TRANSCRIBING CODE-SWITCHED BILINGUAL LECTURES USING DEEP NEURAL NETWORKS WITH UNIT MERGING IN ACOUSTIC MODELING

Ching-Feng Yeh and Lin-Shan Lee

Graduate Institute of Communication Engineering, National Taiwan University

andrew.yeh.1987@gmail.com

ABSTRACT

This paper considers the transcription of the widely observed yet less investigated bilingual code-switched speech: the words or phrases of the guest language are inserted within the utterances of the host language, so the languages are switched back and forth within an utterance, and much less data are available for the guest language. Two approaches utilizing the deep neural network (DNN) were tested and analyzed, including using DNN bottleneck features in HMM/GMM (BF-HMM/GMM) and modeling context-dependent HMM senones by DNN (CD-DNN-HMM). In both cases the unit merging (and recovery) techniques in acoustic modeling were used to handle the data imbalance problem. Improved recognition accuracies were observed with unit merging (and recovery) for the two approaches under different conditions.

Index Terms—Code-switching, Deep Neural Networks, Bilingual, Speech Recognition, Unit Merging

1. INTRODUCTION

In the globalized world today, many people are capable of speaking more than one languages and actually using more than one languages in their daily lives. As a result, very often the speech signals observed in our daily lives include more than one languages. This is why people have tried to extend existing speech recognition technologies primarily developed for monolingual tasks to consider multilingual environments [1][2][3][4][5][6][7].

In general, bilingual speech can be classified into two categories [8]. In the first category, the speaker switches between languages from sentence to sentence. For example, in the sentences, "It's fine. 謝謝你(Thank you).", the first sentence is in English, while the second in Chinese. In the other category, the languages are switched from words to words within a sentence. For example, in the sentence, "這個equation很複雜(This equation is very complicated).", the word "equation" in the guest language of English is embedded in a sentence in the host language of Mandarin. The latter category is very common for speakers of non-English native languages (with English as the guest language and the non-English native language as the host), especially when they speak very good English and many English words (and phrases) are not yet properly translated into their native languages. So when they speak their native languages, very naturally some English words (and phrases) are inserted into the utterances and become a part of the utterances although in different languages. This paper is focused on this second category, and the word "code-switching" in this paper refers to this second category. This phenomenon of code-switching is very common in Asia [8][9][10][11][12][13], for example between English and Mandarin, Korean and Japanese. Taking Mandarin-English code-switching as an example, the English words embedded in the Mandarin utterances are very often composed of phonemes sounding like Mandarin phonemes rather than English phonemes, because the speaker is a native speaker of Mandarin and he is actually speaking a Mandarin sentence. This makes the task very special among speech recognition tasks.

A major concern here is that some acoustic units in different languages are very close, some are similar but slightly different, and some are unique for specific languages. In addition, the highly imbalanced data distribution for involved languages very often causes another different problem, i.e., much more host language data and very limited guest language data, since only words or phrases of the guest language are embedded in the utterances of the host language [8]. This not only makes acoustic modeling for the guest language difficult, but the recognizer naturally tends to take most speech signals as in the host language, because the language model almost always gives higher prior probabilities to the host language words, and the acoustic models for the host language units are better trained with more data and therefore better fitted to the signals. Many approaches in acoustic modeling such as merging acoustic units on different levels across languages in acoustic modeling have been proposed to handle these problems and shown to be very helpful [8][9][10][11][12][13][14].

Recently, the deep neural network (DNN) was shown to be able to improve speech recognition performance significantly [15][16]. The most popular form of DNN application in acoustic modeling is the context-dependent DNN-HMM (CD-DNN-HMM), in which each context-dependent HMM state or senone is modeled by a node of the output layer of the DNN. In addition, bottleneck features from DNN were also used in HMM/GMM (referred to as BF-HMM/GMM here) [17]. In this paper, both CD-DNN-HMM and BF-HMM/GMM are tested and analyzed with code-switched bilingual lectures, with unit merging technique for acoustic modeling used to handle the code-switching issues.

Below after the target corpora description in section 2, the DNN approaches, CD-DNN-HMM and BF-HMM/GMM are introduced in section 3, and the unit merging and recovery approach is presented in section 4. The experiments are discussed in section 5, with conclusion given in section 6.

2. TARGET CORPORA DESCRIPTION

The corpora used for the experiments reported here were the recorded lectures of two courses offered in National Taiwan University in form of spontaneous speech with highly imbalanced Mandarin-English code-switching characteristics (Mandarin as

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

host and English as guest languages) as mentioned above. In the recordings, utterances were produced spontaneously, with plenty of disfluencies such as pauses, hesitations, repairs and repetitions causing additional difficulties for recognition. In the corpora, most English words appeared to be domain-specific terms related to the topics of the lectures, while almost all function words were in Chinese, which implies the accuracy of English words is important. The detailed statistics of the corpora are listed in Table 1. We see the percentage of English (guest language) in the corpora is only 15-19%, or roughly 1.5 hours in training and 5 minutes in adaptation data. So the data imbalance problem is serious.

	Course 1	Course 2
Training Set (hr)	8.53	9.10
Adaptation Set (min)	28.59	30.27
Development Set (min)	129.62	126.81
Testing Set (min)	131.53	133.77
Mandarin / English (%)	80.5 / 19.5	84.8 / 15.2

TABLE 1. *Details for the Target Corpora.*

3. CD-DNN-HMM AND BF-HMM/GMM FOR CODE-SWITCHING ENVIRONMENT

Here we very briefly summarize the basic ideas of CD-DNN-HMM and BF-HMM/GMM used for the code-switching environment considered here.

3.1. Code-switched CD-DNN-HMM

Different from the conventional HMM/GMM, in CD-DNN-HMM each context-dependent HMM state or senone is modeled by a node of the output layer of the DNN instead of a GMM. During recognition, the likelihood for an observed feature with respect to a senone is given by the output value of the node modeling the senone by feeding the observed features as input to the DNN, as shown in Fig. 1(a). The features accepted at the input layer were MFCC in consecutive frames concatenated just as in common settings. Each hidden layer was pre-trained by a restricted Boltzmann machine (RBM) in an unsupervised manner with input being the output of the previous layer. The output layer consists of senones for all Mandarin and English triphones. In these triphones, the central phonemes include all Mandarin phonemes plus English phonemes, and all cross-language context dependency conditions are considered. Modeling senones for different languages by individual networks may lead to poor results due to the very limited data for the guest language in the code-switched corpora. We therefore adopt the concept of multilingual DNN recently proposed [18], in which all layers except the output layer were jointly trained by all data of both languages. The only difference is that here the output layer including senones for the two languages with all code-switching context dependencies are jointly obtained, while in the multilingual DNN [18] the triphone senones for each individual language were separately obtained.

As a classifier, deep neural network (DNN) has been proven to significantly outperform the conventional neural network with less hidden layers [15][16]. By pre-training with restricted Boltzmann machine (RBM), more hidden layers or deeper network structure can be trained sequentially. In acoustic modeling, context-dependent DNN-HMM (CD-DNN-HMM) is the most popular form of DNN application. Different from conventional HMM-GMM structure, each context-dependent HMM state or senone is modeled by a node of the output layer of the DNN instead of a GMM [15]. During the recognition process, the likelihood between an observed feature and a senone is given by the output value of the node modeling the senone by feeding the observation as input to the DNN.

3.2. Code-switched BF-HMM/GMM

This approach originated from the auto-encoding theory of neural networks. By transforming features sequentially from one layer to another, better features could be obtained. Conventionally, the output of the last hidden layer is extracted, with a size much smaller than other hidden layers for generating more compact features. However, it is difficult to decide the size of the last hidden layer, and it takes time to retrain the DNN whenever the size of bottleneck feature is changed. Therefore, fixing the size of the last hidden layer but using a dimension reduction procedure was proposed to extract the DNN-based bottleneck features [17]. The performance of such a system is reported to be comparable to CD-DNN-HMM with the same data, and such a structure is completely compatible to the conventional acoustic modeling framework including many powerful techniques such as MLLR, MAP, MPE and MME, as well as many approaches for cross-lingual acoustic modeling such as unit merging and recovery [8] as presented below. In this work, we use linear discriminant analysis (LDA) for the dimension reduction mentioned above as shown in the upper part of Fig. 1(b) to reduce the dimensionality from the size of last hidden layer (typically thousands) to the size for HMM/GMM (typically tens). These bottleneck features are then used to train the acoustic models for the two languages.

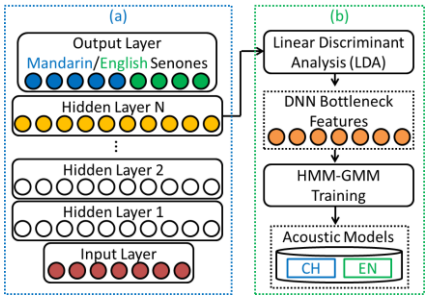

Figure 1. *(a) Code-switched CD-DNN-HMM*
(b) Bottleneck Feature for Code-switched BF-HMM/GMM

4. ACOUSTIC UNIT MERGING AND RECOVERY

Unit merging and recovery has been shown to be very useful in multilingual acoustic modeling [8]. For the code-switched speech considered here, the data imbalance problem can be properly handled in this way because of data sharing between units across languages. Here we consider this approach for both BF-HMM/GMM and CD-DNN-HMM.

4.1. Unit merging and recovery for BF-HMM/GMM

Unit merging and recovery for BF-HMM/GMM is very natural. Considering the difference between Mandarin and English on the linguistic nature and the phoneme level, unit merging on lower levels such as senone and Gaussian levels is performed. The com-

bined triphone model structure developed based on the combined phoneme set of Mandarin and English is shown in Fig. 2.

In Fig. 2, each context-dependent unit (e.g. triphone model, senone or Gaussian) is referred to as for Mandarin or English based on the language the central phoneme of the corresponding context-dependent unit belongs to. In Fig. 2(a), some senones (and Gaussians) are merged and shared (trained with shared data), and used as components for models for both languages. This is helpful since many acoustic segments in different languages produced by the same bilingual speaker may actually sound very similar. So the parameters of these senones can be better estimated if extra Mandarin data can be used together with the very limited English data. Similarly for unit merging on Gaussian level in Fig. 2(b), except the Gaussian is a finer unit so the merging can be more delicate.

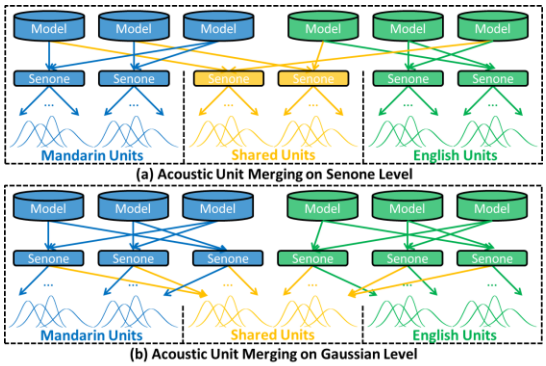

Figure 2. *Unit Merging on Senone and Gaussian Levels for BF-HMM/GMM*

The distance between Gaussians can be evaluated using the symmetrical KL divergence [8][9][12][13]. For distance between senones, we first model each senone by a single Gaussian, and the distance between senones is estimated by the symmetrical KL divergence between the corresponding Gaussians. In this way, a best candidate unit (Gaussian or senone) in the host language (Mandarin) with more data for each unit (Gaussian or senone) in guest language (English) with limited data can be obtained for merging based on the minimum distance criteria. But we also divided the combined phoneme set into four classes: plosive, affricate, voiced consonant and vowel, and unit merging are allowed only within the same class.

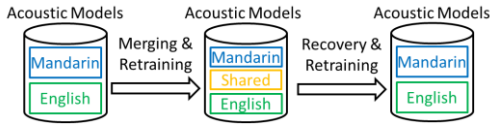

Figure 3. *The Concept of Unit Recovery After Merging*

The unit merging process mentioned above helps with the insufficient data for English units, but also limits the achievable likelihood. The merged units tend to be closer to the Mandarin units rather than the English units, because the former dominates the data. This also limits the likelihood for the corresponding signal segments given the merged units, especially for the English units. This is why we perform unit recovery in addition as in Fig. 3. The merged units were first recovered for both languages by copying all parameters from the merged units, and then an additional run of parameter re-training was applied. Here the influence of data sparseness problem is less serious since the initial parameters have been previously estimated by the shared data during unit merging.

4.2. Senone merging for CD-DNN-HMM

Unit merging for CD-DNN-HMM can be achieved on senone level by replacing the senones in the output layer by the corresponding merged set of senones (some nodes or senones are shared across languages) before DNN training as in Fig. 4. Note that in conventional HMM/GMM, each unit (senone or Gaussian) is to model the local distribution for the specific unit. In contrast, for CD-DNN-HMM here the parameters are shared by all target senones and trained by all training data. Therefore, the data sparseness and imbalance problem may not be as serious here.

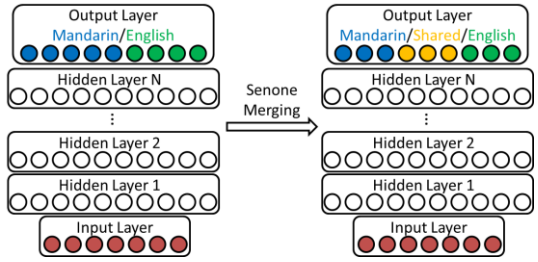

Figure 4. *Senone Merging for CD-DNN-HMM*

5. EXPERIMENTS

5.1. Experimental environment setup

The target corpora have been reported in section 2 and listed in Table 1. The 39 MFCC parameters were used as features for HMM/GMM baseline, and concatenated consecutive 9 frames of such MFCCs as features for CD-DNN-HMM. For HMM/GMM, triphone models obtained with state-clustering using decision tree were used. The senone number for HMM/GMM baseline were 2845 and 3172 for courses 1 and 2, respectively. The number of hidden layers is 4 for all DNN, with 2048 nodes in each hidden layer. For DNN bottleneck features, the feature dimension was reduced from 2048 to 40 by LDA.

The bilingual lexicon used included English words, Chinese words and all commonly used Chinese characters taken as mono-character Chinese words. For language modeling, the background model is trained with a combined corpus including Gigaword, Yahoo! News plus some target-domain related corpora such as master theses in related domains. We used Kneser-Ney trigram language model started with this background model and then adapted with the transcriptions of the training set for the target lectures here.

The way the recognition performance was evaluated followed the earlier work [8] and was very similar to the mixed error rate (MER) used for multilingual speech recognition evaluation later on [7]. When aligning recognition results with the reference transcriptions, insertions, deletions and substitutions were evaluated respectively for each language and summed up for overall evaluation. The basic unit for alignment and calculation was character for Mandarin and word for English. Individual performance for both Mandarin and English is reported. Since many English words are

the key terms in the lectures considered here, the recognition accuracy for English part alone is important.

5.2. Experimental results

The experimental results for both courses 1 and 2 are listed in Table 2. Rows (1)~(5) are for HMM/GMM with conventional MFCCs, with row (1) for the conventional HMM/GMM baseline, row (2) for merging on senone level and row (3) with unit recovery in addition, similarly for rows (4)(5) except on Gaussian level. Similarly for rows (6)~(10) for BF-HMM/GMM except MFCCs replaced by DNN bottleneck features. Rows (11)~(12) are for CD-DNN-HMM with row (12) for senone merging.

First consider rows (1)(6)(11) without unit merging, we can see that regardless of the model structure and the features used, the English accuracy were always significantly lower than Mandarin due to the data imbalance problem for the code-switching bilingual speech. By comparing rows (1) and (11), we see the CD-DNN-HMM greatly outperformed the HMM/GMM baseline using the same MFCC features without any unit merging or recovery. The BF-HMM/GMM with DNN bottleneck features in row (6) was somewhere in between in most cases.

Acoustic Models	Course 1			Course 2		
	Mandarin	English	Overall	Mandarin	English	Overall
(1) HMM/GMM (MFCCs)	75.62	71.63	75.32	83.62	61.87	81.99
(2) HMM/GMM (MFCCs) (MRG, Senone)	75.70	73.70	75.55	83.98	64.08	82.49
(3) HMM/GMM (MFCCs) (MRG+RCV, Senone)	76.11	76.95	76.17	84.34	69.04	83.19
(4) HMM/GMM (MFCCs) (MRG, Gaussian)	75.93	75.39	75.89	84.25	69.00	83.11
(5) HMM/GMM (MFCCs) (MRG+RCV, Gaussian)	76.04	77.50	76.15	84.38	71.94	83.45
(6) BF-HMM/GMM	78.56	74.78	78.27	84.32	56.99	82.27
(7) BF-HMM/GMM (MRG, Senone)	78.62	76.71	78.47	84.38	62.54	82.74
(8) BF-HMM/GMM (MRG+RCV, Senone)	78.47	77.03	78.36	84.30	67.92	83.07
(9) BF-HMM/GMM (MRG, Gaussian)	78.60	79.24	78.65	84.61	69.57	83.48
(10) BF-HMM/GMM (MRG+RCV, Gaussian)	78.72	80.06	78.82	84.70	71.92	83.74
(11) CD-DNN-HMM	79.14	78.86	79.11	85.32	69.04	84.10
(12) CD-DNN-HMM (MRG, Senone)	79.63	79.12	79.59	85.48	69.51	84.28

TABLE 2. *Experimental Results for HMM/GMM systems with Different Features and CD-DNN-HMM systems*

Now compare rows (2)(7)(12) to rows (1)(6)(11). First we see the English accuracy was significantly improved (rows (2)(7) vs. (1)(6)) by senone merging due to data sharing regardless of using MFCCs or DNN bottleneck features. However, for CD-DNN-HMM, the improvement brought by senone merging is relatively limited (rows (12) vs. (11)). A possible explanation for this is that for DNN all parameters are shared by all target classes, so the data sparseness issue is not as serious as in HMM/GMM (rows (1)(2)(6)(7)), for which the parameters are to model the local distributions for the specific senone. Furthermore, by checking rows (3)(8) in addition, we see that performing an extra pass of senone recovery did bring improvement for HMM/GMM (rows (3)(8) vs. (2)(7)) for either MFCCs or DNN bottleneck features. For Gaussian level merging and recovery (rows (4)(5)(9)(10)), we can see the trend is very similar to that for senone level merging and recovery, except with better performance due to the finer structure of the Gaussian level (rows (4)(5)(9)(10) vs. (2)(3)(7)(8)). Furthermore, the DNN bottleneck features always outperformed MFCCs in most cases (rows (7)(8)(9)(10) vs. (2)(3)(4)(5)).

Comparing BF-HMM/GMM with CD-DNN-HMM, we see the best BF-HMM/GMM (merging and recovery on Gaussian level in row (10)) achieved better English accuracy while the best CD-DNN-HMM (merging on senone level in row (12)) achieved better Mandarin and overall accuracies (rows (10) vs. (12)). This is important since Gaussian level merging is feasible only for BF-HMM/GMM and English accuracy is emphasized in this task. Therefore, for the two approaches of using DNN in modeling (CD-DNN-HMM) or for feature extraction (BF-HMM/GMM) considered in the work for the specific task, DNN in modeling gave better overall performance while DNN for feature extraction gave better English accuracy.

When performing unit merging, each English unit (senone or Gaussian) is given a Mandarin unit candidate with minimum distance; and such unit mapping pairs can be ranked according to the distance. So we can choose to merge only a selected percentage of English units with the corresponding Mandarin units starting with those pairs with minimum distances, but not all. The English accuracies obtained in this way for different percentages of English units merged (but not recovered) on senone and Gaussian levels for HMM/GMM (MFCCs) baseline, BF-HMM/GMM and CD-DNN-HMM for Course 1 is shown in Fig. 5. The best values on each curve correspond to the numbers in rows (2)(4)(7)(9)(12) in Table 2, respectively. We can see that the English accuracy in general increased when more English units were merged on either senone or Gaussian levels for HMM/GMM regardless of using MFCCs or DNN bottleneck features, although in some cases too high percentage of merging may not be good. However, for CD-DNN-HMM, the improvement achievable with senone merging was very limited (best at 30 %), and the performance degraded seriously when too many senones were merged. This is consistent with the previous assumption that parameters in DNN are jointly trained by all data, so do not benefit too much from data-sharing, and in fact the modeling ability of DNN was degraded when too few senones are present in the output layer.

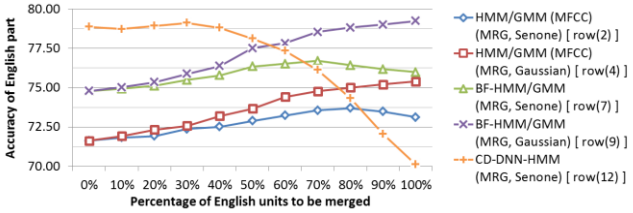

Figure 5. *English Accuracy with Different Merging Percentage of Course 1 for Different Systems*

6. CONCLUSION

Code-switching is a widely observed phenomenon in the daily languages of many people in the globalized world today. In this paper, we evaluated two approaches of transcribing speech of this type utilizing the deep neural network (DNN), including CD-DNN-HMM and BF-HMM/GMM, with the previously proposed unit merging (and recovery) approach applied in addition to handle the data imbalance issue. Experimental results indicated that unit merging (and recovery) techniques are helpful in such DNN configurations. Also, CD-DNN-HMM gave the best overall accuracy, while BF-HMM/GMM gave the best guest language accuracy.

12. REFERENCES

[1] Tanja Schultz, Alex Waibel, "Multilingual and Crosslingual Speech Recognition," *Proc. DARPA Workshop on Broadcast News Transcription and Understanding*, 1998.

[2] Tanja Schultz and Alex Waibel, "Language Independent and Language Adaptive Acoustic Modeling for Speech Recognition," Speech Communication, 2001.

[3] Hui Lin, Li Deng, Jasha Droppo, Dong Yu, and Alex Acero, "Learning Methods in Multilingual Speech Recognition," *NIPS*, 2008.

[4] L. Lamel , M. Adda-decker , J.L. Gauvain, "Issues in Large Vocabulary, Multilingual Speech Recognition," *Proc. Europ. Conf. on Speech Communication and Technology*, 185-188.

[5] Li Deng, "Integrated-multilingual Speech Recognition Using Universal Phonological Features in A Functional Speech Production Model", *ICASSP*, 1997.

[6] S.J. Young, M. Adda-Dekker, X. Aubert, C. Dugast, J.L. Gauvain, D.J. Kershaw, L. Lamel, D.A. Leeuwen, D. Pyea, A.J. Robinson, H.J.M. Steeneken, P.C. Woodland, "Multilingual Large Vocabulary Speech Recognition: the European SQALE Project," *Computer Speech & Language*, 1997.

[7] Ngoc Thang Vu, Dau-Cheng Lyu, Jochen Weiner, Dominic Telaar, Tim Schlippe, Fabian Blaicher, Eng-Siong Chng, Tanja Schultz, Haizhou Li, "A First Speech Recognition System for Mandarin-English Code-switch Conversational Speech," *ICASSP*, 2012.

[8] Ching-Feng Yeh, Chao-Yu Huang and Lin-Shan Lee, "Bilingual Acoustic Model Adaptation by Unit Merging on Different Levels and Cross-level Integration," *Interspeech*, 2011.

[9] Yanmin Qian, Daniel Povey and Jia Lu, "State-level Data Borrowing for Low-resource Speech Recognition based on Subspace GMMs," *Interspeech*, 2011.

[10] B. Mark and E. Barnard, "Phone Clustering Using Bhattacharyya Distance," *in Proc. Of ICSLP*, vol. 4, pp. 2005-2008, 1996.

[11] Anne-Katrin Kienappel, Dieter Geller, and Rolf Bippus, "Cross-language Transfer of Multilingual Phoneme Models," *Automatic Speech Recognition*, 2000.

[12] Houwei Cao, Tan Lee and P.C. Ching, "Cross-lingual Speaker Adaptation via Gaussian Component Mapping," *Interspeech*, 2010.

[13] Chung-Hsien Wu, Han-Ping Shen, and Yan-Ting Yang, "Phone Set Construction based on Context-sensitive Articulatory Attributes for Code-switching Speech Recognition," *ICASSP*, 2012.

[14] David Imseng, Herve Bourlard, Mathew Magimai.-Doss, John Dines, "Language Dependent Universal Phoneme Posterior Estimation for Mixed Language Speech Recognition," *ICASSP*, 2011.

[15] George Dahl, Dong Yu, Li Deng, and Alex Acero, "Context-Dependent Pre-trained Deep Neural Networks for Large Vocabulary Speech Recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, 2012.

[16] Geoffrey Hinton, Li Deng, Dong Yu, George Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara Sainath,, and Brian Kingsbury, "Deep Neural Networks for Acoustic Modeling in Speech Recognition," *IEEE Signal Processing Magazine*, 2012.

[17] Zhi-Jie Yan, Qiang Huo, and Jian Xu, "A Scalable Approach to Using DNN-Derived Features in GMM-HMM Based Acoustic Modeling for LVCSR," *Interspeech*, 2013

[18] Jui-Ting Huang, Jinyu Li, Dong Yu, Li Deng, and Yifan Gong,"Cross-language Knowledge Transfer using Multilingual Deep Neural Network with Shared Hidden Layers," *ICASSP*, 2013.