An Investigation of Deep Neural Networks for Multilingual Speech Recognition Training and Adaptation

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
{sibo.tong, phil.garner, bourland}@idiap.ch

Abstract

Different training and adaptation techniques for multilingual Automatic Speech Recognition (ASR) are explored in the context of hybrid systems, exploiting Deep Neural Networks (DNN) and Hidden Markov Models (HMM). In multilingual DNN training, the hidden layers (possibly extracting bottleneck features) are usually shared across languages, and the output layer can either model multiple sets of language-specific senones or one single universal IPA-based multilingual senone set. Both architectures are investigated, exploiting and comparing different language adaptive training (LAT) techniques originating from successful DNN-based speaker-adaptation. More specifically, speaker adaptive training methods such as Cluster Adaptive Training (CAT) and Learning Hidden Unit Contribution (LHUC) are considered. In addition, a language adaptive output architecture for IPA-based universal DNN is also studied and tested.

Experiments show that LAT improves the performance and adaptation on the top layer further improves the accuracy. By combining state-level minimum Bayes risk (sMBR) sequence training with LAT, we show that a language adaptively trained IPA-based universal DNN outperforms a monolingually sequence trained model.

Index Terms: multilingual ASR, DNN, adaptive training

1. Introduction

Recently, there has been increased interest in rapidly developing high performance automatic speech recognition (ASR) systems for a broad range of languages. Speech recognition systems built with multilingual deep neural networks (DNNs) have been shown to provide consistent advantages especially for low-resourced languages [1, 2, 3, 4]. 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 [3, 5]. The target of the multilingual DNN can be either the universal International Phonetic Alphabet (IPA) based multilingual senones [6] or a layer consisting of separate activations for each language [3, 7, 8].

The shared-hidden-layer multilingual DNN (SHL-MDNN) has been shown to outperform the monolingual DNN by a $3-5\%$ relative word error rate (WER) reduction [3]. In SHL-MDNN, hidden layers are trained to be shared across languages and only the output layer is language-specific. This method allows a shared, language-independent speech representation to be more robustly learned in the lower layers due to the increased training data presented. Meanwhile, DNN is also trained to model one single universal multilingual senone set. Phones of multiple languages are all explicitly mapped to a universal phone set (e.g., IPA) [6, 9]. Thus there is sufficient data to train the universal phones. However, it is usually found that the performance of the universal acoustic models is worse than the language-specific acoustic models unless the amount of training data for the target language is really small [9, 10]. Although the universal model may share data among various languages, mixture of data creates more variation especially for those identical IPA symbols shared among different languages.

Recently, various speaker adaptive training (SAT) approaches based on DNN have been proposed. Cluster adaptive training (CAT) was extended from Gaussian mixture model (GMM) to DNN [11]. It factorizes the hidden layers in DNN into a set of canonical weight matrices and speaker-dependent interpolation parameters. Similar approaches have been proposed independently in [12] and [13]. Researchers have also introduced learning hidden unit contribution (LHUC) to weight hidden unit activations in a speaker- or environment-dependent manner [14]. It was shown that LHUC results in consistent WER reductions for speaker and environment adaptation [15].

Both CAT and LHUC thus use shared parameters to learn a speaker-independent acoustic transformation and use speaker-dependent parameters to model speaker specificities. Inspired by this successful work in speaker adaptation, we hypothesize that language adaptive training (LAT) could model language specificity while keeping the advantage of data sharing across languages. It has been proved that SAT on the bottom layers is more effective than top layers [15, 16]. We also hypothesize that LAT is just the reverse of SAT. Top layers should be more language related. In this paper, CAT and LHUC are for the first time evaluated in the context of LAT. Both SHL-MDNN and universal IPA-based multilingual DNN are investigated. Under the latter framework, we also propose to directly train language-specific linear outputs in a LAT fashion. By combining state-level minimum Bayes risk (sMBR) sequence training, we show that the sequence-trained IPA-based universal model for the first time yields a better result than a monolingually sequence-trained model.

Our main goal is to improve multilingual acoustic modelling by applying LAT. Thus, bootstrapping a language with little or no data from another language by using a tandem approach [2, 4, 17, 18, 19], initialization from an existing network [20, 21] or model adaptation [22] is beyond the scope of this paper. In [23], the authors proposed to use a language feature vector, which is the bottleneck feature of a language recognition DNN, as auxiliary language information to enable LAT. They investigated LAT from feature level and the idea is similar to the tandem setup. To the best of our knowledge, there has been no work investigating LAT from the model level.

The remainder of this paper is organized as follows: in Section 2, we briefly discuss IPA-based multilingual DNN and the architecture of SHL-MDNN. In Section 3, adaptive training ap-

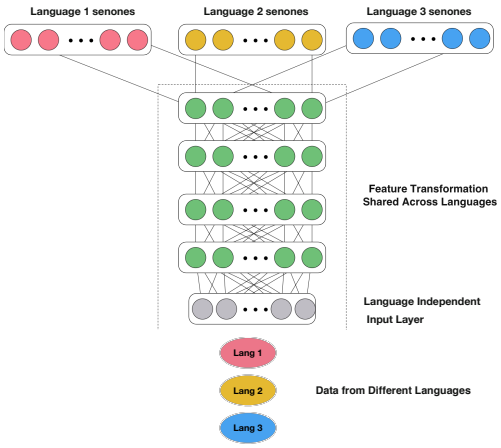

Figure 1: *Architecture of the SHL-MDNN.*

proaches are discussed. Experimental results and analysis are provided in Section 4. Finally, Section 5 concludes the paper.

2. Multilingual Deep Neural Network

2.1. Universal Phone Set Multilingual DNN

The main goal of multilingual acoustic modelling is to share the acoustic data across multiple languages to cover as much as possible the contextual variation in all languages being considered. One way to achieve such data sharing is to define a common phonetic alphabet across all languages. This common 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 tied-state targets for training the multilingual DNN are obtained by training the multilingual GMM-HMM systems and building multilingual decision trees to generate tied-state alignments. During decoding, language-specific language models and lexicons are used for each language separately. This architecture is subsequently denoted as MUL-IPA.

2.2. Shared-Hidden-Layer Multilingual DNN

In addition to modelling one single universal multilingual senone set, the output layer can also model multiple sets of language-specific targets and hidden layers are shared across languages [3, 7, 8]. In [7], the authors proposed to train DNNs on a sequence of target languages, progressively swapping the output layer with each new language. Whilst in [3] and [8], data from all languages is presented in an interleaved fashion during training, with the output layer swapped according to the target language being present. Here, only the architecture in [3] and [8] is discussed and we denote this shared-hidden-layer multilingual DNN as SHL-MDNN following [3].

Fig. 1 depicts the architecture used for multilingual ASR. The input and hidden layers are shared across all the languages. The output layers, however, are not shared. Instead, each language has its own output layer to estimate the posterior probabilities of the senones specific to that language. During recognition, language-specific prior and posterior probabilities are used for decoding.

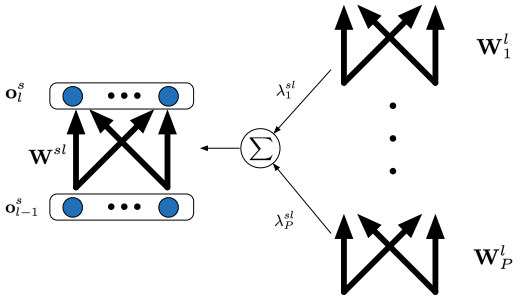

Figure 2: *Architecture of CAT-DNN for one layer.*

3. Adaptive Training Approaches

3.1. Cluster Adaptive Training

In [11], [12] and [13], multiple weight matrices or sub-networks are constructed to form the bases of a canonical parametric space. During adaptation, an interpolation vector, specific to a particular acoustic condition, is used to combine the multiple sub-networks into a single adapted DNN. We refer to this factorized DNN training as cluster adaptive training (CAT) following [11]. CAT was initially proposed for GMM-HMM acoustic models [24], and later extended to DNN by introducing multiple canonical weight matrices for a DNN layer as depicted in Fig. 2. For a specific speaker s, The adapted weight matrix between layer $l-1$ and layer l, $\mathbf{W}^{sl}$, is represented as an interpolation of the canonical DNN matrices:

$$\mathbf{W}^{sl} = \sum_{c=1}^{P} \lambda_c^{sl} \mathbf{W}_c^l \tag{1}$$

where $[\mathbf{W}_1^l, ..., \mathbf{W}_P^l]$ is the set of weight matrix bases between layer $l-1$ and layer l, P is the number of bases, $\boldsymbol{\lambda}^{sl}$ denotes the speaker dependent interpolation vector for layer l and speaker s. Therefore a general form of CAT-layer output can be obtained as following:

$$\mathbf{o}_l^s = \psi(\mathbf{x}_l^s), \tag{2}$$

$$\mathbf{x}_l^s = \sum_{c=1}^{P} \lambda_c^{sl} \mathbf{W}_c^l \mathbf{o}_{l-1}^s + \mathbf{B}^l \boldsymbol{\alpha}^{sl} \tag{3}$$

where $\boldsymbol{\alpha}^{sl}$ is the interpolation vector of bias for speaker s in layer l, $\mathbf{B}^l = [\mathbf{b}_1^l, ..., \mathbf{b}_P^l]$ is the concatenated bias bases and ψ is the hidden unit activation function. During training, all the parameters in a CAT-DNN, including the interpolation vectors and the canonical bases, are trained simultaneously using gradient descent algorithm.

As a speaker adaptive training approach, adaptation to test speakers must be done before testing. Therefore, canonical matrices are retained and interpolation vectors for test speakers are re-estimated using the hypothesis generated from first pass decoding of a speaker independent system. Thus, canonical components are speaker-independent and are shared across all speakers. The interpolation vectors model speaker specificity.

3.2. Learning Hidden Unit Contribution

Learning hidden unit contribution (LHUC) is a method that linearly re-combines hidden units in a speaker- or environment-dependent manner [14, 25]. Given adaptation data, LHUC rescales the contributions (amplitudes) of the hidden units in the model without actually modifying their feature receptors. A

speaker-dependent amplitude function is introduced to modify $\mathbf{o}_i^{sl}$, the hidden unit output of unit i in layer l for speaker s:

$$\mathbf{o}_i^{sl} = \xi(r_i^{sl}) \cdot \psi(\mathbf{w}_i^l \mathbf{o}^{l-1} + b_i^l) \quad (4)$$

$r_i^{sl} \in \mathbb{R}$ is an adaptable speaker-dependent parameter, re-parametrised by a function $\xi : \mathbb{R} \rightarrow \mathbb{R}^+$. A sigmoid function with range $(0, 2)$ is usually used. $\mathbf{w}_i^l$ is the i^{th} row of the corresponding weight matrix $\mathbf{W}^l \in \mathbb{R}^{d_{\mathbf{o}^l} \times d_{\mathbf{o}^{l-1}}}$ where $d_{\mathbf{o}^l}$ is the dimension of vector $\mathbf{o}^l$. b_i^l denotes the bias. ψ is the hidden unit activation function.

In SAT-LHUC, the hidden units are trained to capture both good average representations and speaker-specific representations, by estimating speaker-specific hidden unit amplitudes for each training speaker. Similarly, the speaker-specific scaling parameters are re-estimated for test speakers before testing.

3.3. Language Adaptive Training in Multilingual DNN

Inspired by the success in speaker adaptation, we hypothesize, that the language specificity can be also modelled by a set of language-dependent parameters (e.g., interpolation vectors in CAT and scaling parameters in LHUC). The focus of this work 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.

Compared with the limited amount of available data per speaker in SAT, much more language-specific data can be used in LAT to learn language specificities. Therefore, it is reasonable to use architectures consisting of more adaptable parameters for LAT. Besides the standard CAT and LHUC, a combination of these two is studied. Specifically, layer l is factorized into several canonical sub-layers. Each sub-layer is re-scaled in the LHUC manner and then combined together following

$$\mathbf{o}_l^s = \sum_{c=1}^{P} \xi(\mathbf{r}_c^{sl}) \odot \psi(\mathbf{W}_c^l \mathbf{o}_{l-1}^s + \mathbf{b}_c^l) \quad (5)$$

where $\mathbf{r}_c^{sl}$ is the language-dependent scaling parameters for language s and base c, and $\odot$ denotes a Hadamard product. Thus, more adaptable parameters are constructed which would better model language specificity. This architecture is subsequently denoted as LHUC-CAT.

Considering the sufficient amount of data in multilingual training, and following our assumption that top layers in DNN are more language-related, we propose to train MUL-IPA with language-specific output weights and biases. The output of the DNN for language s, $\mathbf{o}^{sL}$, is calculated as

$$\mathbf{o}^{sL} = \mathrm{softmax}(\mathbf{W}^{sL} \mathbf{o}^{L-1} + \mathbf{b}^{sL}) \quad (6)$$

where L is the output layer, $\mathbf{W}^{sL}$ and $\mathbf{b}^{sL}$ are the language-specific output weight and bias for language s. One advantage of doing so is that a language-specific prior, instead of a universal multilingual prior, can be used during decoding. In DNN-HMM framework, the scaled likelihood is computed as

$$\bar{p}(\mathbf{x}_t | q_t = i) = \frac{p(q_t = i | \mathbf{x}_t)}{p(q_t = i)} \quad (7)$$

where $\mathbf{x}_t$ and q_t are the acoustic observation and the corresponding state at time t, $p(q_t = i)$ is the prior probability of state i. If we substitute Eq. (6) into Eq. (7) and take the logarithm, it yields

$$\log \bar{p}(\mathbf{x}_t | q_t = i) = \mathbf{W}_i^{sL} \mathbf{o}_t^{L-1} + b_i^{sL} - \log p^s(q_t = i) - C_t \quad (8)$$

Table 1: *Statistics of the subset of GlobalPhone languages used in this work: the amounts of speech data for training and evaluation sets are in hours.*

Language	Vocab	PPL	#Phones	#Spkrs	Train	Eval
FR	65k	324	38	100	22.7	2.0
GE	38k	672	41	77	14.9	1.5
PO	62k	58	45	101	22.7	1.8
RU	293k	1310	48	115	21.1	2.4
SP	19k	154	40	100	17.6	1.7

where C_t is the logarithm of the denominator in the softmax function at time t. The incorporation of language-specific prior could explicitly make the bias b_i^{sL} more adapted to language s. This adaptive output architecture for MUL-IPA is further denoted as AO.

4. Experiments

4.1. GlobalPhone Database

In this section, we report experiments on GlobalPhone [26]. In this study, we used the French (FR), German (GE), Portuguese (PO), Russian (RU) and Spanish (SP) datasets from the GlobalPhone corpus. Each language has roughly 20 hours of speech for training and two hours for development and evaluation sets, from a total of about 100 speakers. Because of the limited space, results on evaluation sets are reported. The conclusions hold true on development sets. The trigram language models that we used are publicly available[1]. The detailed statistics for each of the languages is shown in Table 1.

4.2. Setup

We conducted two different sets of experiments by varying the output layer. In the first set of experiments, the SHL-MDNN architecture was used where each language has its corresponding softmax output. Monolingual GMM-HMM systems were trained to obtain the language specific tied-state alignments. The second set of experiments was conducted using the IPA-based universal triphone output. To create the universal phone set, we merged all the monolingual phones which share the same symbol in the IPA table. Multilingual GMM-HMM system was trained and multilingual decision tree was built to generate tied-state alignments.

The Kaldi speech recognition toolkit [27] was used to build all the systems. For each language, we built maximum-likelihood (ML) trained GMM-HMM systems, using 39-dimensional MFCC features (C0-C12, with delta and acceleration coefficients). The number of context-dependent triphone states for each language is 3100 with a total of 50K Gaussians (an average of roughly 16 Gaussians per state). The number of the IPA-based multilingual context-dependent triphone states is 8000 with a total of 150K Gaussians. All the DNNs used in the experiments had 6 hidden layers, each consisting of 2,000 sigmoidal units and were trained from 11 consecutive frames after restricted Bolzmann machine (RBM) pretraining.

4.3. Results

This section presents all the experimental results of our study. We first show the comparison between different multilingual architectures and baseline monolingual systems, which is listed

[1] `http://www.csl.uni-bremen.de/GlobalPhone/`

Table 2: *Comparison between monolingual baseline systems and multilingual training in WER(%).*

system	FR	GE	PO	RU	SP
monolingual baseline	23.2	16.6	19.9	28.8	9.0
MUL-IPA	23.3	18.5	19.3	30.4	9.8
SHL-MDNN	**23.0**	**15.6**	**18.9**	**28.3**	**8.6**

Table 3: *Compare SHL-MDNN and different LAT approaches in WER(%).*

system	FR	GE	PO	RU	SP
SHL-MDNN	23.0	15.6	18.9	28.3	8.6
+CAT-L1	22.9	15.7	19.1	28.3	8.8
+LHUC-CAT-L1	22.9	15.4	19.0	28.3	9.0
+LHUC-L1	22.8	15.6	18.9	28.5	9.0

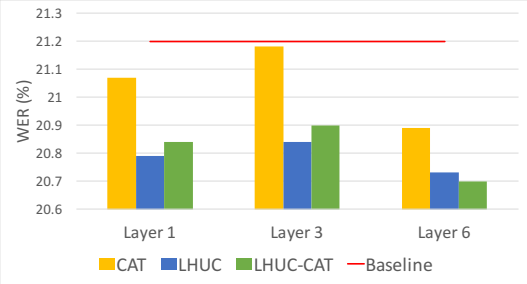

Figure 3: *Overall WER comparison of LAT approaches on different layers.*

in Table 2. It shows that SHL-MDNN achieves improvement over monolingual DNN baseline systems in all languages. Note that the use of multiple output layers in SHL-MDNN is similar to the concept of AO proposed for IPA-based universal DNN in Section 3.3. It can be viewed as training the output layer language-adaptively. The improvement demonstrates the benefit of language adaptive training. However, the multilingual training using universal phone set does not show much improvement and in most cases it is even worse. The result is consistent with previous work [3, 9, 10]. Although the IPA-based multilingual modelling enjoys richer data resources, it has a larger set of units to model as well. Moreover, identical IPA symbols across languages may not correspond to acoustic similarity.

4.3.1. Additional Language Adaptive Training in SHL-MDNN

Although SHL-MDNN has the adaptively trained output layer, we want to investigate that if additional LAT on previous layers could bring more gain. Standard CAT, LHUC-CAT and LHUC were conducted and only the first hidden layer was adapted. Three bases were used for CAT and LHUC-CAT in all the experiments. Large number of bases will make the amount of canonical parameters explode, which will easily lead to overfitting with insufficient training data. Table 3 shows that LAT on the first layer fails to give any further gain. As we hypothesize, in multilingual DNN, the layers close to the output are more language-related. The bottom hidden layers shared across languages are learned to project the acoustic feature to a universal phonetic space. Since the language specificity is mainly modeled in the language-dependent output layer, the LAT in the bottom layers becomes trivial.

4.3.2. Language Adaptive Training in MUL-IPA

Similarly, a series of experiments was conducted using IPA-based multilingual senones as the target. Following our hypothesis, language adaptive training was applied on different hidden layers. Fig. 3 describes the overall WER among all the five languages. It indicates that all the LAT approaches help improve the recognition performance. Adaptation on the last hidden layer further improves the accuracy. However, LAT on the middle layer doesn't perform as well as that on the bottom or top layer. LHUC and LHUC-CAT perform equally well and yield better results than standard CAT, which also demonstrates our hypothesis that more adaptation parameters would lead to more robust language specific modelling.

Table 4: *Compare LHUC, LHUC-CAT and MUL-IPA-AO in WER(%). The last row shows the relative improvement of MUL-IPA-AO over monolingual baseline.*

system	FR	GE	PO	RU	SP
monolingual baseline	23.2	16.6	19.9	28.8	9.0
MUL-IPA	23.3	18.5	19.3	30.4	9.8
+LHUC-L6	23.1	18.0	19.0	29.2	9.7
+LHUC-CAT-L6	22.9	18.2	19.1	**29.1**	9.5
+AO	**22.8**	**16.3**	**18.5**	29.2	**9.0**
relative improvement	1.7%	1.8%	7.0%	-1.4%	0%

Table 5: *Results of combining sMBR and LAT in WER(%). The last row shows the relative improvement of MUL-IPA-sMBR-AO over monolingual-sMBR.*

system	FR	GE	PO	RU	SP
monolingual-sMBR	22.6	15.4	18.4	**27.6**	8.3
MUL-IPA-sMBR	22.3	16.2	17.7	29.0	8.6
+AO	**21.9**	**14.8**	**17.1**	27.8	**7.9**
relative improvement	3.1%	3.9%	7.1%	-0.7%	4.8%

Given the fact that adaptation on the top layer with more adaptation parameters leads to better performance, MUL-IPA-AO is put into comparison. The results are listed in Table 4. As is expected, MUL-IPA-AO yields the best result and it performs equally well as the monolingual baseline.

Since MUL-IPA-AO outperforms other LAT techniques, we further investigate whether it is complementary with sequence training. In this work, state-level minimum Bayes risk (sMBR) sequence training was combined with AO language adaptive training. During sMBR training, alignments and lattices were generated using a language-specific prior, lexicon and language model. Table 5 shows that the sequence-trained LAT multilingual model outperforms sequence-trained monolingual systems in most languages. One interesting finding is that multilingual LAT yields more improvements after sequence training. Sequence-level LAT better captures both the universal acoustic representations and language specificities.

5. Acknowledgement

This paper was supported by the H2020 project SUMMA.

6. Conclusions

Several language adaptive training approaches were investigated under both SHL-MDNN and IPA-based multilingual DNN architectures. It was demonstrated that SAT approaches work also for language adaptation. Adaptation on top layers with more adaptation parameters further improves the accuracy. By combining AO and sMBR sequence training, the IPA-based universal network was shown for the first time to outperform state of the art monolingual DNN-based systems.

7. References

[1] N. T. Vu and T. Schultz, "Multilingual multilayer perceptron for rapid language adaptation between and across language families," in *Proceedings of Interspeech*, 2013.

[2] Z. Tüske, J. Pinto, D. Willett, and R. Schlüter, "Investigation on cross-and multilingual MLP features under matched and mismatched acoustical conditions," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013.

[3] 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 *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013.

[4] M. J. F. Gales, K. M. Knill, A. Ragni, and S. P. Rath, "Speech recognition and keyword spotting for low resource languages: Babel project research at CUED," *Spoken Language Technologies for Under-Resourced Languages*, 2014.

[5] K. Veselỳ, M. Karafiát, F. Grézl, M. Janda, and E. Egorova, "The language-independent bottleneck features," in *Proceedings of the IEEE Workshop on Spoken Language Technology*, 2012.

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

[7] A. Ghoshal, P. Swietojanski, and S. Renals, "Multilingual training of deep neural networks," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013.

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

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

[10] D. Chen and B. K.-W. Mak, "Multitask learning of deep neural networks for low-resource speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2015.

[11] T. Tan, Y. Qian, M. Yin, Y. Zhuang, and K. Yu, "Cluster adaptive training for deep neural network," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015.

[12] M. Delcroix, K. Kinoshita, T. Hori, and T. Nakatani, "Context adaptive deep neural networks for fast acoustic model adaptation," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015.

[13] C. Wu and M. J. F. Gales, "Multi-basis adaptive neural network for rapid adaptation in speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015.

[14] P. Swietojanski and S. Renals, "Learning hidden unit contributions for unsupervised speaker adaptation of neural network acoustic models," in *Proceedings of the IEEE Workshop on Spoken Language Technology*, 2014.

[15] P. Swietojanski, J. Li, and S. Renals, "Learning hidden unit contributions for unsupervised acoustic model adaptation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2016.

[16] T. Tan, Y. Qian, and K. Yu, "Cluster adaptive training for deep neural network based acoustic model," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2016.

[17] S. Thomas, S. Ganapathy, and H. Hermansky, "Cross-lingual and multi-stream posterior features for low resource LVCSR systems," in *Proceedings of Interspeech*, 2010.

[18] F. Grézl, M. Karafiát, and K. Veselỳ, "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*, 2014.

[19] P. Bell, J. Driesen, and S. Renals, "Cross-lingual adaptation with multi-task adaptive networks." in *Proceedings of Interspeech*, 2014.

[20] S. Thomas, S. Ganapathy, and H. Hermansky, "Multilingual MLP features for low-resource LVCSR systems," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2012.

[21] N. T. Vu, W. Breiter, F. Metze, and T. Schultz, "An investigation on initialization schemes for multilayer perceptron training using multilingual data and their effect on ASR performance," 2012.

[22] J. Zheng and A. Stolcke, "fMPE-MAP: improved discriminative adaptation for modeling new domains." in *Proceedings of Interspeech*.

[23] M. Müller, S. Stüker, and A. Waibel, "Language adaptive DNNs for improved low resource speech recognition," in *Proceedings of Interspeech*, 2016.

[24] M. J. F. Gales, "Cluster adaptive training of hidden markov models," *IEEE transactions on speech and audio processing*, 2000.

[25] P. Swietojanski and S. Renais, "SAT-LHUC: Speaker adaptive training for learning hidden unit contributions," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016.

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

[27] 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 *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, 2011.