Interspeech 2018
2-6 September 2018, Hyderabad

Development of the CUHK Dysarthric Speech Recognition System for the UASpeech Corpus

Jianwei Yu[1]*, *Xurong Xie*[2]*, *Shansong Liu*[1], *Shoukang Hu*[1], *Max.W.Y.LAM*[1], *Xixin Wu*[1], *Ka Ho Wong,*[1] *Xunying Liu*[1], *Helen Meng*[1]

[1]Department of Systems Engineering and Engineering Management,
[2]Department of Electric Engineering
The Chinese University of Hong Kong, Hong Kong SAR, China

{jwyu,ssliu,skhu,xyliu,hmmeng}@se.cuhk.edu.hk, xrxie@ee.cuhk.edu.hk

Abstract

Dysarthric speech recognition is a highly challenging task. The articulatory motor control problems associated with neuromotor conditions produce large mismatch against normal speech. In addition, such data is difficult to collect in large quantities. This paper presents the development of the Chinese University of Hong Kong automatic speech recognition (ASR) system for the Universal Access Speech (UASpeech) [1]. A range of deep neural network (DNN) acoustic models and their more advanced variants based on time delayed neural networks (TDNNs) and long short-term memory recurrent neural networks (LSTM-RNNs) were developed. Speaker adaptation by learning hidden unit contributions (LHUC) was used. A semi-supervised complementary auto-encoder system was further constructed to improve the bottleneck feature extraction. Two out-of-domain (OOD) ASR systems separately trained on broadcast news and switchboard data were cross domain adapted towards the UASpeech data and adopted in system combination. The final combined system gave an overall word accuracy of 69.4% on the 16-speaker test set.

Index Terms: dysarthric speech, speech recognition, cross-domain adaptation, system combination, auto-encoder

1. Introduction

Dysarthria is a type of speech disorder associated with neuromotor conditions. The underlying wide causes of dysarthria include neurological conditions such as Parkinson disease, amyotrophic lateral sclerosis, or cerebral palsy, and brain damages due to stroke or head injuries. Dysarthria results in a loss of controlling of speech articulators during production. This produces a large mismatch against normal speech. Hence, commercial automatic speech recognition systems constructed using normal speech are unsuitable to be directly employed [2, 3]. It also introduces a large variation among dysarthric speakers of different levels of severity. In addition, dysarthric speech data is also difficult to collect in large quantities for ASR system development.

For these reasons, there has been increasing research interest in recent years to develop dysarthric speech recognition systems of competitive performance [4, 5, 6, 7, 8, 9, 10, 11]. In order to make best use of the often limited amounts of dysarthric data, the suitable choice of acoustic models plays a crucial role. A great number of the earlier systems were based on hidden Markov models [6, 12, 8, 9, 7]. With the successful application of deep learning techniques in recognizing normal speech, a few previous works attempted to apply deep neural networks (DNNs) for dysarthric speech, e.g., to improve bottleneck feature extraction of tandem DNN-HMMs [9, 10]. Consistent performance improvements over HMM based acoustic models were reported [13, 9]. However, compared with state-of-the-art large vocabulary speech recognition systems, the use of more advanced forms of DNN architecture [14, 15] and their associated adaptation techniques [16] was limited. In order to reduce the mismatch of large amounts of out-of-domain (OOD) normal speech against dysarthric data, multi-level adaptive neural networks (MLANs) were proposed to transform normal speech data into in-domain like features and used to augment the limited dysarthric training materials [10].

This paper presents an initial development attempt at the Chinese University of Hong Kong to develop an automatic speech recognition (ASR) system for the Universal Access Speech (UASpeech) database. Improved speech segmentation was first performed by removing excessive silence at the start and end of dysarthric speech utterances. As there is a lack of coverage of all test set words in the UASpeech training data, it is difficult to use an end-to-end based modelling techniques represented by, for example, encoder-decoder recurrent neural networks with attention [17].

In our work, a range of deep neural network (DNN) acoustic models [18] with a deep and stacked architecture were first constructed. Their more advanced variants based on time delayed neural networks (TDNNs) [19] and long short-term memory recurrent neural networks (LSTM-RNNs) [20] were then developed to explore the potential benefit from longer range context modelling. Speaker adaptive training and adaptation by learning hidden unit contributions (LHUC) [16] was further used to handle inter-speaker variability. A semi-supervised complementary auto-encoder (CAE) [21] system was also applied to improve bottleneck feature extraction for stacked DNN systems.

In order to reduce the cost of system development using both in-domain dysarthric and large amounts of OOD normal speech, two OOD ASR systems separately trained on 144 hours of broadcast news and 300 hours of switchboard data were cross domain adapted to the UASpeech data. These were then used in system combination with in-domain data trained systems to leverage their diversity. The final combined system gave an overall word recognition accuracy of 69.4% on the 16 speaker test set. Compared with the previously published best accuracy of 65.2% on this task in [11], this corresponds to a total of 4.2%

Equal contribution was made between the first two authors. This research was supported by MSRA grant no. 6904412 and Chinese University of Hong Kong (CUHK) grant no. 4055065. The authors would like to thank Dr. Heidi Christensen and Prof. Mark Hasegawa-Johnson for insight discussion leading to this research.

10.21437/Interspeech.2018-1541

absolute (12.1% relative) word error rate reduction.

The rest of this paper is organized as follows. The task description for system development is presented in section 2. Section 3 describes GMM-HMM, tandem and hybrid systems. A wide range of advanced DNN acoustic models and adaptation techniques were introduced in section 4. Section 5 describes the construction of OOD cross adaptation system. The system combination approach is described in section 6. Experimental result of each system could be found in each corresponding section (3,4,5,6). The last section concludes and discusses possible future work.

2. Task Description

In-domain acoustic models were trained on UASpeech data, which is one of the largest databases available for English dysarthric speech. The UASpeech corpus is comprised of 16 dysarthric speakers and 13 normal speakers. Speakers were required to repeat 455 distinct words including 155 common words and 300 uncommon words. These words were distributed into three blocks. Each block contains the common words and one third of uncommon words. Approximately 126,000 sentences were involved in the training and test sets after we manually labeled some records missing word-level labels. We treated the block 1 and block 3 of all speakers as the training set, and the remaining block 2 consisting of only the dysarthric speakers as the test set. The UASpeech task is an isolated word recognition task. Following the decoding strategy proposed by [9]. A uniform language model was adopted with a word grammar network containing silence models at the start and end, and all possible test words in parallel.

All systems were decoded using the HTK [22] large vocabulary decoder HDecode. A modified HTK decoder which can utilize state posterior probabilities produced by DNNs using the Kaldi toolkit [23] was used. This decoder was used in all systems to produce both word lattices and subsequent confusion networks (CN) [24] for system combination.

3. In-domain Acoustic Modelling

3.1. GMM-HMM system

An initial phonetic decision tree clustered triphone GMM-HMM acoustic model with 2k tied states and 16 Gaussians per state was developed with the original audio segmentation. 39 dimensional PLP features augmented with their first and second order differentials were used. Following the re-align strategy in [9], this GMM-HMM (GH) system was then applied to re-align all the training and test data to remove excessive amounts of silence at the start and end of each utterance. After silence stripping, 0.2 seconds silence was reserved at both the start and the end of each utterance. The baseline GMM-HMM system was then re-trained with the silence stripped data (GHS). Using the re-segmented audio data in both system training and evaluation improved the word recognition accuracy by 6.9% absolute, as are shown in the first two lines of table 1.

3.2. Hybrid DNN-HMM system

In this step, state level alignment produced by the baseline system was prepared for training hybrid DNN (HD) system. Phonetic decision tree clustered tied states derived from the GMM-HMM system based on re-segmented audio data were treated as DNN output targets. The structure of the hybrid system contains 6 hidden layers with sigmoid activation function. Each hidden layer has 2000 nodes. Cross-entropy training of this hybrid system is initialized by a layer-wise discriminative pre-training using context-dependent (CD) states as the targets. The input of this system is a cascading of 9 consecutive 40-dimensional filter bank feature followed by its first-order difference vector. The performance of HD system is shown in the last line of table 1, which is 12.5% higher than the baseline GHS system.

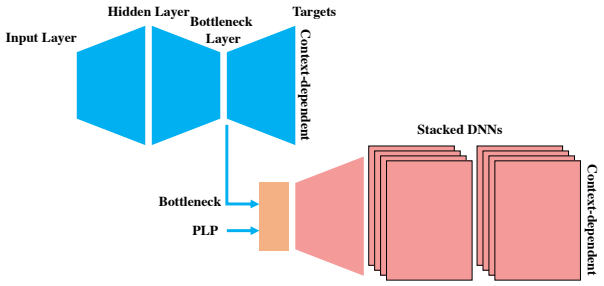

Figure 1: *An illustration of Stacked DNN system*

3.3. Tandem GMM-HMM system

The bottleneck (BN) features were obtained from a front-end DNN using the same architecture with the hybrid DNN except for the last but one layer, which contains 39 nodes and was used for producing BN features. The input features and the training criterion are the same as hybrid DNN. Tandem GMM-HMM systems used concatenated features, including BN features and 52-dimensional PLP+$\Delta + \Delta^2 + \Delta^3$ features. Cepstral mean normalization (CMN) and cepstral variance normalization (CVN) were applied. This is followed by feature transformations, including heteroscedastic linear discriminant analysis (HLDA) [25] for PLP features and global semi-tied transform [26] for BN features. A speaker independent model using Minimum Phone Error (MPE) [27] criterion was built first. A CMLLR [28] based MPE tandem SAT [29] system was also constructed. The results in table 1 indicates that, using tandem feature gives 3.1% accuracy improvement over GHS system and the adaptive training technology provides further 5% increment.

Table 1: *Performance of HMM, tandem and hybrid NN systems*

ID	SYSTEM DESCRIPTION	ACC
GH	GMM-HMM	46.7
GHS	GMM-HMM + Sllience stripping	53.6
TD	Tandem DNN	56.7
TD-SAT	Tandem DNN SAT	61.7
HD	Hybrid DNN	66.1

4. Stacked DNN systems

Figure 1 illustrates the basic architecture of stacked systems. Compared with normal systems, the stacked DNN systems use concatenated features as input and DNN architectures as state-level estimators, such kind of systems provide better modeling for longer range temporal context, whcih is beneficial for better classification of speech data. In this paper, all stacked systems perform lattice rescoring based on the lattices generated by hy-

brid DNN system. The result of this section is shown in table 2.

4.1. Stacked DNN-HMM system

Stacked hybrid system takes advantages of both hybrid system and tandem system. Such system uses concatenated features as inputs and, in the meantime, uses front-end DNN to estimate the probability of the output states. To utilize long time span information, the input of the hybrid DNN is a concatenation of 9 consecutive feature vectors. Similar to hybrid system, the DNN of stacked hybrid system is also initialized by layer-wise pretraining with CI targets. Fine-tuning is done using cross-entropy criterion with CD targets. The number of CD states is the same as in the tandem system. As the first two lines in table 2 showes, the stacked hybrid system (SHD) improves the frame accuracy around 1% over HD system.

4.2. TDNN and RNN system

For acoustic modeling, long context of features may provide better estimation of the target. Therefore, TDNN and RNN were employed to build the stacked hybrid systems. The TDNN has six hidden layers with 1000 nodes and input context of $([-2,+2],\{-1,2\},\{-3,3\},\{-3,3\},\{-7,2\},\{0\})$. Namely, a total of 29 frames of context between $[-16,12]$ were utilized for each time instant. However, determining the context configuration of TDNN may not be an easy job. RNN with Bi-directional LSTM (BLST) layers can automatically maintain the useful information from long context. Here the RNN is consist of four BLSTM layers with 500 cells on each direction. The network training procedures are the same as DNNs. The TDNN system and LSTM are labeled as K-TDNN and K-LSTM in table 2 respectively.

4.3. Complementary auto-encoder bottleneck features

Even for one speaker, the acoustic condition for each word or pronunciation may not be the same. This kind of variation could be uncertain and difficult to define. In this system, complementary auto-encoder (CAE) can be utilized to model the acoustic variability without explicitly using any prior knowledge of the acoustic condition. The structure of the CAE system is illustrated in figure 2 and figure 3. In the system, the CAE consisted of three LSTM sub networks: a auxiliary target encoder (ATE) with 256 cells, a reconstruction decoder (RD) with 256 cells, and a complementary feature encoder (CE) with 128 cells. For the CE, moving average of the LSTM outputs with a 21-frame context window is constricted by a 39 dimensional bottleneck layer on the top. For the training stage, as the left part in figure 2 shows, the auxiliary target encoder firstly used phoneme sequence alignment as input to model the target information. The optimal auxiliary target was found to be on word levels and subsequently used in all CAE system training. Then, as the right part of figure 2 shows, CE encoded the acoustic variability from the stacked acoustic features as complement to the target for reconstructing the acoustic features on the RD. Figure 3 illustrates the extaction and use of CAE bottleneck (BN) features in stacked DNN system. Only CE was used to extracted the BN features. Finally, for acoustic modeling, the DNN was trained by using the concatenation of the stacked acoustic features and the CAE BN features as input. The performance of CAE system is shown in the second section of table 2. The details of the CAE system could be found in [21], which was also submitted to 2018 interspeech.

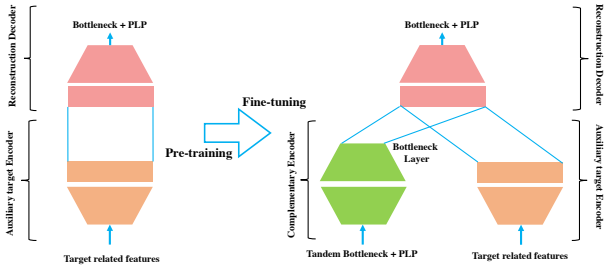

Figure 2: *An illustration of the training process of CAE system*

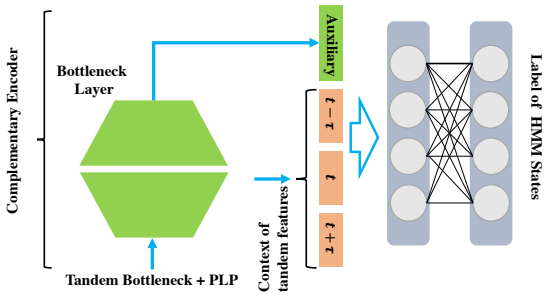

Figure 3: *The extraction and use of CAE bottleneck features in stacked DNN system*

4.4. Multi-task learning with monophone target

The results shown in table 2 suggest a weak correlation between validation data frame level accuracy and test set word recognition accuracy. One possible reason behind this may be due to the unreliable state alignment used in various DNN variant system construction. This in particular has a larger impact on the performance of systems incorporating longer contexts, e.g., the BLSTM system (K-LSTM in table 2) and the CAE (CAE in table 2) system. For both systems, higher validation data frame prediction accuracy was obtained over the baseline stacked DNN system (line 1 in table 2). However, no recognition performance improvement was obtained using either system. In order to address this issue, a secondary, simpler monophone labels based auxilary task was also used in a multi-task learning based framework [30] when developing the BLSTM and CAE systems (line K-LSTM-M and K-CAE-M in table 2). This was found to give consistent recognition performance improvements over the BLSTM and CAE systems.

4.5. LHUC speaker adaptation

For dysarthric speech acoustic modeling, acoustic level speaker variability affects the performance significantly. Since the SHD system holds the best performance in above stacked system, speaker adaptation technologies ware employed based on SHD system. In the system, LHUC scaling[16] was used for speaker adaptive training and adaptation on the stacked DNN-HMM system. For adaptation, the LHUC scaling parameters for each speaker were updated once per utterance. As the last two lines in table 2 shows, the LHUC systems provide improvement about 0.2% and 0.5% over the SHD system respectively.

Table 2: *Performance of stacked NN systems*

ID	SYSTEM	FRAME ACC	ACC
SHD	Stacked Hybrid DNN	41.5	67.1
K-TDNN	TDNN	43.0	66.0
K-LSTM	BLSTM	54.0	61.5
CAE	CAE	49.5	63
K-LSTM-M	BLSTM+ MTL	49.2	64.5
CAE-M	CAE+MTL	41.9	67.1
K-LHUC	LHUC	-	67.3
K-LHUC-SAT	LHUC+SAT	-	67.8

5. Cross domain adaptation

Two tandem OOD systems were used in this task. The acoustic model of Switchboard (SWBD) data was trained on 300-hour conversational telephone speech from Switchboard I, while the acoustic model of broadcast news (BNE) was trained using 144-hour BNE speech dataset from 1996&1997 Hub-4 English. These two OOD systems share the same architecture of the SAT system described in section 3.3. Cross adaptation [31] takes the recognition outputs from UASpeech stacked hybrid HD system as supervision. For both OOD systems, mean MLLR, diagonal covariance MLLR and feature space CMLLR transforms were used in cross adaptation to the HD in-domain system decoding outputs. For the SWBD OOD system (TDS), a total of 65 (1 silence, 64 speech) transforms were used (TDS-C in table 3) in each of the three kinds of adaptation methods mentioned above. For the BNE OOD system (TDB), a larger set of transforms (1 silence, 9 speech) were found to give the best performance and subsequently used in cross adaptation (TDB-C in table 3). Cross adaptation method can transform the mean and variance of the OOD-GMMs to be suitable for describing the distribution of the dysarthric speech, and thus provide a significant improved recognition performance. Compared with MLAN system[10], the cross adaption technique does not need to extract features from both in-domain and OOD data and then train a model using these features. The effectiveness of the cross-adaptation method is shown in Table 3. Word accuracy rates of un-adapted SWBD system and BNE system are 7.8% and 22.9%, respectively. The adapted systems provide recognition accuracy by 67.6% and 68.1%, separately.

Table 3: *Performance of Out-of-domain systems*

ID	SYSTEM	CROSS ADAPTATION	ACC
TDS	SWBD tandem DNN	NO	7.8
TDS-C		YES	67.6
TDB	BNE tandem DNN	NO	22.9
TDB-C		YES	68.1

6. System Combination

State-of-the art LVCSR systems often use system combination techniques [32, 33, 34]. The technique we used is hypothesis level combination [33, 34]. Confusion network combination (CNC) [24], which can exploit the consensus using voting or confidence measures among component systems, is one of such approaches. The condition of applying CNC technique is that the systems chosen to be combined should have similar performance but possess distinct characteristics. If one component system apparently performs poorer than other component systems or all the component systems have equivalent characteristics, the combined system will show no competitiveness in system combination. As figure 4 shows, in this paper, we selected in-domain HD, SHD, TDS-C and TDB-C systems trained by htk tools and CAE-M, K-LHUC-SAT systems trained using kaldi tools to do confusion network-based system combination. The second section in Table 4 shows the accuracy of the combined systems. The in-domain combined system used 4 different in-domain systems (ID-CNC) provides 68.5% word level accuracy. The system (ID+OOD-CNC) combined all the six systems provides the best performance by 69.4%. The first section in table 3 cites the recent representative system performance on the whole UASpeech test data. To the best of our knowledge, the best published result on the UASpeech data is the tandem speaker adaptation training system (Sheffield-15) reported in [11]. Over this system, our best performing ID+OOD-CNC system in this paper improves the absolute word accuracy by 4.2% (correspond to 12.1% relative word error rate reduction).

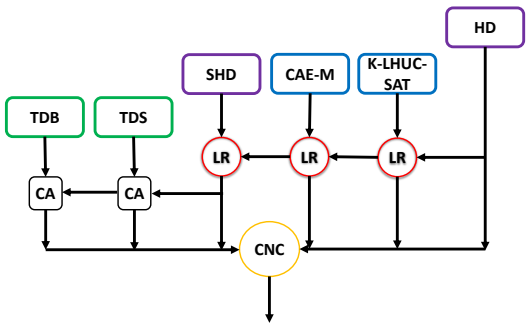

Figure 4: *System combination. LR, CA and CNC denotes lattice rescoring,cross-domain and confusion network respectively*

Table 4: *Performance of previous systems [11, 10] and our system combination*

ID	SYSTEMS	ACC
Sheffield-12[10]	Tandem ML-SI+SD-MAP	57.9
Sheffield-13[10]	MLAN ML-SI+SD-MAP	62.5
Sheffield-15[11]	Tandem-MLLR-MAP-SAT	65.2
ID-CNC	HD + SHD + K-LHUC-SAT + CAE-M	68.5
ID+OOD-CNC	HD + SHD + K-LHUC-SAT + CAE-M+ TDB-C + TDS-C	69.4

7. Conclusions

This paper presents the development of the Chinese University of Hong Kong an automatic speech recognition (ASR) system for the UASpeech task. In comparison to the previous state-of-the-art system[11], our best system presents a significant increment on the word accuracy rate, up to 4.2%. The future work will focus on dysarthric speech restoration technology and multi-channel speech recognition.

8. References

[1] H. Kim, M. Hasegawa-Johnson, A. Perlman, J. Gunderson, T. S. Huang, K. Watkin, and S. Frame, "Dysarthric speech database for universal access research," in *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[2] P. C. Doyle, H. A. Leeper, A.-L. Kotler, N. Thomas-Stonell *et al.*, "Dysarthric speech: A comparison of computerized speech recognition and listener intelligility," *Journal of rehabilitation research and development*, vol. 34, no. 3, p. 309, 1997.

[3] V. Young and A. Mihailidis, "Difficulties in automatic speech recognition of dysarthric speakers and implications for speech-based applications used by the elderly: A literature review," *Assistive Technology*, vol. 22, no. 2, pp. 99–112, 2010.

[4] R. Sriranjani, M. R. Reddy, and S. Umesh, "Improved acoustic modeling for automatic dysarthric speech recognition," in *Communications (NCC), 2015 Twenty First National Conference on*. IEEE, 2015, pp. 1–6.

[5] S. Gillespie, Y.-Y. Logan, E. Moore, J. Laures-Gore, S. Russell, and R. Patel, "Cross-database models for the classification of dysarthria presence," *Proc. Interspeech 2017*, pp. 3127–3131, 2017.

[6] J. Deller Jr, D. Hsu, and L. J. Ferrier, "On the use of hidden markov modelling for recognition of dysarthric speech," *Computer Methods and Programs in Biomedicine*, vol. 35, no. 2, pp. 125–139, 1991.

[7] M. Hasegawa-Johnson, J. Gunderson, A. Perlman, and T. Huang, "Hmm-based and svm-based recognition of the speech of talkers with spastic dysarthria," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, vol. 3. IEEE, 2006, pp. III–III.

[8] S. L. Christina, P. Vijayalakshmi, and T. Nagarajan, "Hmm-based speech recognition system for the dysarthric speech evaluation of articulatory subsystem," in *Recent Trends In Information Technology (ICRTIT), 2012 International Conference on*. IEEE, 2012, pp. 54–59.

[9] H. Christensen, S. Cunningham, C. Fox, P. Green, and T. Hain, "A comparative study of adaptive, automatic recognition of disordered speech," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[10] H. Christensen, M. Aniol, P. Bell, P. Green, T. Hain, S. King, and P. Swietojanski, "Combining in-domain and out-of-domain speech data for automatic recognition of disordered speech."

[11] S. Sehgal and S. Cunningham, "Model adaptation and adaptive training for the recognition of dysarthric speech," in *Proceedings of SLPAT 2015: 6th Workshop on Speech and Language Processing for Assistive Technologies*, 2015, pp. 65–71.

[12] M. J. Kim, J. Wang, and H. Kim, "Dysarthric speech recognition using kullback-leibler divergence-based hidden markov model." in *INTERSPEECH*, 2016, pp. 2671–2675.

[13] T. Nakashika, T. Yoshioka, T. Takiguchi, Y. Ariki, S. Duffner, and C. Garcia, "Dysarthric speech recognition using a convolutive bottleneck network," in *Signal Processing (ICSP), 2014 12th International Conference on*. IEEE, 2014, pp. 505–509.

[14] K. J. Han, S. Hahm, B.-H. Kim, J. Kim, and I. Lane, "Deep learning-based telephony speech recognition in the wild," in *Proc. Interspeech*, 2017, pp. 1323–1327.

[15] Y. Zhang, W. Chan, and N. Jaitly, "Very deep convolutional networks for end-to-end speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 4845–4849.

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

[17] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Advances in neural information processing systems*, 2015, pp. 577–585.

[18] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[19] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. J. Lang, "Phoneme recognition using time-delay neural networks," in *Readings in speech recognition*. Elsevier, 1990, pp. 393–404.

[20] A. Graves, A.-R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on*. IEEE, 2013, pp. 6645–6649.

[21] X. Xie, X. Liu, T. Lee, J. Yu, and L. Wang, "Complementry auto-encoder bottleneck features for asr acoustic modeling," in *submission to interspeech 2018*.

[22] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey *et al.*, "The htk book," *Cambridge university engineering department*, vol. 3, p. 175, 2002.

[23] 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 *IEEE 2011 workshop on automatic speech recognition and understanding*, no. EPFL-CONF-192584. IEEE Signal Processing Society, 2011.

[24] G. Evermann and P. Woodland, "Posterior probability decoding, confidence estimation and system combination," in *Proc. Speech Transcription Workshop*, vol. 27. Baltimore, 2000, pp. 78–81.

[25] N. Kumar and A. G. Andreou, "Heteroscedastic discriminant analysis and reduced rank hmms for improved speech recognition," *Speech communication*, vol. 26, no. 4, pp. 283–297, 1998.

[26] M. J. Gales, "Semi-tied covariance matrices for hidden markov models," *IEEE transactions on speech and audio processing*, vol. 7, no. 3, pp. 272–281, 1999.

[27] D. Povey and P. C. Woodland, "Minimum phone error and i-smoothing for improved discriminative training," in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, vol. 1. IEEE, 2002, pp. I–105.

[28] M. J. Gales, "Maximum likelihood linear transformations for hmm-based speech recognition," *Computer speech & language*, vol. 12, no. 2, pp. 75–98, 1998.

[29] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A compact model for speaker-adaptive training," in *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, vol. 2. IEEE, 1996, pp. 1137–1140.

[30] M. L. Seltzer and J. Droppo, "Multi-task learning in deep neural networks for improved phoneme recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 6965–6969.

[31] P. C. Woodland, C. J. Leggetter, J. Odell, V. Valtchev, and S. J. Young, "The 1994 htk large vocabulary speech recognition system," in *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, vol. 1. IEEE, 1995, pp. 73–76.

[32] X. Liu, M. J. Gales, and P. C. Woodland, "Language model cross adaptation for lvcsr system combination," *Computer Speech & Language*, vol. 27, no. 4, pp. 928–942, 2013.

[33] R. Schwartz, T. Colthurst, N. Duta, H. Gish, R. Iyer, C.-L. Kao, D. Liu, O. Kimball, J. Ma, J. Makhoul *et al.*, "Speech recognition in multiple languages and domains: the 2003 bbn/limsi ears system," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, vol. 3. IEEE, 2004, pp. iii–753.

[34] P. Woodland, H. Chan, G. Evermann, M. Gales, D. Kim, X. Liu, D. Mrva, K. Sim, L. Wang, K. Yu *et al.*, "Superears: Multi-site broadcast news system," in *Rich Transcription (RT-04F) Workshop*, 2004.