Interspeech 2018
2-6 September 2018, Hyderabad

Improving Cross-Lingual Knowledge Transferability Using Multilingual TDNN-BLSTM with Language-Dependent Pre-Final Layer

Siyuan Feng, Tan Lee

Deparment of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong
siyuanfeng@link.cuhk.edu.hk, tanlee@ee.cuhk.edu.hk

Abstract

Multilingual acoustic modeling for improved automatic speech recognition (ASR) has been extensively researched. It's widely acknowledged that the shared-hidden-layer multilingual deep neural network (SHL-MDNN) acoustic model (AM) could outperform the conventional monolingual AM, due to its effectiveness in cross-lingual knowledge transfer. In this work, two research aspects are investigated, with the goal of improving multilingual acoustic modeling. Firstly, in the SHL-MDNN architecture, the shared hidden layer configuration is replaced by a combined TDNN-BLSTM structure. Secondly, the improvement of cross-lingual knowledge transferability is achieved through adding the proposed language-dependent pre-final layer under each network output. The pre-final layer, rarely adopted in past works, is expected to increase nonlinear modeling capability between universal transformed features generated by shared hidden layers and language-specific outputs. Experiments are carried out with CUSENT, WSJ and RASC-863 corpora, covering Cantonese, English and Mandarin. A Cantonese ASR task is chosen for evaluation. Experimental results show that SHL-MTDNN-BLSTM achieves the best performance. The proposed additional language-dependent pre-final layer brings moderate while consistent performance gains in various multilingual training corpora settings, thus demonstrates its effectiveness in improving cross-lingual knowledge transferability.

Index Terms: multilingual acoustic model, TDNN, BLSTM, cross-lingual knowledge transfer

1. Introduction

In recent years there has been a significant research interest in developing technologies for multilingual speech modeling, especially in the areas of acoustic modeling for automatic speech recognition (ASR) [1–6], keyword search [7], speech synthesis [8], and speech emphasis detection [9]. One of the driving purposes is to enable cross-lingual knowledge transfer [1], as motivated by humans that it is very natural to learn a language by borrowing information from other language resources. While different languages have distinctive linguistic properties, speech sounds can be produced may have significant overlap, because the basic mechanism of speech production is largely language-independent. Moreover, although large amounts of transcribed data for major languages like English or Mandarin are made commercially available nowadays [10, 11], there are plenty of low-resource languages lacking speech and linguistic resources. Multilingual modeling approaches could alleviate the data scarcity problem in low-resource acoustic [3, 5, 8] and language [12] modeling.

This work focuses on multilingual acoustic modeling, which aims at exploiting out-of-domain language resources to improve acoustic model (AM) for a target language. It has been widely acknowledged that multilingual acoustic modeling, especially in the context of deep neural network hidden Markov model (DNN-HMM) [13], could reduce word error rate (WER) for ASR tasks compared with its monolingual counterpart [1–3, 5]. Basically, there are two methods in the development of a multilingual DNN-HMM AM, namely, feature-based and model-based methods [14, 15]. For feature-based method, a bottleneck network is trained with multilingual corpora and used to extract bottleneck features (BNFs) for a target language, followed by downstream DNN-HMM acoustic modeling. For model-based method, multilingual resources are jointly employed to train a DNN-HMM AM, either sequentially [2] or in parallel [1]. The two methods could also be incorporated within one model [14].

The shared-hidden-layer multilingual DNN (SHL-MDNN) architecture, in which hidden layers are made common across many languages while the softmax layers are language dependent [1], is a milestone in the history of multilingual acoustic modeling. It enables joint optimization of DNN parameters by multiple languages simultaneously. The shared hidden layer architecture is regarded as a language-independent feature transform, which has been proved to work well for all languages involved during training [1]. In recent years, there were studies on combining SHL-MDNN with advanced machine learning techniques. For example, Zhou et al. [5] replaced hidden layers of SHL-MDNN by long short-term memory (LSTM) with residual learning, and achieved improvements in terms of ASR performance. Nevertheless, most researchers working on SHL-MDNN based architecture take it for granted, without doubting whether it is optimal in terms of cross-lingual knowledge transfer by sharing all hidden layers among multiple languages. Yosinski et al. [16] made an investigation to quantify the transferability of each hidden layer in image classification tasks, and found out transfer learning from a base task to a target task achieves the best performance when hidden layers except the last layer are transfered, followed by adding one layer on top, and retraining by the target task. Motivated by this, it naturally raises a question to us: Will the SHL-MDNN AM perform better if we add one language-dependent hidden layer before each block-softmax layer? To our best knowledge it is the first time the effectiveness of setting the language-dependent pre-final layer has been explicitly researched.

Time delay neural network (TDNN) [17] and (bidirectional) LSTM recurrent neural network ((B)LSTM-RNN) [18, 19] are structures able to capture long term temporal dependencies. Past works have shown their advantages over conventional structures such as Gaussian mixture model (GMM) or feed-forward neural network (FFNN)[1] on large vocabulary continuous speech recognition (LVCSR) [17–19], as well as speaker

[1]FFNN structure will be denoted as *DNN* without causing confusion.

10.21437/Interspeech.2018-1182

[20] and language recognition [21], probably due to their ability to make use of longer contextual information. Recently, network combination for further improving AMs has been actively investigated, especially the combination of TDNN and (B)LSTM [22, 23]. Cheng et al. [22] found out that a network with interleaving layers of TDNN and LSTM (TDNN-LSTM) outperforms BLSTM in terms of ASR performance. In the latest Arabic MGB-3 Challenge [24], a TDNN-BLSTM AM was adopted in Aalto system [23] and reported achieving better ASR performance than TDNN and TDNN-LSTM. This system performed the best in this challenge. Motivated by the works above, it drives us to investigate on the efficacy of a TDNN-BLSTM in multilingual acoustic modeling.

The rest of the paper is organized as follows. Section 2 briefly introduces TDNN and BLSTM structures and describes the proposed TDNN-BLSTM applied to SHL-MDNN with the additional pre-final layer. Section 3 introduces multilingual corpora and experimental setup. Experimental results and analyses are discussed in Section 4. Section 5 draws the conclusion.

2. SHL-MTDNN-BLTSM with pre-final layer

2.1. TDNN-BLSTM

An LSTM network transforms an input sequence $\boldsymbol{x} = (\boldsymbol{x_1}, \ldots, \boldsymbol{x_T})$ to an output sequence $\boldsymbol{y} = (y_1, \ldots, y_T)$ by computing equations from $t = 1$ to T [18]:

$$i_t = \sigma(W_{ix}\boldsymbol{x_t} + W_{ih}h_{t-1} + W_{ic}c_{t-1} + b_i), \quad (1)$$
$$f_t = \sigma(W_{fx}\boldsymbol{x_t} + W_{fh}h_{t-1} + W_{fc}c_{t-1} + b_f), \quad (2)$$
$$c_t = f_t \odot c_{t-1} + i_t \odot \tanh(W_{cx}\boldsymbol{x_t} + W_{ch}h_{t-1} + b_c), \quad (3)$$
$$o_t = \sigma(W_{ox}\boldsymbol{x_t} + W_{oh}h_{t-1} + W_{oc}c_t + b_o), \quad (4)$$
$$h_t = o_t \odot \tanh(c_t), \quad (5)$$
$$y_t = \phi(W_{yh}h_t + b_y), \quad (6)$$

where i_t, f_t, o_t, c_t, h_t are input gate, forget gate, output gate, cell activation and cell output activation vectors at time t, with the same size. $\odot$ is element-wise dot product, σ is sigmoid function, ϕ is the network output activation function, often in the form of softmax. As a commonly used alternative in practice, projected LSTM (LSTMP) could reduce computational complexity [18]. With LSTMP structure, the equations above change slightly, the h_t is replaced with r_t and the following is added:

$$r_t = W_{rh}h_t, \quad (7)$$
$$y_t = \phi(W_{yr}r_t + b_y), \quad (8)$$

where r_t is recurrent projection.

The shortcoming of LSTMs lies in the fact that it could only make use of previous context information. A BLSTM network extends to exploiting both previous and future context information by constructing pairs of forward and backward LSTM layers together before network output [19]. (B)LSTM could be trained by a truncated backpropagation through time (BPTT) algorithm [25].

A TDNN network captures long term temporal dependencies, with training times comparable to vanilla DNNs [17]. Unlike DNNs, TDNN transforms are tied across time steps. A typical TDNN model is shown as in Figure 1. The input context to layer 2, 3 and 4 are $\{-1, 0, 1\}, \{-1, 1\}$ and $\{-2, 2\}$. As can be seen that a higher layer could learn a wider range of temporal context. In this work, TDNN training algorithm follows greedy layer-wise supervised training with preconditioned stochastic gradient descent (SGD) updates, exponential learning rate schedule and mixing-up, following studies in [26].

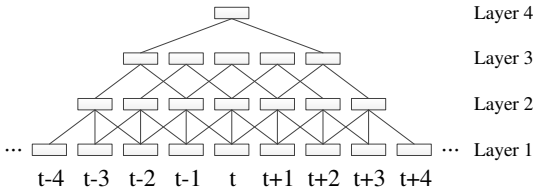

Figure 1: *A typical TDNN structure*

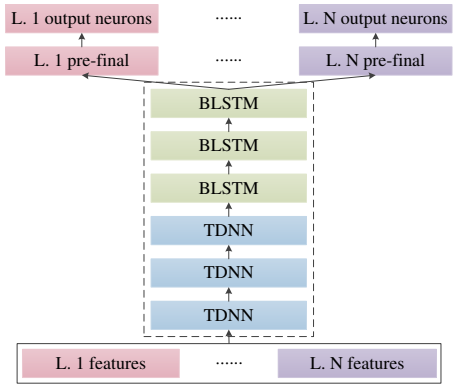

Figure 2: *The proposed SHL-MTDNN-BLSTM structure*

The proposed TDNN-BLSTM structure combines TDNN and BLSTM by constructing forward and backward pairs of LSTM layers on top of TDNN layers, as illustrated inside dashed box in Figure 2.

2.2. SHL-MTDNN-BLSTM with pre-final layer

The proposed TDNN-BLSTM applied to SHL-MDNN architecture is denoted as SHL-MTDNN-BLSTM, and illustrated in Figure 2. Different from the widely adopted SHL-MDNN architecture without concerning any specific hidden layer configuration [1, 5, 14], in this work, a language-dependent pre-final hidden layer is proposed to add in under each block-softmax output layer. We argue that in the conventional SHL-MDNN, it is unknown to us whether the direct connection between hidden layers and the block-softmax layer is fully capable of modeling the mappings between language-independent universal transformed features and the language-dependent HMM state classification task, especially while multilingual corpora used for AM training have a diverse range of phonetic and linguistic properties. The proposed pre-final layer increases nonlinear modeling capability between universal transformed features and language-dependent outputs, thus is expected to facilitate effective cross-lingual knowledge transfer.

Let $\boldsymbol{x_m^i}$ and y_m^i denote feature vector and target context-dependent HMM (CD-HMM) state label of the m-th training example in the k-th minibatch, where i denotes the language identity of $\{\boldsymbol{x_m^i}, y_m^i\}$. The total loss value of the k-th minibatch, $L(k)$, is defined as,

$$L(k) = \sum_{m=1}^{M} \omega^i l[\theta^i(\boldsymbol{x_m^i}), y_m^i], \quad (9)$$

where M is minibatch size, ω^i is task weight of the i-th lan-

Table 1: *Information about CA, EN and MA corpora*

Language:	CA	EN	MA
Training hours:	19.3	81.5	105.3
Test hours:	0.6	0.7	5.9

guage, θ^i denotes nonlinear transform from input to the i-th block-softmax output, l is loss function, cross-entropy [27] adopted in this paper. During network training, CD-HMM state classification errors within a certain language are back-propagated through the corresponding language-dependent pre-final layer and shared TDNN-BLSTM layers, while parameters of block-softmax layers and pre-final layers for the other languages keep unchanged.

3. Experimental setup

3.1. Multilingual corpora

The datasets used in this work cover three languages: CUSENT in Cantonese (CA), Wall Street Journal (WSJ) in English (EN) and RASC-863 in Mandarin (MA). CUSENT is a read speech corpus developed by The Chinese University of Hong Kong [28]. There are $20,378$ training utterances from 68 speakers, and 799 test utterances from other 8 speakers. WSJ is a read speech corpus [10]. The set *si284* is selected as training data, including $37,416$ utterances from 283 speakers. The set *eval92* is selected as test data, including 333 utterances from other 8 speakers. RASC-863 is a read speech corpus containing $89,003$ training utterances from 154 speakers, and $5,146$ test utterances from other 8 speakers [11]. Detailed information about the multilingual corpora is listed in Table 1.

3.2. Feature extraction and alignment generation

Mel-frequency cepstral coefficients (MFCCs) without cepstral truncation are used as input features, i.e., 40-dimensional MFCCs are computed at each time step [29]. MFCCs are spliced with a specific context size for a certain neural network as will be discussed in Section 4.1, and further appended with 100-dimensional i-vectors to perform instantaneous speaker adaptation. The i-vectors are extracted in an online version, where only frames prior to the current frame, including previous utterances of the same speaker, are used. A speed-perturbation method is used to augment training speech data three-fold, with speed factors of $0.9, 1.0$ and 1.1 [30].

Target labels for both monolingual and multilingual DNN-HMM hybrid AM training are state level phone alignments. A monolingual CD-GMM-HMM AM for each language is trained beforehand to generate alignments for training data, including original and speed-perturbed speech. These GMM-HMMs are based on 39-dimensional MFCCs+Δ+$\Delta\Delta$, and processed with linear discriminant analysis (LDA), maximum likelihood linear transform (MLLT) and feature-space maximum likelihood linear regression (fMLLR).

It is worth noting that basic acoustic units defined for MA in this work are different from those for EN and CA. This is mainly because we hoped to facilitate this study by re-using previously developed GMM-HMM systems. In Mandarin, each character is pronounced as a monosyllable, which can be composed of an *Initial* (onset) and a *Final* (rime), according to Hanyu Pinyin System [31]. *Initials* and *Finals* are adopted as the basic acoustic units for MA. For EN and CA, phone based acoustic units are adopted.

Table 2: *Context configurations for TDNN and TDNN-BLSTM*

	TDNN	TDNN-BLSTM
Layer	Layer context (TDNN) or LSTMP	
1	$\{-2,-1,0,1,2\}$	$\{-2,-1,0,1,2\}$
2	$\{-1,0,1\}$	$\{0\}$
3	$\{-1,0,1\}$	$\{-1,0,1\}$
4	$\{-3,0,3\}$	$\{-1,0,1\}$
5	$\{-6,-3,0\}$	LSTMP-forward
6	–	LSTMP-backward
7	–	LSTMP-forward
8	–	LSTMP-backward
9	–	LSTMP-forward
10	–	LSTMP-backward
11	–	LSTMP-forward
12	–	LSTMP-backward

4. Experiments

4.1. Baseline systems

Baseline systems include monolingual DNN, TDNN, BLSTM and TDNN-BLSTM, all trained with CUSENT training set. There are 300 training utterances randomly selected as validation data, in order to prevent overfitting. These models are trained on a cross-entropy criterion. For feature extraction and acoustic modeling, we use Kaldi toolkit [32].

DNN contains 6 hidden layers with 1024 neurons per layer, with ReLU activation and batch normalization (ReLU-batchnorm) [33]. Input MFCCs are spliced with ± 5. TDNN contains 5 ReLU-batchnorm layers with 1024 neurons per layer. BLSTM contains 4 pairs of forward and backward LTSMP layers [18], with 1024-dimensional cells and 256-dimensional recurrent projections. TDNN-BLSTM is composed of TDNN and forward-backward LSTMP layers, where TDNN layer width is 1024, LSTMP cell and recurrent projection dimensions are 1024 and 256, respectively. The temporal context configurations of TDNN and TDNN-BLSTM are summarized in Table 2. For DNN and TDNN, the number of training epochs is 3, learning rate starts from 1.5×10^{-2} to 1.5×10^{-3} with exponential decay, minibatch size is 256. For BLSTM and TDNN-BLSTM models, the number of training epochs is 6, minibatch size is 128, learning rate starts from 3×10^{-3} to 3×10^{-4}, also with exponential decay. A dropout method is adopted during network training, in order to improve generalization [22]. Dropout probability $p(n)$ with respect to training iterations n is piecewise linear, as described below,

$$p(n) = \begin{cases} 0.2 \times \frac{n}{N}, & 0 \le n \le \frac{N}{2} \\ 0.2 \times (1 - \frac{n}{N}), & \frac{N}{2} \le n \le N \end{cases} \quad (10)$$

where N is the number of iterations. Note that the choices of training hyperparameters and TDNN temporal contexts basically follow Kaldi default settings.

4.2. Results and analyses

For multilingual acoustic modeling, DNN, TDNN, BLSTM and TDNN-BLSTM are implemented in the SHL-MDNN architecture. Layer configurations are set the same as in baseline monolingual models, plus the language-dependent pre-final layer as proposed in Section 2.2. The pre-final layer is implemented by ReLU-renorm in Kaldi *nnet3* recipe, with 1024 neurons. Learning rate and minibatch sizes for multilingual AM training are consistent with corresponding baseline model settings. The number of training epochs for DNN and TDNN is 2, BLSTM and TDNN-BLSTM is 4. The dropout probability is constant

Table 3: *SERs of baseline and multilingual systems with optimized language weights*

Model	Training language(s)	#parameters	SER%
DNN	CA	8.3M	7.37
TDNN	CA	15.4M	6.34
BLSTM	CA	42.8M	6.67
TDNN-BLSTM	CA	56.4M	6.31
DNN	CA, EN:0.7, 0.3	14.2M	6.54
TDNN	CA, EN:0.6, 0.4	21.1M	5.99
BLSTM	CA, EN:0.7, 0.3	48.5M	6.45
TDNN-BLSTM	CA, EN:0.8, 0.2	62.0M	5.79
TDNN	CA, MA:0.8, 0.2	20.1M	5.93
TDNN-BLSTM	CA, MA:0.9, 0.1	60.9M	6.21
TDNN	CA, EN, MA:0.65, 0.25, 0.1	24.6M	5.75
TDNN-BLSTM	CA, EN, MA:0.65, 0.2, 0.15	65.5M	**5.50**

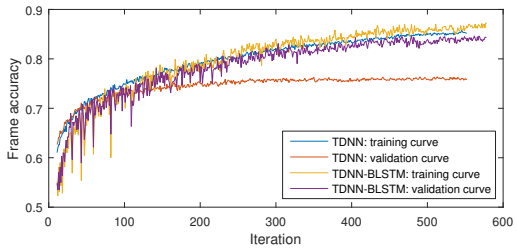

Figure 3: *Frame accuracy curves during TDNN and TDNN-BLSTM training with merged CA, EN and MA corpora*

0.1 for forward-backward LSTMP layers, and 0 for DNN or TDNN layers, as we find this to be optimal in multilingual acoustic modeling. EN and/or MA are merged with CA to form various multilingual training corpora. Similar to baseline model training, for each corpus, there are 300 training utterances randomly selected as validation set. Syllable error rates (SERs) of CA test set is chosen for evaluation. A syllable trigram language model trained with transcriptions of CA training data is used during decoding, using SRILM toolkit [34]. SERs of baseline and multilingual systems are listed in Table 3. Note that in this Table, SERs of multilingual systems with only the optimized language weight are listed. The corresponding language weight is specified right after training language identities. Figure 3 compares frame accuracy curves during training of TDNN and TDNN-BLSTM, with merged CA, EN and MA corpora. From Table 3 and Figure 3, the following observations are made:

(1) Multilingual models of DNN, TDNN, BLSTM and TDNN-BLSTM using merged CA and EN corpora outperform their monolingually trained counterparts, with relative improvements ranging from 3.3% to 11.3%. This result is generally in line with past works [1, 5].

(2) TDNN-BLSTM achieves the best Cantonese ASR performance among all neural network structures in both monolingual and multilingual training, with SERs 6.31% and 5.50%, respectively. Moreover, by observing frame accuracy curves, the generalization capability of the trained TDNN-BLSTM is stronger than TDNN.

(3) Compared with TDNN-BLSTM and TDNN, it can be observed that the achieved SER reduction from monolingual (CA) training to multilingual training is larger for TDNN-BLSTM than for TDNN. This shows the stronger modeling capability of TDNN-BLSTM as compared to TDNN, especially in the context of multilingual acoustic modeling. On the other hand, the model size of TDNN-BLSTM is significantly larger than TDNN, hence much more computational resources are required.

Table 4: *SERs of SHL-MTDNN-BLSTM with/without pre-final layer*

Training languages: Weights	With pre-final	Without pre-final
CA, EN, MA: 0.65, 0.2, 0.15	5.50	5.52
CA, EN: 0.8, 0.2	5.79	6.00
CA, MA: 0.9, 0.1	6.21	6.54

4.3. The effectiveness of the pre-final layer

The effectiveness of language-dependent pre-final layer is of great interest to us. We report our experimental results based on SHL-MTDNN-BLSTM model with the optional pre-final layer. Various training language identities are adopted to train SHL-MTDNN-BLSTM, as illustrated in Table 4. CA test set is chosen for evaluation. Experimental results are summarized as in Table 4. Note that language weights listed in this Table are all optimized ones.

As can be seen from Table 4, SHL-MTDNN-BLSTM with the pre-final layer brings moderate while consistent ASR performance improvements in various multilingual corpora settings, in comparison with those without the pre-final layer. This indicates that the conventional SHL-MDNN architecture may be suboptimal in modeling the mappings between language-independent features and language-dependent outputs. The pre-final layer could alleviate this problem and facilitate effective cross-lingual knowledge transfer through increasing nonlinear modeling capability between shared hidden layers and network outputs.

5. Conclusions and future works

This paper presents a study on improving multilingual acoustic modeling for ASR, in the context of SHL-MDNN AM architecture. Two research aspects are investigated. Firstly, the shared hidden layer configuration is replaced with the more advanced, TDNN-BLSTM structure. Secondly, the SHL-MDNN architecture is modified by adding the proposed language-dependent pre-final layer under each block-softmax output layer, with the goal of improving cross-lingual knowledge transferability. Experiments are carried out with multilingual corpora covering Cantonese, English and Mandarin. A Cantonese ASR task is selected for evaluation. Experimental results show that multilingual AMs consistently outperform their monolingual counterparts. Multilingual TDNN-BLSTM model trained with the three corpora achieves the best ASR performance, meanwhile preserving strong generalization capability. The language-dependent pre-final layer could bring moderate while consistent improvements in various multilingual training corpora settings, thus demonstrates its effectiveness in improving cross-lingual knowledge transferability.

Future works include detailed investigation on whether and how the effectiveness of our proposed pre-final layer is influenced by diverse linguistic dissimilarities within different groups of multilingual corpora, and the effect of basic acoustic unit definition mismatch on multilingual acoustic modeling.

6. Acknowledgements

We thank Yuzhong Wu for providing processed RASC-863 dataset. This research is partially supported by a GRF project grant (Ref: CUHK 14227216) from Hong Kong Research Grants Council.

7. References

[1] J.-T. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in *Proc. ICASSP*, 2013, pp. 7304–7308.

[2] A. Ghoshal, P. Swietojanski, and S. Renals, "Multilingual training of deep neural networks," in *Proc. ICASSP*, 2013, pp. 7319–7323.

[3] M. Karafiát, M. K. Baskar, P. Matějka, K. Veselỳ, F. Grézl, and J. Černocky, "Multilingual BLSTM and speaker-specific vector adaptation in 2016 BUT babel system," in *Proc. SLT*, 2016, pp. 637–643.

[4] S. Tong, P. N. Garner, and H. Bourlard, "An investigation of deep neural networks for multilingual speech recognition training and adaptation," in *Proc. INTERSPEECH*, 2017, pp. 714–718.

[5] S. Zhou, Y. Zhao, S. Xu, and B. Xu, "Multilingual recurrent neural networks with residual learning for low-resource speech recognition," in *Proc. INTERSPEECH*, 2017, pp. 704–708.

[6] J. Ma, F. Keith, T. Ng, M.-h. Siu, and O. Kimball, "Improving deliverable speech-to-text systems with multilingual knowledge transfer," in *Proc. INTERSPEECH*, 2017, pp. 127–131.

[7] C. Ni, L. Wang, C.-C. Leung, F. Rao, L. Lu, B. Ma, and H. Li, "Rapid update of multilingual deep neural network for low-resource keyword search," in *Proc. INTERSPEECH*, 2016, pp. 3698–3702.

[8] Q. Yu, P. Liu, Z. Wu, S. K. Ang, H. Meng, and L. Cai, "Learning cross-lingual information with multilingual BLSTM for speech synthesis of low-resource languages," in *Proc. ICASSP*, 2016, pp. 5545–5549.

[9] Y. Ning, Z. Wu, R. Li, J. Jia, M. Xu, H. Meng, and L. Cai, "Learning cross-lingual knowledge with multilingual BLSTM for emphasis detection with limited training data," in *Proc. ICASSP*, 2017, pp. 5615–5619.

[10] D. B. Paul and J. M. Baker, "The design for the Wall Street Journal-based CSR corpus," in *Proc. workshop on Speech and Natural Language*. Association for Computational Linguistics, 1992, pp. 357–362.

[11] A. Li, Z. Yin, T. Wang, Q. Fang, and F. Hu, "RASC863-a Chinese speech corpus with four regional accents," in *Proc. O-COCOSDA*, 2004.

[12] A. Ragni, E. Dakin, X. Chen, M. J. Gales, and K. M. Knill, "Multi-language neural network language models." in *Proc. INTERSPEECH*, 2016, pp. 3042–3046.

[13] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Trans. ASLP*, vol. 20, no. 1, pp. 30–42, 2012.

[14] M. Karafiát, M. K. Baskar, P. Matejka, K. Veselỳ, and F. Grézl, "2016 BUT babel system: Multilingual BLSTM acoustic model with i-vector based adaptation," in *Proc. INTERSPEECH*, 2017, pp. 719–723.

[15] T. Sercu, G. Saon, J. Cui, X. Cui, B. Ramabhadran, B. Kingsbury, and A. Sethy, "Network architectures for multilingual speech representation learning," in *Proc. ICASSP*, 2017, pp. 5295–5299.

[16] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Advances in neural information processing systems*, 2014, pp. 3320–3328.

[17] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Proc. INTERSPEECH*, 2015, pp. 3214–3218.

[18] H. Sak, A. W. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling." in *Proc. INTERSPEECH*, 2014, pp. 338–342.

[19] A. Graves, N. Jaitly, and A.-r. Mohamed, "Hybrid speech recognition with deep bidirectional LSTM," in *Proc. ASRU*, 2013, pp. 273–278.

[20] H. Zheng, S. Zhang, and W. Liu, "Exploring robustness of DNN/RNN for extracting speaker Baum-Welch statistics in mismatched conditions," in *Proc. INTERSPEECH*, 2015, pp. 1161–1165.

[21] Y. Qian, K. Evanini, P. L. Lange, R. A. Pugh, R. Ubale, and F. K. Soong, "Improving native language (L1) identifation with better VAD and TDNN trained separately on native and non-native English corpora," in *Proc. ASRU*, 2017, pp. 606–613.

[22] G. Cheng, V. Peddinti, D. Povey, V. Manohar, S. Khudanpur, and Y. Yan, "An exploration of dropout with LSTMs," in *Proc. INTERSPEECH*, 2017, pp. 1586–1590.

[23] P. Smit, S. Gangireddy, S. Enarvi, S. Virpioja, and M. Kurimo, "Aalto system for the 2017 arabic multi-genre broadcast challenge," in *Proc. ASRU*, 2017, pp. 338–345.

[24] A. Ali, S. Vogel, and S. Renals, "Speech recognition challenge in the wild: Arabic MGB-3," in *Proc. ASRU*, 2017, pp. 316–322.

[25] R. J. Williams and J. Peng, "An efficient gradient-based algorithm for on-line training of recurrent network trajectories," *Neural computation*, vol. 2, no. 4, pp. 490–501, 1990.

[26] X. Zhang, J. Trmal, D. Povey, and S. Khudanpur, "Improving deep neural network acoustic models using generalized maxout networks," in *Proc. ICASSP*, 2014, pp. 215–219.

[27] K. Veselỳ, A. Ghoshal, L. Burget, and D. Povey, "Sequence-discriminative training of deep neural networks." in *Proc. INTERSPEECH*, 2013, pp. 2345–2349.

[28] T. Lee, W. K. Lo, P. Ching, and H. Meng, "Spoken language resources for Cantonese speech processing," *Speech Communication*, vol. 36, no. 3, pp. 327–342, 2002.

[29] D. Povey, X. Zhang, and S. Khudanpur, "Parallel training of deep neural networks with natural gradient and parameter averaging," *arXiv preprint*, 2014.

[30] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition." in *Proc. INTERSPEECH*, 2015, pp. 3586–3589.

[31] Wikipedia, "Pinyin table — wikipedia, the free encyclopedia," 2018, [Online; accessed 19-March-2018]. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Pinyin_table&oldid=818338658

[32] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *Proc. ASRU*, 2011.

[33] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. ICML*, 2015, pp. 448–456.

[34] A. Stolcke, "SRILM – an extensible language modeling toolkit," in *Proc. ICSLP*, 2002, pp. 901–904.