An Investigation of Adaptation Techniques for Building Acoustic Models for Hearing-impaired Children in a CAPT Application

Yingke Zhu, Brian Mak

Department of Computer Science and Engineering
The Hong Kong University of Science and Technology
{yzhuav,mak}@cse.ust.hk

Abstract

This paper describes our efforts in building an Android-based computer-assisted pronunciation training application for the local hearing-impaired (HI) children whose mother tongue is Cantonese. Since Cantonese HI children represent only a minority population in the world, the greatest challenge to the undertaking is the lack of their speech data and the difficulty in collecting sufficient speech data from them for acoustic modeling. We took the approach of building HI children acoustic model from normal-hearing (NH) adults model by adaptation using limited amount of adaptation data. Various feature-based and model-based adaptation methods were investigated. They include linear input networks (LIN) and its variants, Kullback-Leibler divergence (KLD) regularization, and learning hidden unit contributions (LHUC). We report results on phoneme recognition error rate (PER) as well as initial consonant recognition error rate (ICER) because the application currently focuses on the articulation of the initial consonants. The best results show that a combination of KLD and LIN-Nblock may reduce PER and ICER by a relative 11% and 16% respectively.

Index Terms: acoustic modeling, adaptation, deep neural network, computer-assisted pronunciation training

1. Introduction

Compared to traditional learning in classrooms, computer-assisted pronunciation training (CAPT) systems have the benefit of providing immediate feedbacks in a stress-free environment to the learners. CAPT can also be practiced anywhere and whenever a learner desires. In fact, many CAPT systems [1, 2] have been developed over the past decade for native speakers learning their first languages [3, 4] as well as non-native speakers [2, 5, 6] learning a second language. This paper investigates acoustic modeling for hearing-impaired (HI) children, whose mother tongue is Cantonese, for the development of a CAPT system that runs on Android-based mobile devices, such as a mobile smartphone or a tablet. Besides overall phoneme recognition accuracy, our CAPT application also focuses more on the articulation accuracy of initial consonants since it has been found that HI speakers produce more consonant errors than NH speakers [7, 8]. Nevertheless, this paper describes mainly our efforts on acoustic modeling of HI children's speech and not on the implementation of the application which is based on the system described in [9].

The target users of our CAPT application are a minority group of HI children of Hong Kong, aged between 6 to 12, and the mother tongue of their parents is Cantonese. It is well known that it is not easy to collect a large amount of speech data from young kids; collecting large amount of speech data from HI children is even harder. Even with the help of a local society for the deaf, we managed to collect only about one hour of Cantonese speech from 36 HI children. Hence, a major challenge we encountered is how to build an acoustic model for HI children with very limited amount of speech data. On the other hand, adult Cantonese speech corpora are already available, and further collection of speech data from NH adults is much easier. Consequently, our strategy is to first train an acoustic model using a large amount of NH adults speech data, and then investigate various adaptation techniques to transform the NH adults acoustic model to work for HI children using the limited amount of collected HI children speeches. We refer this as task adaptation. Compared with standard speaker adaptation, task adaptation does not estimate or store speaker-dependent parameters for any specific speaker. Instead, it uses the whole set of adaptation data to create a task-specific model from a model developed for another task by minimizing the mismatch between data from the two tasks. Two metrics are used to gauge the adaptation techniques: phoneme error rate (PER) and recognition error rate of the initial consonant (ICER) of Cantonese words.

2. DNN-HMM hybrid system

The hybrid deep neural network and hidden Markov model (DNN-HMM) has shown to give better performance in many automatic speech recognition (ASR) tasks over Gaussian mixture model (GMM) HMM in recent years. A DNN-HMM takes an observation $\mathbf{x}$ which usually consists of several contextual frames of acoustic features as input, and performs nonlinear transformations through L layers of perceptrons. For the lth hidden layer with $1 \le l \le L-1$, we have $h_i^l = \sigma(z_i^l) = \sigma\left((\mathbf{w}_i^l)^T \cdot \mathbf{v}^l + b_i^l\right)$, where b_i^l, z_i^l, and h_i^l are the bias, excitation and output of its ith neuron; $\mathbf{v}^l = \mathbf{h}^{l-1}$ is the input vector to the lth hidden layer; $\mathbf{w}_i^l$ is the weight vector associated with the ith neuron; $\sigma(x) = 1/\left(1+e^{-x}\right)$ is the sigmoid function. A softmax layer L is then added to the top and the DNN is trained to produce the posterior probabilities $p(y = s|\mathbf{x}) = p(y = s|\mathbf{v}^L)$ of HMM senones $s \in \{1, 2, \ldots, S\}$ (where S is the total number of senones). A DNN is typically trained by back-propagation by minimizing the following cross entropy,

$$\overline{D} = \frac{1}{N}\sum_{t=1}^{N} D(\mathbf{x}_t) = \frac{1}{N}\sum_{t=1}^{N}\sum_{y=1}^{S} \widetilde{p}(y|\mathbf{x}_t)\log p(y|\mathbf{x}_t)\ , \quad (1)$$

where N is the number of training samples and $\widetilde{p}(y|\mathbf{x}_t)$ is the target probability for the observation $\mathbf{x}_t$ at frame t. Usually hard alignment is used to produce the training labels, reducing $\widetilde{p}(y|\mathbf{x}_t)$ to $\delta(y = s_t)$, where δ is the Kronecker delta function and s_t is the label of the t-th frame.

3. DNN adaptation techniques

ASR often faces the problem that the test speech may not match well with what the speech recognizer was trained on. To alleviate such mismatch, various adaptation techniques have been developed for DNNs. They can be classified into feature-based methods [10, 11, 12, 13], addition of auxiliary features [14, 15, 16], and model-based adaptation [17, 18, 19].

3.1. KL divergence regularization

Kullback-Leibler divergence (KLD) regularization proposed by [17] adapts the model conservatively by requiring the state distribution estimated by the adapted model to be close to the distribution of the speaker-independent (SI) model. The constraint is realized by adding a KLD regularization term to the original cross-entropy optimization criterion in Eq. (1). Hence, we have

$$\hat{D} = (1-\rho)\overline{D} + \rho \cdot \frac{1}{N}\sum_{t=1}^{N}\sum_{y=1}^{S} p^{SI}(y|\mathbf{x}_t)\log p(y|\mathbf{x}_t)\ , \quad (2)$$

where $p^{SI}(y|\mathbf{x}_t)$ and $p(y|\mathbf{x}_t)$ are the posterior probabilities estimated by the SI model and the adapted model respectively, and ρ is the regularization parameter. If we define $\hat{p}(y|x_t)$ as

$$(1-\rho)\cdot\widetilde{p}(y|\mathbf{x}_t) + \rho \cdot p^{SI}(y|\mathbf{x}_t)\ , \quad (3)$$

then Eq. (2) may be rewritten as

$$\hat{D} = \frac{1}{N}\sum_{t=1}^{N}\sum_{y=1}^{S}\hat{p}(y|\mathbf{x}_t)\log p(y|\mathbf{x}_t). \quad (4)$$

Eq. (4) implies that KLD adaptation is equivalent to changing the target distribution from $\widetilde{p}(y|\mathbf{x}_t)$ to $\hat{p}(y|\mathbf{x}_t)$, which is a linear interpolation between the probabilities derived from the empirical distribution of the adaptation data and the distribution estimated from the SI model. As a result, KLD adaptation can be performed on DNNs using the conventional BP algorithm after modifying the target distribution from $\widetilde{p}(\cdot)$ to $\hat{p}(\cdot)$. The interpolation weight ρ has to be determined empirically using a separate development data set. When $\rho = 0$, KLD adaptation simply uses the SI model as the initial model and re-trains the model with the adaptation data. On the other hand, when $\rho = 1$, the method is reduced to unsupervised adaptation.

3.2. Linear input network

Linear transformation is a common approach to DNN adaptation. Methods based on linear transformation differ in the parameters that are being transformed. For example, the linear transformation can be applied to the input acoustic features (LIN), the outputs of hidden layers (LHN), or the inputs to the softmax layer (LON). Among them, LIN has been shown to give better performance [11, 13] and is investigated in this paper.

3.2.1. LIN

The linear input network (LIN) adaptation method assumes that the mismatch between training and testing conditions could be captured in the feature space, and the speaker-dependent (SD) features can be linearly transformed to match the SI features. Specifically, the LIN method augments an SI DNN with a linear layer to transform the input features. The augmented layer has the same dimension as the SI DNN's original input layer, and takes the identity activation function. In DNN-HMM, the input usually consists of several adjacent frames for a total of, say, N frames. Thus, if the dimension of the acoustic vector of each speech frame is D, then the LIN (including the biases) will have a dimension of $ND \times (ND+1)$. During adaptation, the LIN is first initialized to an identity matrix with zero biases, and standard BP algorithm is employed to update only the LIN parameters by minimizing the cross-entropy criterion while keeping the SI DNN parameters intact.

3.2.2. LIN-Nblock

Since one major challenge in our current task adaptation problem is the limited amount of adaptation data from local HI children, robust estimation of the large number of parameters in LIN is a concern. Thus, we also investigate a LIN variant, the LIN-Nblock (which is referred to as LINblk(I) in [11]) which reduces the number of parameters by applying a structural constraint to the network. Whereas LIN tries to capture both inter-frame and intra-frame relationship among the features, LIN-Nblock only captures intra-frame feature relationship: LIN-Nblock applies a different transformation to the features in each input frame. As a result, the number of network parameters is reduced by a factor of N to $ND(D+1)$.

The adaptation process of LIN-Nblock is exactly the same as that of LIN except that the LIN-Nblock is comprised of N smaller matrices of dimension $D\times(D+1)$. The N matrices are initialized to identity matrices with zero biases before training.

3.3. Learning hidden-unit contributions (LHUC)

LHUC [18, 19] is a model-based adaptation technique which learns speaker-specific scaling factors for each hidden unit. LHUC can be considered as a special case of the LHN method in which the linear transformation of the outputs from a hidden layer is a diagonal matrix. Specifically, for a test speaker, LHUC modifies the output of the ith unit in the lth hidden layer in an SI DNN by a SD scaling factor a_i^l as follows:

$$h_i^l = a_i^l \cdot \sigma(z_i^l)\ . \quad (5)$$

When the scaling factor a_i^l is set to 1.0, the SD model is equivalent to the SI model. The scaling factors given by Eq. (5) are unbounded and their training can be unstable. In [19], these scaling factors are trained incrementally in a procedure similar to DNN pre-training. In [18], the scaling factor is bounded by computing it from a sigmoid function with an amplitude of 2.0:

$$h_i^l = \sigma'(r_i^l)\cdot\sigma(z_i^l)\ , \quad (6)$$

where $\sigma'(r_i^l) = 2/(1+e^{-r_i^l})$. Since the number of hidden nodes in a DNN is much smaller than the number of weights in a layer, the number of parameters to be estimated in LHUC is much smaller than that in LIN, and is also slightly smaller to that in LIN-Nblock. Hence, LHUC may be preferred if the amount of adaptation data is relatively small.

4. Our CAPT system

Cantonese is the major Chinese dialect spoken in Hong Kong. It is also the most popular dialect among the group of 'Yue' dialects spoken in the Southern region of China. Each Chinese character is a syllable, and each syllable consists of an Initial and a Final. In Cantonese, there are 19 initial consonants if the Initial is not null, and 53 Finals. The Final is comprised of a vowel and an optional ending consonant, and there are 18 vowels and 6 ending consonants.

The current CAPT system was developed for the Hong Kong Society for the Deaf (HKSOD), which is a non-governmental organization. There are listening and speaking exercises of ∼400 Cantonese words. Special attention is given to the pronunciation of the 19 initial consonants. In the example below (Figure 1), a subject is required to tell the difference (in both listening or speaking exercises) between two very similar words that differ only in their initial consonants. Therefore, we employ two metrics in reporting the system performance: overall phoneme error rate (PER) and initial consonant error rate (ICER). Moreover, in the speaking exercise, we treat the pronunciation assessment as a phoneme verification problem using the PLASER technology described in [9] and report the CAPT assessment performance in terms of the equal error rate (EER).

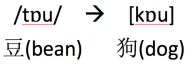

Figure 1: *A minimal-pair exercise example.*

Table 1: *Partitioning of NH adults and HI children data.*

Corpus	Data set	#Speakers	Gender (M/F)	#Utterances	Amount of Data (hrs)
NH adults	train	127	67/60	37,252	30.4
	dev	22	10/12	4,320	2.4
	test	17	9/8	3,985	2
	Total	166	86/80	45,557	35
HI children	adapt	18	13/5	963	0.51
	dev	9	6/3	412	0.22
	test	9	6/3	505	0.27
	Total	36	25/11	1,880	1.00

5. Experimental evaluations

The various adaptation methods were investigated to build acoustic models for hearing-impaired (HI) children from normal-hearing (NH) adults models for the CAPT system.

5.1. Speech corpora

Three Cantonese corpora were used for experiments: two corpora were collected from NH adults and one corpus from HI children in Hong Kong whose native language is Cantonese. Table 1 summarizes how the NH adults and HI children data were partitioned for training the baseline adults model and adapting to the children model.

5.1.1. Normal-hearing adults corpora

Two NH adults corpora were used. One was the CUSENT corpus [20]. It consists of 20 hours of speeches from 34 male and 34 female speakers, who read texts from 5100 different sentences extracted from several Hong Kong newspapers. It was collected inside a quiet recording room using a head-mounted microphone. The second corpus was collected by HKSOD from 52 female and 46 male speakers for a total of 13 hours of speech data. The corpus consists of Cantonese words spoken over the built-in microphone of a mobile phone or tablet. In summary, the two NH adults corpora together provide a total of ∼35 hours of speech from 166 speakers (86 females and 80 males).

5.1.2. Hearing impaired children corpus

The HI children corpus was also collected by HKSOD in the same way as the NH adults corpus. The corpus consists of only about 1 hour of speech from 11 HI girls and 25 HI boys aged between 6 and 12.

5.2. Acoustic modeling

For acoustic modeling, we used a DNN with 4 hidden layers of 2048 sigmoid units per layer. A softmax layer of 132 output units was stacked onto the DNN and classified an input frame to one of the 132 monophone HMM states. We chose to model monophones instead of other context-dependent units such as triphones, because the testing materials consist of only ∼400 isolated Cantonese words with limited triphone contexts. If triphones were modeled, many triphone states will have no adaptation data, resulting in poor performance. We had verified the decision by training a triphone system which indeed gave poor phoneme recognition performance on the HI children test data.

The DNN inputs are the standard 39-dimensional MFCC vectors with a context of 11 frames centered at the current frame. Phoneme recognition was done with a unigram phoneme language model. All experiments were conducted using the open-source Kaldi toolkit [21].

Table 2: *Performance of the DNN-HMM baseline system trained on NH adults data only.*

Test Set	Overall PER (%)	Consonant PER (%)	Vowel PER (%)	ICER (%)
NH adults	31.1	33.5	27.6	21.6
HI children	73.0	65.6	83.7	58.4

Table 3: *Adaptation performance on HI children test set.*

Adaptation	Overall PER (%)	Consonant PER (%)	Vowel PER (%)	ICER (%)
Baseline	73.0	65.6	83.7	58.4
KLD ($\rho = 0.5$)	65.8	57.7	77.4	49.5
LIN-Nblock	68.0	60.9	78.3	53.3
LIN	67.8	60.9	77.7	52.9
LIN + bias	67.5	60.3	77.8	52.5
LIN-Nblock + bias	66.4	60.1	75.4	52.5
LHUC	66.7	60.9	75.1	52.8
KLD+LHUC	65.4	58.9	74.8	51.1
KLD+LIN-Nblock+bias	65.1	57.5	76.0	49.5
KLD+LIN-Nblock	65.0	57.3	76.1	49.1

5.3. Phoneme recognition results

The baseline system was trained using only NH adults data, and its performances on both NH adults and HI children test data are shown in Table 2 in terms of both (overall) phoneme error rate (PER) and initial consonant error rate (ICER). It can be seen that the performance of the baseline model drops dramatically when testing on the mismatched HI children data set. It is also interesting to find that while the baseline NH adults model recognizes vowels better than consonants on NH test data, its performance is opposite on the HI children data. Moreover, among the consonants, initial consonants are better recognized.

The performance of the various adaptation methods is summarized in Table 3, which is further discussed below.

5.3.1. Adaptation results of KLD regularization

Figure 2 illustrates the KLD adaptation performance on the test data while the regularization parameter ρ was varied from 0.0

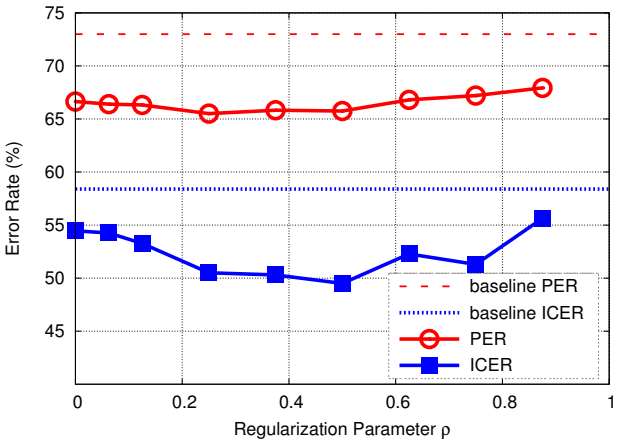

Figure 2: *Adaptation performance of KLD regularization*

to 1.0; the behavior is the same on the development data. Significant improvements are obtained regardless of the value of ρ. The finding is different from that in [17], which reported worse performance from KLD adaptation with small amount of adaptation data (5 or 10 utterances per speaker) and small ρ. The reason is that compared to speaker adaptation in [17], we are doing task adaptation with relatively sufficient adaptation data. This is also confirmed by the improvement obtained with $\rho = 0.0$ when even no regularization was applied. The best performance is obtained with $\rho = 0.5$, and PER (ICER) is reduced by an absolute 7.25% (8.90%).

5.3.2. *LIN adaptation results*

As shown in Table 3, LIN and its variants also significantly improve the performance of the baseline NH adults model on the HI children test data. The results show that it is important to adapt the biases as well: while LIN and LIN-Nblock have similar improvement of $\sim$5% absolute, the best result is obtained by LIN-Nblock+bias which reduces the PER (ICER) by an absolute 6.5% (5.9%).

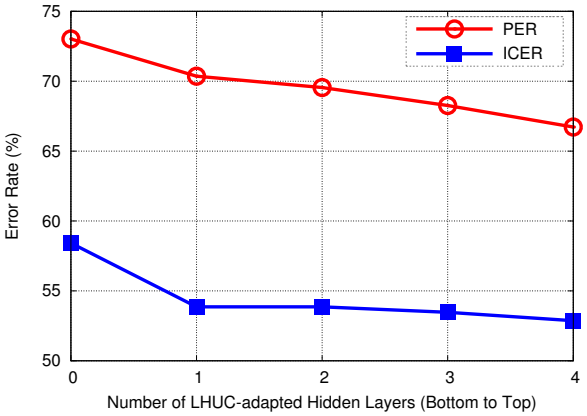

Figure 3: *LHUC with different number of hidden layers.*

5.3.3. *LHUC adaptation results*

From Table 3, the LHUC adaptation performance is very similar to that of LIN-Nblock+bias, and is just slightly worse than the latter. We further investigated the effect of LHUC adaptation with different number of hidden layers and the results are plotted in Figure 3. It is observed that the overall PER reduces almost linearly with the number of adapted hidden layers, starting from the bottom layer. On the other hand, most improvement in ICER reduction is obtained from adapting the bottom layer alone. Thus, it seems the bottom layer output is more relevant to the discrimination of the initial consonants.

5.3.4. *System combinations*

Finally we checked if the various adaptation methods are complementary by combining them through joint adaptation. From Table 3, we see that joint adaptation of KLD with LIN or LHUC may give slightly better results. The best results are obtained by combining KLD regularization and LIN-Nblock adaptation which reduces PER (ICER) by an absolute 8% (9.3%).

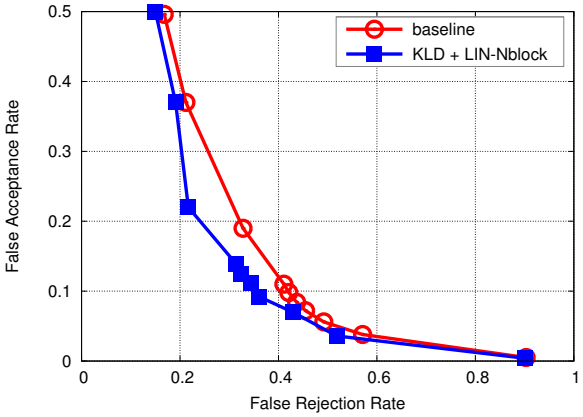

Figure 4: *CAPT performance: baseline vs. adapted system.*

5.4. CAPT of initial consonants

Our CAPT system focuses on learning the articulation of initial consonants, and gives a binary correct/incorrect response on their pronunciation produced by an HI child. From the DET curves in Figure 4, the EER is reduced by KLD+LIN-Nblock adaptation from 27% to 22%.

6. Conclusions

We investigated various speaker adaptation techniques for task adaptation: adapting an NH adults acoustic model to work for HI children in a mobile CAPT application. The major challenges are (1) the acoustic characteristics of HI children speech are very different from those of NH speakers in the original model; (2) the amount of adaptation data is very limited. We present adaptation results from KLD regularization, LHUC, LIN, and their combinations. Among the three methods, if they were applied alone, KLD regularization gave the best performance. Further improvement could be achieved from the joint adaptation of KLD and LIN-Nblock, reducing PER and ICER by a relative 11% and 16% respectively. The EER for the assessment of initial consonants was reduced by a relative 18.5%.

7. Acknowledgements

The work described in this paper was partially supported by grants from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project Nos. HKUST616513 and HKUST16206714), and partially by a grant from WeChat (Project No. 1516144-0).

8. References

[1] M. Eskenazi, "An overview of spoken language technology for education," *Speech Communication*, vol. 51, no. 10, pp. 832 – 844, 2009, Spoken Language Technology for Education — Spoken Language.

[2] C. Cucchiarini, A. Neri, and H. Strik, "Oral proficiency training in Dutch L2: The contribution of ASR-based corrective feedback," *Speech Communication*, vol. 51, no. 10, pp. 853 – 863, 2009, Spoken Language Technology for Education — Spoken Language.

[3] S. Wei, G. Hu, Y. Hu, and R.-H. Wang, "A new method for mispronunciation detection using support vector machine based on pronunciation space models," *Speech Communication*, vol. 51, no. 10, pp. 896 – 905, 2009, Spoken Language Technology for Education — Spoken Language.

[4] P. Price, J. Tepperman, M. Iseli, T. Duong, M. Black, S. Wang, C. K. Boscardin, M. Heritage, P. D. Pearson, S. Narayanan, and A. Alwan, "Assessment of emerging reading skills in young native speakers and language learners," *Speech Communication*, vol. 51, no. 10, pp. 968 – 984, 2009, Spoken Language Technology for Education — Spoken Language.

[5] Y. Ohkawa, M. Suzuki, H. Ogasawara, A. Ito, and S. Makino, "A speaker adaptation method for non-native speech using learners native utterances for computer-assisted language learning systems," *Speech Communication*, vol. 51, no. 10, pp. 875 – 882, 2009, Spoken Language Technology for Education — Spoken Language.

[6] K. Zechner, D. Higgins, X. Xi, and D. M. Williamson, "Automatic scoring of non-native spontaneous speech in tests of spoken English," *Speech Communication*, vol. 51, no. 10, pp. 883 – 895, 2009, Spoken Language Technology for Education — Spoken Language.

[7] M. J. Osberger and N. S. McGarr, "Speech production characteristics of the hearing impaired," *Speech and language: Advances in basic research and practice*, pp. 227–267, 1982.

[8] B. E. Walden and A. A. Montgomery, "Dimensions of consonant perception in normal and hearing-impaired listeners," *Journal of Speech and Hearing Research*, vol. 18, pp. 444 – 455, 1975.

[9] B. Mak, M. H. Siu, M. Ng, Y. C. Tam, Y. C. Chan, K. W. Chan, K. Y. Leung, S. Ho, F. H. Chong, J. Wong, and J. Lo, "PLASER: Pronunciation learning via automatic speech recognition," in *Proceedings of HLT-NAACL*, Edmonton, Canada, May 2003.

[10] Y. Xiao, Z. Zhang, S. Cai, J. Pan, and Y. Yan, "A initial attempt on task-specific adaptation for deep neural network-based large vocabulary continuous speech recognition," in *Proceedings of Interspeech*, 2012.

[11] B. Li and K. C. Sim, "Comparison of discriminative input and output transformations for speaker adaptation in the hybrid NN/HMM systems," in *Proceedings of Interspeech*, 2010.

[12] F. Seide, G. Li, X. Chen, and D. Yu, "Feature engineering in context-dependent deep neural networks for conversational speech transcription," in *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop*, 2011, pp. 24–29.

[13] K. Yao, D. Yu, F. Seide, H. Su, L. Deng, and Y. Gong, "Adaptation of context-dependent deep neural networks for automatic speech recognition," in *IEEE Workshop on Spoken Language Technology*, 2012, pp. 366–369.

[14] M. Karafiát, L. Burget, P. Matějka, O. Glembek, and J. Černockỳ, "iVector-based discriminative adaptation for automatic speech recognition," in *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop*. IEEE, 2011, pp. 152–157.

[15] G. Saon, H. Soltau, D. Nahamoo, and M. Picheny, "Speaker adaptation of neural network acoustic models using i-vectors." in *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop*, 2013, pp. 55–59.

[16] P. Karanasou, Y. Wang, M. J. Gales, and P. C. Woodland, "Adaptation of deep neural network acoustic models using factorised i-vectors." in *Proceedings of Interspeech*, 2014, pp. 2180–2184.

[17] D. Yu, K. Yao, H. Su, G. Li, and F. Seide, "KL-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2013, pp. 7893–7897.

[18] C. Zhang and P. Woodland, "DNN speaker adaptation using parameterised sigmoid and relu hidden activation functions," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2016, pp. 5300–5304.

[19] P. Swietojanski and S. Renals, "Learning hidden unit contributions for unsupervised speaker adaptation of neural network acoustic models," in *IEEE Workshop on Spoken Language Technology*, 2014, pp. 171–176.

[20] T. L. W. K. Lo and P. C. Ching, "Development of Cantonese spoken language corpora for speech applications," in *Proceedings of the International Symposium of Chinese Spoken Language Processing*, 1998, pp. 102–107.

[21] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," in *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop*, Dec. 2011.