The NICT ASR System for IWSLT 2013

Chien-Lin Huang, Paul R. Dixon, Shigeki Matsuda, Youzheng Wu, Xugang Lu, Masahiro Saiko, Chiori Hori

Spoken Language Communication Laboratory
National Institute of Information and Communications Technology, Kyoto, Japan
chien-lin.huang@nict.go.jp

Abstract

This study presents the NICT automatic speech recognition (ASR) system submitted for the IWSLT 2013 ASR evaluation. We apply two types of acoustic features and three types of acoustic models to the NICT ASR system. Our system is comprised of six subsystems with different acoustic features and models. This study reports the individual results and fusion of systems and highlights the improvements made by our proposed methods that include the automatic segmentation of audio data, language model adaptation, speaker adaptive training of deep neural network models, and the NICT SprinTra decoder. Our experimental results indicated that our proposed methods offer good performance improvements on lecture speech recognition tasks. Our results denoted a 13.5% word error rate on the IWSLT 2013 ASR English test data set.

1. Introduction

The IWSLT 2013 Automatic Speech Recognition is an ongoing evaluation whose goal is to automatically transcribe TED [1] talks from audio to text [1]. TED is a nonprofit organization that promotes the dissemination of ideas. People can access TED talks on its website. Due to speech disfluency, emotional speech, noisy speech, different channels and speakers, the automatic transcription of TED talks is challenging. This year, the evaluation contains English and German speech materials as well as the automatic and mandatory segmentation of audio data. Since some talks are with non-native speakers, this year's evaluations are particularly challenging.

Automatic speech recognition has been widely applied in different kinds of applications [2]-[4]. To achieve better speech recognition performance, many techniques [5]-[9] have been proposed to address the problems in speech recognition. Cui et al. [5] presented a new semi-supervised learning method that exploits cross-view transfer learning for speech recognition through a committee machine that consists of multiple views learned from different acoustic features and randomized decision trees. A multi-objective scheme is generalized to a unified semi-supervised learning framework that can be interpreted into a variety of learning strategies under different weighting schemes. Huang et al. [6] proposed a joint analysis approach which simultaneously considers the vocal tract length normalization and the averaged temporal information of cepstral features. The Gaussian mixture model estimates conditional parameters in a data-driven manner. Chelba et al. [8] reviewed an approach to acoustic modeling that borrows from n-gram language modeling to increase both the amount of training data and the model size to approximately 100 times larger than the current sizes used in ASR. They experimented with contexts that span seven or more context-independent phones, and up to 620 mixture components per state. Hinton et al. [9] provided an overview of deep neural networks (DNNs) for acoustic modeling. Most speech recognition systems use hidden Markov models (HMMs) to deal with the temporal variability of speech and Gaussian mixture models (GMMs) to determine how well each state of each HMM fits a frame or a short window of frames of coefficients that represents the acoustic input. DNNs trained using new methods have outperformed GMMs on a variety of speech recognition benchmarks. In addition, Kaldi[2] [10] is an open-source toolkit of ASR written in C++. The core library support state-of-the-art techniques of modeling and feature extraction including DNN models, subspace Gaussian mixture models (SGMMs), decoder of finite-state transducers, and so on. In this study, we adopt Kaldi and NICT SprinTra for ASR system development and investigate speech recognition techniques on data analysis, feature extraction, acoustic and language models, and speech decoders.

The rest of this paper is organized as follows. Section 2 introduces data analysis and segmentation. We present the construction of combining multiple features and models for lecture speech recognition in Section 3. In Section 4, we describe our experiment setup, experiment results as well as a discussion of the results. Finally, we conclude this work in Section 5.

2. Data Analysis and Segmentation

We used three types of speech data to build acoustic models: the Wall Street Journal (WSJ), HUB4 English Broadcast news, and collected TED talks. We obtained WSJ and HUB4 from the Linguistic Data Consortium (LDC[3]). We crawled 760 TED talks from its online website published before December 31, 2010. The data are summarized in Table 1. WSJ is read speech. HUB4 is spontaneous broadcast news speech. TED is lecture style speech. Totally, we have about 300 hours of speech to build acoustic models with transcripts.

Both WSJ and HUB4 provide manual transcripts that can be directly used for acoustic model training. Text captions or subtitles of TED are provided with the speech recording, but speech segmentation and word alignment are not available. We used the SailAlign toolkit for speech segmentation and speech-text alignment [11]. SailAlign, which provides decoder-based segmentation with acoustic and language model adaptation, runs with HTK in which the acoustic model is trained by WSJ. Based on the segmentation results, the

[1] http://www.ted.com/

[2] http://kaldi.sourceforge.net/

[3] http://www.ldc.uppen.edu/

Table 1: Details of acoustic training data.

Name	Data	Type	Hours
TED	-	Lecture	167.8
HUB4	LDC97S44, LDC98S71	Broadcast	62.9
WSJ	LDC93S6B, LDC94S13B	Read	81.1

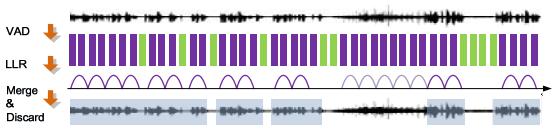

Figure 1: Illustration of the automatic segmentation of audio data.

speech-text alignment can be viewed as text-text alignment using dynamic programming to minimize the distance between reference and hypothesized texts.

In this study, the techniques of speaker clustering and automatic segmentation are applied to training and test audio data sets. First, speaker clustering has been widely adopted for clustering speech data based on speaker characteristics so that speaker-based cepstral mean normalization (CMN) and speaker adaptive training (SAT) [12] can be applied for better automatic speech recognition performance. TED talks are not always monologue; they might include interviews or conversations. We apply the vector space strategy to represent spoken utterances and conduct speaker clustering to group the spoken utterances into a number of speaker clusters in each talk. Experimental analysis is available in our earlier study [13].

Second, the length of a TED talk may range from 3 to 18 minutes with speech, laugh, applause, music, etc. For a good speech transcription, we apply the automatic segmentation processing to the audio data to remove non-speech segments (Fig. 1). Energy-based voice activity detection (VAD) is first used to detect the voice segments. Then the log-likelihood score with sliding windows is computed to detect speech/non-speech segments based on two GMMs trained using labeled speech/non-speech data. Finally, we merge the speech segments with a short interval between them and discard short segments. Merging and discard are based on a threshold of 170 ms.

3. System Description

3.1. Feature Extraction

Feature extraction is crucial to estimate numerical representation from speech samples. In this study, we extracted two sets of acoustic features to build acoustic models. The first set is Mel-frequency cepstral coefficient (MFCC), which is popular in speech recognition applications [14]. In MFCC feature extraction, 16-KHz speech input is coded with 13-demensional MFCCs with a 25ms window and a 10ms frame-shift. Each frame of the speech data is represented by a 39-dimensional feature vector that consists of 13 MFCCs with their deltas and double-deltas. Nine consecutive feature frames are spliced and projected to 40 dimensions using linear discriminant analysis (LDA) and maximum likelihood linear transformation (MLLT). The second acoustic feature is a perceptual linear predictive cepstrum (PLP) [15], which has the same LDA and MLLT. Both have 40 dimensions.

3.2. Subsystem Descriptions

The HMM models were with maximum 10,000 tied states and 160,000 Gaussian mixture components. We investigated three kinds of acoustic models: training of maximum mutual information (MMI), SGMM, and DNN.

Maximum Mutual Information Training: We maximized the auxiliary function in the M-steps of the EM estimation of the HMM parameters. The likelihood of the data given HMM is bound to increase when the value of the auxiliary function increases. In model space MMI training, we maximize a model's correctness by formulating an objective function and penalizing confusable models to the true model [16]. fMMI is feature space discriminative training with the same objective function as model space MMI training. After applying a global matrix, a high dimension feature vector is projected and added to the original features. In this study, we first apply speaker adaptive training on a triphone HMM system. Then discriminative training is applied with a feature space boosted fMMI followed by tree rebuilding and model space MMI training with indirect differential [17].

Subspace GMM Training: The subspace Gaussian mixture model is a compact representation of a large collection of a mixture of Gaussian models [18]-[20]. SGMM's basic idea is that all phonetic states share a common GMM structure, but the means and mixture weights vary in the total parameter space. Since most parameters are shared, we have more robust parameter estimation. We initialize the model by training a single GMM on all the speech classes that are pooled together. This is the universal background model (UBM). We use a total of 800 Gaussians in the UBM. Before SGMM training, SAT is used on the triphone system that is related to MLLR adaptation.

DNN Training: The deep neural networks are feed-forward, artificial neural networks that show more than one hidden layers between inputs and outputs [9, 21]. Recently, DNN has become a popular technique because it indicates good results for modeling speech acoustics. Many studies show that neural network based HMMs significantly outperform Gaussian mixture model based HMMs. In this study, starting from a DNN trained using cross-entropy, sequence discriminative training is then applied based on the state level minimum Bayesian risk criterion (sMBR) [22]. sMBR's objective function is explicitly designed to minimize the expected error corresponding to state labels, but we minimize the cross-entropy at the frame-level. We build DNNs by using five hidden layers and 2100 neurons (the structure is 300-2100-2100-2100-2100-2100-8070) (Fig. 2). DNN's input features are obtained by splicing together 15 frames (seven on each side of the current frame) and projected down to 300 dimensions using LDA. To better fit new speakers and environments, DNN acoustic models have been further adapted for specific talks using speaker adaptive training. Due to the limited amount of data in each talk, an efficient and effective method of speaker adaptive training of DNN models is only to adapt the middle layer (the third hidden layer). Speaker adaptation for DNN is difficult. In

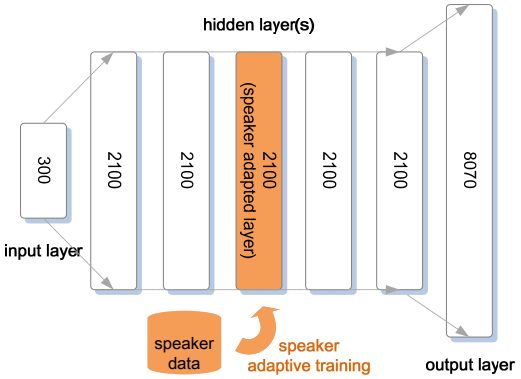

Figure 2: Illustration of speaker adaptive training of deep neural network models.

most studies, a speaker independent DNN (SI-DNN) is first trained. Then a speaker adaptation DNN is done by retraining the DNN parameters for different speakers either on all layers or some specific layers in the DNN [23, 24].

In this study, we propose a new speaker adaptive DNN training framework (SAT-DNN). We first assume that speaker specific processing is done in one layer in the DNN. All other layers are related to the speaker independent processing. Based on this assumption, we constructed a DNN with one layer as a speaker dependent layer, and the other layers are shared cross all speakers. In the DNN training, the parameters related to the speaker dependent layer are modified for each speaker while the parameters for all the shared layers are updated for all speakers. Explicitly specifying one layer as a speaker dependent layer in training focuses the training much more on speaker adaptation in DNN.

3.3. N-best ROVER

We considered a combination of two subsystems of MMI and SGMM in last year's evaluation [25]. This year, we built six subsystems using three types of acoustic models with two types of acoustic features. We integrated multiple complementary features and models for a better performance (Fig. 3). Several methods can be used to combine different recognition results. One popular approach is called recognizer output voting error reduction (ROVER) [26, 27]. Cui et al. [5] applied ROVER as a decision committee that votes for the labels of unlabeled data by cross validation. The combination can be carried out at the text output level as an n-best ROVER by output voting. We combine all decoding directories by composing the lattices. In this paper, different combination weights are applied to MMI, SGMM and DNN subsystems with 0.25, 0.25, and 0.5, respectively.

3.4. Language Model Adaptation and RNN Rescoring

We used the CMU pronouncing dictionary which has 133.3K words. We extended 39 phones of the dictionary to a 336 monophone set based on the accent and position information. The language models (LM) are modified Kneser-Ney smoothed 4-gram LMs trained on official data using the SRILM toolkit [28]. We used two different pruning 4-gram

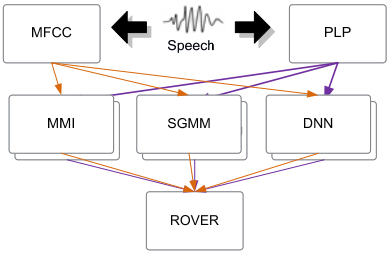

Figure 3: The combination of multiple systems for speech recognition using ROVER.

LMs in our experiments. The small 4-gram LM has 212 MB, and the big 4-gram LM has 9.6 GB and its perplexity is 115.4. Due to hardware and software limits, speech is decoded using the small 4-gram LM and rescored using the big 4-gram LM on MMI and SGMM subsystems. We use the first pass decoding results to adapt the language models that are used for second pass decoding [29]. In addition to conventional 4-gram LMs, we also applied a recurrent neural network (RNN) based LM [30] to rescore the n-best results. The sigmoidal recurrent network was built with the RNN-LM toolkit [31].

3.5. NICT SprinTra Decoder

In this paper, the ASR decoding process was based on weighted finite state transducers (WFSTs) [32], which integrate the acoustic and language models at the lattice level. We used the NICT SprinTra decoder, which has two major advantages [33]. First, the NICT SprinTra has smaller memory requirement and shows much faster decoding speed than the Kaldi decoder. Both NICT SprinTra and Kaldi use OpenFST [4] tools and library [34], but we use different structures to build the decoding graph. We also computed the so-called real-time (RT) factor. On the small 4-gram LM, SprinTra's decoding time was about 0.729×RT measured on an Intel Xeon CPU at 2.6GHz. This is better than the 1.023×RT of Kaldi and about a 30% difference in decoding time. Running on the big 4-gram LM, NICT SprinTra is ten times faster than Kaldi. Second, since the NICT SprinTra decoder decodes speech using the one pass method without language model rescoring, it is more accurate than decoding using language model rescoring. The word error rates vary from 0.1% to 0.3% between NICT SprinTra and Kali. This also denotes the gain using the big 4-gram LM decoding or rescoring.

4. Experiments

4.1. Training of Different Acoustic Data Sets

We experimented on the IWSLT 2013 ASR English test data set, which contained 4.5 hours of lecture speech, with 28 talks including 14 males and 14 females. There were at least eight non-native speakers (four males and four females) and one child. The effect of reverberation can be found in ten lectures. Non-native speech may be the main reason for the decrease of recognition accuracy. System performance was assessed using Word Error Rate (WER). Table 2 shows the results of

[4] http://www.openfst.org/

Table 2: MFCC-DNN subsystem results on training of different acoustic data sets.

Data	TED	TED+HUB4	TED+HUB4+WSJ
WER	16.9%	16.1%	15.7%

Table 3: Improvements by adding of different techniques on the IWSLT ASR 2013 English test data set.

System	WER	Reduction
MFCC-DNN baseline	15.7%	-
+ Six ROVER subsystems	14.8%	5.7%
+ Automatic segmentation	14.3%	3.4%
+ LM adaptation	14.1%	1.4%
+ SAT on DNN	13.5%	4.3%

the MFCC-DNN subsystem using different training data sets. All results were conducted on the entire lecture without any segmentation. Our experiments indicated that more data improved performance. Only the TED training was not good enough to recognize the TED speech of the IWSLT 2013 ASR English test data set. We achieved 15.7% WER using TED+HUB4+WSJ for the single MFCC-DNN subsystem, although HUB4 and WSJ were different types of speech from TED. We used the 15.7% WER result as the baseline in the following experiments.

4.2. Step-by-Step Improvements

Based on an MFCC-DNN baseline of 15.7% WER, Table 3 summarizes the step-by-step WER reductions with our proposed methods. First, the WER can be reduced to 14.8% using six ROVER subsystems. Due to error propagation and non-speech segments, the entire lecture decoding indicated poor performance. Adding an automatic segmentation technique reduced the WER from 14.8% to 14.5%, or 3.4% relative WER reduction. In addition, WER reductions of 1.4% and 4.3% were achieved for LM adaptation and SAT on DNN. Both adaptation methods were used to adjust models to better fit new speakers and environments. Our proposed methods offered more than 10% WER reduction on average. Our best result was 13.5% WER on the IWSLT 2013 ASR English test data set. Note that the application order of these techniques impacted the gain. For example, to get good speech transcriptions for adaptation, the LM adaptation technique is based on automatic segmentation results of audio data and six ROVER subsystems. In addition, the single MFCC-DNN subsystem indicated about 1.0% absolute WER reduction using the automatic segmentation of audio data, LM adaptation, and SAT on DNN.

4.3. Subsystems and ROVER Results

Table 4 shows the speech recognition evaluation of a combination of multiple features and models. Our experiments suggest the following observations. First, the MFCC and PLP features indicated similar results in most cases. Second, we evaluated the results of individual subsystems (1S). The DNN acoustic models significantly outperformed SGMM and MMI. Even the SGMM and MMI performances were much worse than DNN, and a combination of six subsystems (6S) further reduced the WER using ROVER. Compared with the 13.5% WER of six ROVER subsystems, the best result of the single MFCC-DNN system was 14.0% WER. The ROVER result was about 13.9% if we only considered MFCC features on three acoustic models (3S). Interestingly, we can obtain 13.9% WER using ROVERs of MFCC-DNN and PLP-DNN.

Table 4: Results (in %WER) of different subsystems and ROVER on the IWSLT ASR 2013 English test data set.

Feature	Model	1S	3S	6S
MFCC	MMI	19.7%	13.9%	13.5%
	SGMM	20.4%		
	DNN	14.0%		
PLP	MMI	20.2%	14.0%	
	SGMM	20.6%		
	DNN	14.1%		

4.4. Summary Results

Table 5 indicated the detailed results of each talk on the IWSLT ASR 2013 English test data set. Non-native speakers have the higher error rate in most cases. The WER is lower than 5% in the best condition but over 30% in the worst condition. Due to child voices and non-native speakers, talkid1699 denoted the worst recognition result. Furthermore, the IWSLT ASR 2011 (tst2011) and 2012 (tst2012) test data sets were used as progressive tests. There are eight and 11 talks in tst2011 and tst2012, respectively. Compared with this year's result of 13.5% WER, 7.7% and 8.2% WER results were achieved for tst2011 and tst2012 using our proposed approaches.

5. Conclusions

In this study, we propose a combination of multiple features and models for lecture speech recognition. We build six subsystems using three types of acoustic models (MMI, SGMM, and DNN) with two types of acoustic features (MFCC and PLP). The n-best ROVER denotes a good solution for a subsystem combination. We discover techniques of discriminative training and the adaptation of both acoustic and language models show great contributions to ASR. We propose the automatic segmentation of audio data, language model adaptation, speaker adaptive training of DNN models, and NICT SprinTra decoder. The results of our proposed methods demonstrate good performance improvement on the IWSLT 2013 ASR data set. There is still room for improvement when considering both good and a large amount of data.

6. References

[1] M. Cettolo, J. Niehues, S. Stüker, L. Bentivogli, and M. Federico, "Report on the 10th IWSLT Evaluation Campaign," *International Workshop on Spoken Language Translation (IWSLT)*, 2013.

Table 5: Detailed results of each talk on the IWSLT ASR 2013 English test data set.

Speaker	Gender	Dialect	# Second	# Sentence	# Word	% Corr	% Sub	% Del	% Ins	% Err	% S.Err
talkid1518	male	non-native, Deutsch	588	112	1247	89.4	8.5	2.1	1.7	12.3	53.6
talkid1520	male	south Asian	640	123	1416	88.7	8.1	3.2	2.1	13.4	52.0
talkid1532	female	native	569	113	1529	89.3	7.6	3.1	2.2	12.9	69.0
talkid1534	male	native	343	66	1165	81.6	9.2	9.2	0.9	19.3	84.8
talkid1539	male	African American	250	31	546	92.9	5.7	1.5	2.7	9.9	64.5
talkid1541	female	native	1141	247	2567	96.6	2.7	0.7	0.8	4.1	26.7
talkid1548	male	native	381	75	1083	88.8	7.5	3.7	1.9	13.1	62.7
talkid1553	female	native	359	79	804	96.3	2.6	1.1	0.6	4.4	31.6
talkid1592	female	native	251	39	670	98.4	1.2	0.4	0.7	2.4	30.8
talkid1600	female	African American	301	79	702	90.2	6.6	3.3	1.6	11.4	46.8
talkid1610	male	Italian	382	67	949	94.4	4.0	1.6	0.9	6.5	52.2
talkid1617	male	native	1009	172	2221	87.4	10.4	2.2	2.3	14.9	56.4
talkid1634	male	native	199	50	550	94.5	2.5	2.9	0.2	5.6	36.0
talkid1637	female	native	671	166	2021	96.0	2.5	1.5	0.4	4.4	34.9
talkid1640	male	native	544	108	1632	89.5	5.8	4.7	0.4	10.9	52.8
talkid1646	female	African American	508	80	927	83.7	13.7	2.6	2.5	18.8	71.3
talkid1647	female	native	558	106	1796	90.4	5.9	3.7	1.8	11.4	57.5
talkid1649	male	native	1047	247	3250	84.6	8.1	7.3	0.9	16.3	62.3
talkid1651	female	native	693	187	1739	95.3	2.7	2.0	0.7	5.4	29.9
talkid1654	female	native	940	244	2284	97.4	1.7	0.9	0.4	3.0	20.1
talkid1658	female	native	1077	209	2997	89.0	4.9	6.1	1.1	12.1	57.4
talkid1659	female	non-native, Egypt	562	112	1017	91.4	6.8	1.8	3.1	11.7	48.2
talkid1665	male	non-native, Deutsch	391	69	896	79.2	15.2	5.6	2.5	23.2	75.4
talkid1666	female	non-native, Afghanistan	554	150	1049	95.7	2.7	1.6	0.9	5.1	28.0
talkid1673	male	native	1072	235	3430	79.7	12.0	8.3	1.6	21.9	72.3
talkid1685	male	African American	614	139	1623	72.8	17.7	9.5	1.6	28.8	82.0
talkid1694	female	north Korean	715	115	1617	74.3	21.6	4.1	6.5	32.2	87.8
talkid1699	male	Kenya child	425	126	1021	64.7	26.0	9.3	2.4	37.6	78.6

[2] J.-R. Ding, C.-L. Huang, J.-K. Lin, J.-F. Yang, and C.-H. Wu, "Interactive Multimedia Mirror System Design," *IEEE Trans. Consumer Electronics*, vol. 54, no. 3, pp. 972–980, 2008.

[3] C.-L. Huang and C.-H. Wu, "Spoken Document Retrieval Using Multi-Level Knowledge and Semantic Verification," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2551–2560, 2007.

[4] C.-H. Wu, C.-H. Hsieh, and C.-L. Huang, "Speech Sentence Compression Based on Speech Segment Extraction and Concatenation," *IEEE Trans. Multimedia*, vol. 9, no. 2, pp. 434–438, 2007.

[5] X. Cui, J. Huang, and J.-T. Chien, "Multi-View and Multi-Objective Semi-Supervised Learning for HMM-Based Automatic Speech Recognition," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 20, no. 7, pp. 447–460, 2012.

[6] C.-L. Huang, C. Hori, H. Kashioka, and B. Ma, "Joint Analysis of Vocal Tract Length and Temporal Information for Robust Speech Recognition," in *Proc. of ICASSP*, 2013.

[7] C.-L. Huang and C.-H. Wu, "Generation of Phonetic Units for Mixed-Language Speech Recognition Based on Acoustic and Contextual Analysis," *IEEE Trans. Computers*, vol. 56, no. 9, pp. 1225–1233, 2007.

[8] C. Chelba, P. Xu, F. Pereira, and T. Richardson, "Large Scale Distributed Acoustic Modeling with Back-Off N-Grams," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 21, no. 6, pp. 1158–1169, 2013.

[9] G. Hinton, L. Deng, D. Yu, G.E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T.N. Sainath, and B. Kingsbury, "Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups," *IEEE Signal Processing Magazine*, pp. 82–97, 2012.

[10] 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 *Proc. of ASRU*, 2011.

[11] A. Katsamanis, M. Black, P. Georgiou, L. Goldstein and S. Narayanan, "SailAlign: Robust Long Speech-Text Alignment," in *Proc. of Workshop on New Tools and Methods for Very-Large Scale Phonetics Research*, 2011.

[12] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A Compact Model for Speaker-Adaptive Training," in *Proc. International Conference on Spoken Language Processing (ICSLP)*, vol. 2, pp. 1137–1140, 1996.

[13] C.-L. Huang, C. Hori, H. Kashioka, and B. Ma, "Speaker Clustering Using Vector Representation with Long-Term

Feature for Lecture Speech Recognition," in *Proc. of ICASSP*, 2013.
[14] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, N.J.: Prentice Hall, 1993.
[15] H. Hermansky, "Perceptual Linear Predictive (PLP) analysis of speech," *Journal of the Acoustic Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.
[16] D. Povey, D. Kanevsky, B. Kingsbury, B. Ramabhadran, G. Saon, and K. Visweswariah, "Boosted MMI for model and feature-space discriminative training," in *Proc. of ICASSP*, 2008.
[17] D. Povey, S. M. Chu, J. Pelecanos, and H. Soltau, "Approaches to Speech Recognition based on Speaker Recognition Techniques," chapter in forthcoming GALE book.
[18] X. Zhang, K. Demuynck, D.V. Compernolle, and H.V. Hamme, "Subspace-GMM Acoustic Models for Under-Resourced Languages: Feasibility Study," in *Proc. of SLTU*, 2012.
[19] L. Burget, M. Agarwal, P. Akyazi, K. Feng, A. Ghoshal, O. Glembek, N.K. Goel, M. Karafiat, A. Rastrow, R.C. Rose, P. Schwarz, and S. Thomas, "Subspace Gaussian Mixture Models for Speech Recognition," in *Proc. of ICASSP*, 2010.
[20] N.T. Vu, T. Schultz, and D. Povey, "Modeling gender dependency in the Subspace GMM framework," in *Proc. of ICASSP*, 2012.
[21] A. K. Jain and J. Mao, "Artificial Neural Networks: A Tutorial," *IEEE Computer*, vol. 29, no. 3, pp. 31–44, 1996.
[22] K. Vesely, A. Ghoshal, L. Burget, and D. Povey, "Sequence-discriminative training of deep neural networks," in *Proc. of Interspeech*, 2013.
[23] H. Liao, "Speaker Adaptation of context dependent deep neural networks," in *Proc. of ICASSP*, 2013.
[24] 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 *Proc. of ICASSP*, 2013.
[25] H. Yamamoto, Y. Wu, C.-L. Huang, X. Lu, P. R. Dixon, S. Matsuda, C. Hori, and H. Kashioka, "The NICT ASR System for IWSLT2012," in *Proc. of IWSLT*, 2012.
[26] J. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer Output Voting Error Reduction (ROVER)," in *Proc. IEEE Workshop Automatic Speech Recognition Understanding*, pp. 347–354, 1997.
[27] X. Cui, J. Xue, B. Xiang, and B. Zhou, "A study of bootstrapping with multiple acoustic features for improved automatic speech recognition," in *Proc. of Interspeech*, pp. 240–243, 2009.
[28] A. Stolcke, "SRILM - An Extensible Language Modeling Toolkit," in *Proc. of ICSLP*, vol. 2, pp. 901–904, 2002.
[29] Y. Wu, K. Abe, P.R. Dixon, C. Hori, and H. Kashioka, "Leveraging Social Annotation for Topic Language Model Adaptation," in *Proc. of Interspeech*, 2012.
[30] T. Mikolov, M. Karafiát, L. Burget, J. ˇCernocký, and S. Khudanpur, "Recurrent neural network based language model," in *Proc. of Interspeech*, 2010.
[31] T. Mikolov, S. Kombrink, A. Deoras, L. Burget, and J. Černocký, "RNNLM - Recurrent Neural Network Language Modeling Toolkit," in *Proc. of ASRU*, 2011.
[32] M. Mohri, F. Pereira, and M. Riley, "Weighted finite-state transducers in speech recognition," *Computer Speech and Language*, vol. 20, no. 1, pp. 69–88, 2002.
[33] P.R. Dixon, C. Hori, and H. Kashioka, "Development of the SprinTra WFST Speech Decoder," *NICT Research Journal*, pp 15-20, 2012
[34] C. Allauzen, M. Riley, J. Schalkwyk, W. Skut, and M. Mohri, "OpenFst: a general and efficient weighted finite-state transducer library," in *Proc. of ICAA*, pp. 11–23, 2007.