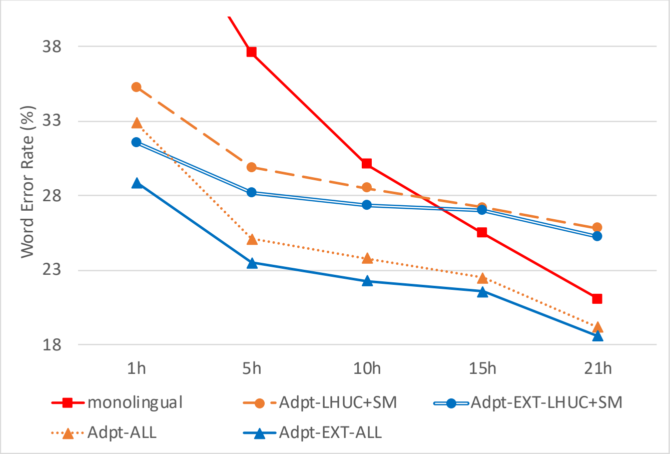

Figure 4: WERs (%) of different cross-lingual adaptation approaches. The WER of monolingual CTC model on 1 hour data is above 50% and exceeds the graph region. All models were trained with dropout.

From the figure, it can be found that adapting the whole network outperforms monolingual CTC training in all cases. It is difficult to train a good CTC model from scratch using less than 5 hours of data. However, the adaptation from a multilingual model can still achieve good performance. When the adaptation data is more than 15 hours, monolingual training beats the adaptation on only the output layer and the LHUC layers. Moreover, updating all the parameters still performs better than only re-training the output layer and the LHUC layers in all cases. We hence make the anecdotal inference that the BLSTM layers are more interdependent than those of the DNN (Huang et al., 2013); stronger inference would require more focused experiments. If we compare the blue lines and the orange ones, consistent improvement can be observed from extending the multilingual output layer. Although the difference becomes marginal with the increase of the adaptation data, it yields about 12% relative improvement on 1 hour adaptation data.

There are 19 extra unseen phonemes in Portuguese while 26 phonemes have been observed in multilingual training. As an example, we analyzed the phoneme error rate (PER) with respect to the overlapped phonemes and the new, unseen, phonemes separately on the development set during CTC training. The analysis was conducted on both adapting all the parameters and only the output layer plus LHUC layers, as plotted in Figure 5 and Figure 6. It shows that adaptation after extending the multilingual output layer keeps the same performance

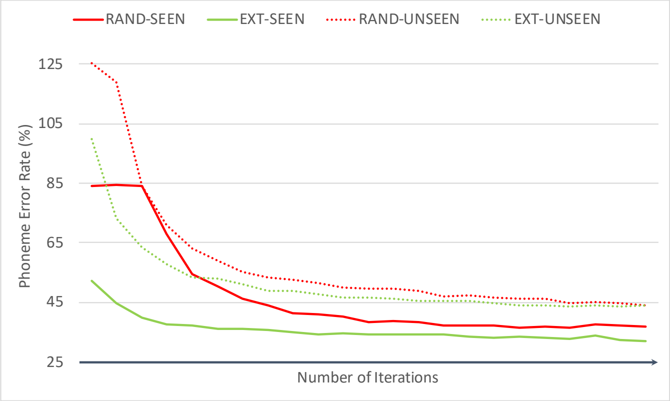

Figure 5: PERs (%) with respect to overlapped phonemes (SEEN) and new phonemes (UNSEEN) on PO development set. RAND denotes randomly initializing a new output layer before adaptation and EXT represents extending the multilingual output layer to the target language. The adaptation was performed by updating only the output layer and the LHUC layers on 1 hour data.

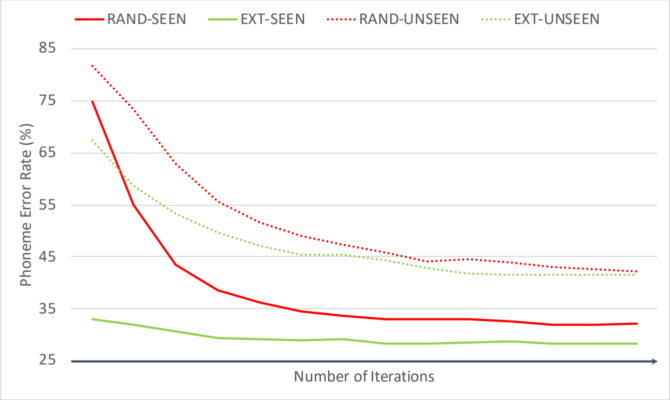

Figure 6: PERs (%) with respect to overlapped phonemes (SEEN) and new phonemes (UNSEEN) on PO development set. The adaptation was performed by updating the whole network on 1 hour data.

on unseen phonemes and converges much faster and better on seen phonemes. Although the adaptation data is limited, the extended model already has strong knowledge about the overlapped phonemes learned from multilingual training, and it is also able to catch up on new phonemes quickly.

6.3.5. *Comparison with DNN-based Cross-lingual Adaptation*

We also compared our best CTC-based cross-lingual adaptation with DNN/HMM-based adaptation approaches, as depicted in Figure 7. In the DNN-based adaptation, the multilingual DNN trained on the same multilingual data was used as seed model. We then replaced the multilingual output layer with Portuguese targets. The Portuguese context-dependent states and alignments were obtained from GMM/HMM systems trained on the corresponding amount of adaptation data. The adaptation was then performed by 1) updating the whole network, 2) Estimating the new output layer plus the LHUC layers while keeping the other parameters fixed and 3) updating only the output layer. Dropout was not applied for DNN since performance degradation was observed with dropout in our experiments.

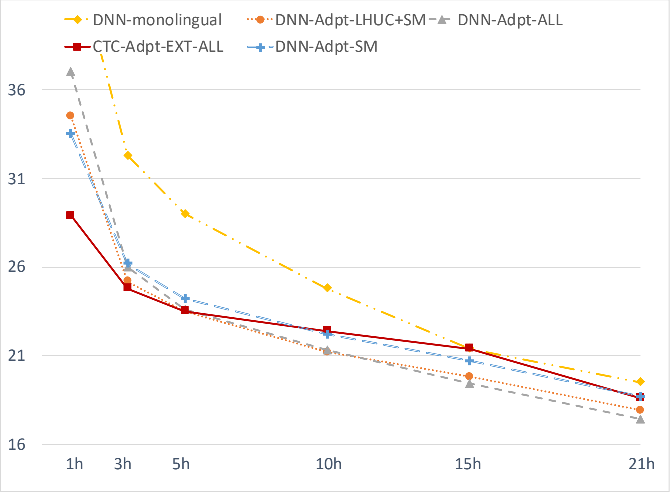

Figure 7: Comparison between CTC-based and DNN/HMM-based cross-lingual adaptation in WER(%). DNN-Adpt-ALL denotes updating all the parameters in DNN and DNN-Adpt-LHUC+SM represents updating the output softmax layer and the LHUC layers. DNN-Adpt-SM is only updating the output layer. The WER of monolingual DNN model on 1 hour data is above 40% and exceeds the graph region.

As shown in the figure, if comparing the DNN-based cross-lingual adaptation approaches, we can find that updating the output layer and the LHUC layers generally outperforms only updating the output layer except on the 1 hour data case. Updating all the parameters performs better than updating the output layer and the LHUC layers when more data is available but the difference is not significant. Meanwhile, updating the whole DNN also performs better than the CTC-based cross-lingual adaptation when adaptation data is more than 5 hours. However, CTC-based adaptation outperforms DNN/HMM based approaches when data is less than 3 hours. One conjecture is the CTC model retains the information about the phonemes that have been well modeled in multilingual training. Thus, it can be easily adapted and performs better than retraining the output layer from scratch in DNN. Given the fact that CTC training outperforms DNN/HMM hybrid modeling when sufficient data is available, we hypothesize CTC-based cross-lingual adaptation can surpass DNN-based approaches again if more data can be used for adaptation. We leave this for the future work.

7. Acknowledgement

This work has been conducted with the support of the European Community H2020 “Research and Innovation Action”funding, under “Scalable Understanding of Multilingual Media” (SUMMA) project n. 688139.

8. Conclusions

It was demonstrated that a universal phoneme-based multilingual CTC model can achieve competitive performance with DNN-based multilingual models. The universal phoneme-based multilingual CTC is extensible to new phonemes during cross-lingual adaptation. The extended model converges faster and better on overlapped phonemes and also catch up quickly on newly added phonemes. Combined with dropout during cross-lingual adaptation, the CTC-based model shows competitive performance with DNN/HMM-based adaptation on limited data.

References

Billa, J., 2017. Improving LSTM-CTC based ASR performance in domains with limited training data. arXiv preprint arXiv:1707.00722.

Dalmia, S., Sanabria, R., Metze, F., Black, A. W., 2018. Sequence-based multilingual low resource speech recognition. arXiv preprint arXiv:1802.07420.

Dupont, S., Ris, C., Deroo, O., Poitoux, S., 2005. Feature extraction and acoustic modeling: an approach for improved generalization across languages and accents. In: Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding.

Ghoshal, A., Swietojanski, P., Renals, S., 2013. Multilingual training of deep neural networks. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing.

Graves, A., Fernández, S., Gomez, F., Schmidhuber, J., 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In: Proceedings of the 23rd international conference on Machine learning.

Graves, A., Jaitly, N., 2014. Towards end-to-end speech recognition with recurrent neural networks. In: Proceedings of the 31st International Conference on Machine Learning. pp. 1764–1772.

Grézl, F., Karafiát, M., Veselỳ, K., 2014. Adaptation of multilingual stacked bottle-neck neural network structure for new language. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing.

Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., Prenger, R., Satheesh, S., Sengupta, S., Coates, A., et al., 2014. Deep speech: Scaling up end-to-end speech recognition. arXiv preprint arXiv:1412.5567.

Heigold, G., Vanhoucke, V., Senior, A., Nguyen, P., Ranzato, M., Devin, M., Dean, J., 2013. Multilingual acoustic models using distributed deep neural networks. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., et al., 2012. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal Processing Magazine.

Huang, J.-T., Li, J., Yu, D., Deng, L., Gong, Y., 2013. Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing.

Imseng, D., Bourlard, H., Dines, J., Garner, P. N., Magimai.-Doss, M., August 2011. Improving non-native ASR through stochastic multilingual phoneme space transformations. In: Proceedings of Interspeech. Florence, Italy.

Kim, S., Seltzer, M. L., 2017. Towards language-universal end-to-end speech recognition. arXiv preprint arXiv:1711.02207.

Knill, K., Gales, M. J., Rath, S. P., Woodland, P. C., Zhang, C., Zhang, S.-X., 2013. Investigation of multilingual deep neural networks for spoken term detection. In: Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding.

Kunze, J., Kirsch, L., Kurenkov, I., Krug, A., Johannsmeier, J., Stober, S., 2017. Transfer learning for speech recognition on a budget. arXiv preprint arXiv:1706.00290.

Lamel, L. F., Gauvain, J.-L., Eskénazi, M., et al., 1991. BREF, a large vocabulary spoken corpus for French.

Lin, H., Deng, L., Yu, D., Gong, Y.-f., Acero, A., Lee, C.-H., 2009. A study on multilingual acoustic modeling for large vocabulary ASR. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing.

Miao, Y., Gowayyed, M., Metze, F., 2015. EESEN: End-to-end speech recognition using deep RNN models and WFST-based decoding. In: Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding.

Miao, Y., Gowayyed, M., Na, X., Ko, T., Metze, F., Waibel, A., 2016. An empirical exploration of CTC acoustic models. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing.

Morgan, N., Bourlard, H., 1990. Continuous speech recognition using multilayer perceptrons with hidden markov models. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing.

Morgan, N., Bourlard, H., 1995. Neural networks for statistical recognition of continuous speech. Proceedings of the IEEE.

Müller, M., Stüker, S., Waibel, A., 2016. Language adaptive DNNs for improved low resource speech recognition. In: Proceedings of Interspeech.

Müller, M., Stüker, S., Waibel, A., 2017a. Language adaptive multilingual CTC speech recognition. In: International Conference on Speech and Computer.

Müller, M., Stüker, S., Waibel, A., 2017b. Multilingual adaptation of RNN based ASR systems. arXiv preprint arXiv:1711.04569.

Müller, M., Stüker, S., Waibel, A., 2017c. Phonemic and graphemic multilingual CTC based speech recognition. arXiv preprint arXiv:1711.04564.

Paul, D. B., Baker, J. M., 1992. The design for the Wall Street Journal-based CSR corpus. In: Proceedings of the workshop on Speech and Natural Language.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., et al., 2011. The Kaldi speech recognition toolkit. In: Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding.

Rao, K., Sak, H., 2017. Multi-accent speech recognition with hierarchical grapheme based models. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing.

Reimers, N., Gurevych, I., 2017. Optimal hyperparameters for deep LSTM-networks for sequence labeling tasks. arXiv preprint arXiv:1707.06799.

Sak, H., Senior, A., Rao, K., Irsoy, O., Graves, A., Beaufays, F., Schalkwyk, J., 2015. Learning acoustic frame labeling for speech recognition with recurrent neural networks. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing.

Scanzio, S., Laface, P., Fissore, L., Gemello, R., Mana, F., 2008. On the use of a multilingual neural network front-end. In: Ninth Annual Conference of the International Speech Communication Association.

Scharenborg, O., Ciannella, F., Palaskar, S., Black, A., Metze, F., Ondel, L., Hasegawa-Johnson, M., 2017. Building an ASR system for a low-research language through the adaptation of a high-resource language ASR system: Preliminary results.

Schultz, T., Vu, N. T., Schlippe, T., 2013. GlobalPhone: A multilingual text & speech database in 20 languages. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing.

Schultz, T., Waibel, A., 2000. Polyphone decision tree specialization for language adaptation. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing.

Semeniuta, S., Severyn, A., Barth, E., 2016. Recurrent dropout without memory loss. In: International Conference on Computational Linguistics.

Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. Journal of machine learning research.

Swietojanski, P., Renals, S., 2014. Learning hidden unit contributions for unsupervised speaker adaptation of neural network acoustic models. In: Proceedings of the IEEE Workshop on Spoken Language Technology.

Swietojanski, P., Renals, S., 2016. SAT-LHUC: Speaker adaptive training for learning hidden unit contributions. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing.

Thomas, S., Ganapathy, S., Hermansky, H., 2012. Multilingual MLP features for low-resource LVCSR systems. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing.

Tong, S., Garner, P. N., Bourlard, H., 2017. An investigation of deep neural networks for multilingual speech recognition training and adaptation. In: Proceedings of Interspeech.

Tong, S., Garner, P. N., Bourlard, H., 2018. Multilingual training and cross-lingual adaptation on CTC-based acoustic model. arXiv preprint arXiv:1412.5567, also published as Idiap research report 01-2018.
URL http://publications.idiap.ch/index.php/publications/show/3748

Toshniwal, S., Sainath, T. N., Weiss, R. J., Li, B., Moreno, P., Weinstein, E., Rao, K., 2017. Multilingual speech recognition with a single end-to-end model. arXiv preprint arXiv:1711.01694.

Vu, N. T., Imseng, D., Povey, D., Motlicek, P., Schultz, T., Bourlard, H., 2014. Multilingual deep neural network based acoustic modeling for rapid language adaptation. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing.

Weninger, F., Schuller, B., Eyben, F., Wöllmer, M., Rigoll, G., 2014. A broadcast news corpus for evaluation and tuning of German LVCSR systems. arXiv preprint arXiv:1412.4616.

Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M., Stolcke, A., Yu, D., Zweig, G., 2017. The Microsoft 2016 conversational speech recognition system. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing.

Yi, J., Ni, H., Wen, Z., Liu, B., Tao, J., 2016. CTC regularized model adaptation for improving LSTM RNN based multi-accent mandarin speech recognition. In: Chinese Spoken Language Processing (ISCSLP), 2016 10th International Symposium on.

Zeyer, A., Beck, E., Schlüter, R., Ney, H., 2017. CTC in the context of generalized full-sum HMM training. In: Proceedings of Interspeech.

Zweig, G., Yu, C., Droppo, J., Stolcke, A., 2017. Advances in all-neural speech recognition. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing.