SWITCHING TO AND COMBINING OFFLINE-ADAPTED CLUSTER ACOUSTIC MODELS BASED ON UNSUPERVISED SEGMENT CLASSIFICATION

Jintao Jiang [1], *Hassan Sawaf* [2]

[1] Applications Technology (AppTek), 6867 Elm Street, Suite 300, McLean, VA 22101, USA
[2] eBay Inc., 2065 Hamilton Avenue, San Jose, CA 95125, USA
jjiang@apptek.com, hsawaf@ebay.com

Abstract

The performance of automatic speech recognition system degrades significantly when the incoming audio differs from training data. Maximum likelihood linear regression has been widely used for unsupervised adaptation, usually in a multiple-pass recognition process. Here we present a novel adaptation framework for which the offline, supervised, high-quality adaptation is applied to clustered channel/speaker conditions that are defined with automatic and manual clustering of the training data. Upon online recognition, each speech segment is classified into one of the training clusters in an unsupervised way, and the corresponding top acoustic models are used for recognition. Recognition lattice outputs are combined. Experiments are performed on the Wall Street Journal data, and a 37.5% relative reduction of Word Error Rate is reported. The proposed approach is also compared with a general speaker adaptive training approach.

Index Terms: MLLR, CMLLR, clustering, ROVER, SAT

1. Introduction

The performance of automatic speech recognition (ASR) system degrades significantly when the incoming audio differs from training data in terms of channel, speaker, and noise conditions. In a practical ASR system, maximum likelihood linear regression (MLLR) [1] and constrained maximum likelihood linear regression (CMLLR) [2] have been widely used for unsupervised adaptation, e.g., usually in a multiple-pass recognition process [3, 4, 5]. Adaptation in such systems usually relies on a limited amount of data, and the recognition outputs that are often unreliable when there is a mismatch between the training and test data [6, 7, 8]. Nevertheless, the supervised online adaption is usually not a convenient solution, especially when the channel and speaker condition changes frequently as in the broadcast news speech. In this paper, we present a novel adaptation framework that utilizes offline, supervised, high-quality adaptation with CMLLR followed by MLLR and online speech segment classification (or model selection). The first difference between the proposed approach and those in [6, 7, 8] is how the models are combined. In [6, 7, 8], the models are combined through the estimation/approximation of a new model from the offline models, while in this work, the offline models are intact and recognition results are then combined. A second difference is that in [6, 7, 8], to estimate new model parameters, either supervised training is needed or there are a lot of data for unsupervised estimation. While in this work, each single utterance can be classified in an unsupervised manner. A third difference is how the sub-models are selected. A novel approach is proposed here to use unsupervised acoustic scoring. In a related work [9], the adaptation was performed online by transforming the selected training speakers' acoustic data to the test data.

Experiments in this paper are performed using the RWTH® (Aachen University) ASR toolkit [10, 11]. Initially, a general acoustic model (AM) is trained using the Wall Street Journal (WSJ0 and WSJ1) training data. The training data are then automatically clustered based on Bayesian Information Criterion [12], after concatenating the speaker-based utterances. For each cluster or speaker, a supervised CMLLR is performed, followed by a supervised MLLR, resulting in a feature transformation matrix for CMLLR and Gaussian mean transformation matrices for MLLR. During testing, an incoming audio is recognized using the general acoustic model; the first-pass recognition also produces a word alignment; the alignment is applied to each sub-model to produce an accumulated acoustic score; these scores are compared, and the top-two models are selected. The second-pass recognition is then performed using the pre-computed CMLLR matrix and MLLR matrices that correspond to the selected automatically derived cluster. There are four adapted recognition processes (CMLLR vs CMLLR-MLLR and top best model vs second best model). Recognition lattice outputs from these four processes are then combined through the Recognizer Output Voting Error Reduction (ROVER) [13]. Finally, an evaluation on the WSJ0 test data shows a 37.5% relative reduction of Word Error Rate (WER). The approach is also compared with the general CMLLR-based SAT.

There are several advantages of this proposed approach. (1) Offline, supervised, high-quality adaptation is utilized for the online adaptation based on a feature-space segment classification. (2) The approach can be easily extended to include more clusters (channel/speaker characteristics) with a simple CMLLR and CMLLR-MLLR adaptation. (3) The online AM switching (i.e., applying CMLLR and MLLR matrices) can be realized without much additional computational overhead and in a supervised fashion. (4) The approach can lead to a potential diarization of channel/speaker conditions. (5) The proposed approach, in theory, can be extended to the deep neural networks models [14] without much effort, while it is not so straightforward in [6, 7, 8].

This paper is organized as follows. Section 2 describes the baseline ASR system and overall system design. Section 3 describes the automatic clustering of training data. CMLLR and MLLR adaptation is described in Section 4. Model selection or segment classification is presented in Section 5. The overall

978-1-4673-6997-8/15/$31.00 ©2015 IEEE

experiment setup including the ROVER system combination, test results are presented in Section 6. The paper is then concluded with discussion and future work in Section 7.

2. Baseline ASR system

The RWTH ASR toolkit [10, 11] is used for carrying out the experiments described in this paper. The overall system is illustrated in Figure 1.

2.1. Frontend feature extraction

A 16-kHz audio is pre-emphasized (α=1), segmented into frames (Hamming windowing, a frame shift of 10 ms, and a frame duration of 25 ms), and followed by a 512-point fast Fourier transform (FFT). The obtained spectral amplitude is then warped according to Mel scale and integrated within each of 20 triangular filters (a bandwidth of 268.258 Hz). A discrete cosine transform is applied to the logarithm of the Mel-scale filterbank outputs to obtain the 16-dimensional mel-frequency cepstral coefficients (MFCC). The mean normalization is carried out for each utterance. Across all the training data, a text-independent variance normalization matrix is computed.

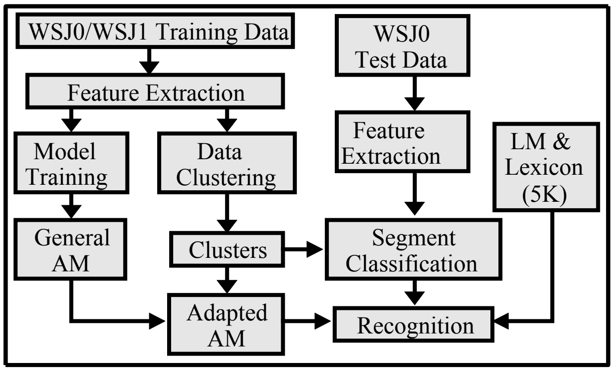

Figure 1: An overview of the proposed system.

For the initial linear segmentation, an energy term is computed for each frame right after the FFT and is then normalized across the utterance. For the initial monophone model training and decision tree training, the 16-dimensional MFCCs are first variance-normalized and then augmented by 16 first-order linear regression coefficients and one second-order linear regression coefficient, leading to a 33-dimension feature vector. Finally, a linear discriminant analysis (LDA) transform with 45-dimension output is trained with nine consecutive 16-dimension feature frames (left four frames), resulting in a 144 x 45 transformation matrix.

2.2. General AM training

The current system has 46 phonemes, including a silence, non-speech, and pause. Triphone AMs are trained based on the context of these phonemes. Across-word modeling is used to model contexts between words. Each model contains three states. Forward, skip, and lop transitions between states are set globally except that the silence model has its own transitions.

Linear segmentation is performed using the normalized energy, and the 33-dimension feature is used to produce the initial monophone Gaussian models. This initial model is then re-trained for 21 iterations. In the next step, the monophone model is further retrained for 20 iterations with a splitting step every three iterations. At the end of this training, the produced alignment is used to calculate simple Gaussian distributions for every triphone state. These distributions are then used to calculate the tying based on a top-down fashion with 144 phonetic questions, resulting in phonetic decision trees (classification and regression trees) with 1001 triphone states (including silence). The same alignment is then used to calculate the LDA transform matrix.

Triphone models are trained using the LDA features, the decision trees, and the monophone alignment. The training is iterated for 24 times with a splitting every three iterations. Thus, the maximum number of densities for each mixture is 256. The obtained general AM is later used for the CMLLR and MLLR adaptation.

3. Acoustic Data Clustering

Automatic clustering is performed using the Bayesian Information Criterion [12]. Each cluster is modeled as a Gaussian distribution:

$$N(\mu,\Sigma) \qquad (1)$$

To decide whether to join/divide into two clusters is based on maximum likelihood ratio with a penalty term:

$$\begin{aligned} R = N \cdot \log(|\Sigma|) - N_1 \cdot \log(|\Sigma_1|) - N_2 \cdot \log(|\Sigma_2|) \\ -\frac{1}{2}\left[d + \frac{1}{2} \cdot d \cdot (d+1)\right] \cdot \lambda \cdot \log(N) \end{aligned} \qquad (2)$$

where $\sum$, $\sum_1$, and $\sum_2$ are the sample covariance matrices from both clusters (N frames), the first cluster (N_1 frames), and the second cluster (N_2 frames), respectively. λ is the panel weight that is 4 in this study. d is the dimension of the LDA features that are used for computing the covariance matrices, which is 45. For this study, the number of clusters is arbitrarily set to 25 to have sufficient training data in each cluster (about 3 hours) and to have sufficient clusters to cover different acoustic conditions.

Utterances from each individual speaker are concatenated first. These concatenated files are then clustered. The automatic clustering of training data can be considered a multiple-view of the data. In the future, different views of the data can be added, for example, a view from the accent, a view from estimated vocal-tract length, etc. Later, the recognition lattice outputs from these views can be combined through ROVER.

4. Cluster-Based Adaptation

The general AM described in Section 2.2 is adapted to each cluster and speaker. Specifically, for each cluster/speaker, a CMLLR adaptation is applied with the triphone mixture models and LDA features. The monophone alignment with the one Gaussian density per mixture is used for the CMLLR adaptation. Following the CMLLR, a MLLR adaptation is applied to each cluster using the triphone mixture models and the CMLLR-adapted LDA features. The monophone alignment with the maximum 256 Gaussian densities per mixture is used for the MLLR adaptation. Both rotation matrices and offset vectors are estimated. The same decision trees from the general AM training is used for the MLLR class definition. The number of classes is set to 64. The minimum observation for each class is 50 seconds, and for silence, the number is 10 seconds.

5. Segment Classification

Segment classification is performed in an unsupervised fashion. For each segment, a first-pass recognition using the general acoustic model is applied. The process produces a word alignment, that is, each frame (f_n) will have an acoustic model mixture identify ($m_{n,j}$). With each CMLLR cluster model (T_k), each frame feature (f_n) is transformed with the corresponding CMLLR matrix (T_k), and its acoustic score is evaluated against the general acoustic model ($m_{n,j}$) and is then accumulated.

For the CMLLR, the models with the top-best and second-best scores are selected. The top selected model identities will be transferred to the CMLLR-MLLR case. That is, the model selection is only based on the CMLLR models. The top-two models are selected so as to approximate the new speaker condition. In [6, 7, 8], the projection of all cluster models is used to approximate the new speaker condition. The AMs adapted to such clusters are then used to recognize the test utterance. The MLLR transformation of the general AM can be performed offline, the CMLLR transformation is relatively fast, and the evaluation of feature again a single mixture is also fast. Therefore, the main computation complexity comes from the first-pass recognition.

6. Experiment Setup and Test Results

6.1. Experiment setup

Experiments are carried out on the Wall Street Journal corpus. Training is conducted on the WSJ0/WSJ1 SI-284-speaker corpus (37414 utterances and 283 actual speakers) and testing on the WSJ0 1992 development and evaluation sets. The test set includes eight speakers, 330 sentences, and 5,353 words in total. Both training and test data are adapted from the Kaldi tutorial processing [15].

LDA features (dimension of 45) are used for recognition. The general AM has 222,425 densities and a global covariance diagonal matrix. Each mixture has a maximum of 256 densities.

Recognition is performed with the single instruction multiple data (SIMD) diagonal maximum feature scoring. During recognition, the penalties for loop, forward, skip, and exit are 3, 0, infinity, and 0, respectively, while for silence, they are 0, 3, infinity, and 20, respectively. Word conditioned tree search is carried out for which the AM pruning threshold is set to 240 and the language model pruning threshold is set to 180. Language model 3-gram lookahead is used to improve both the speed and performance [10].

For each utterance, first-pass recognition is performed with the general acoustic model. The recognition result is used to select the top two models in the CMLLR and CMLLR-MLLR families, respectively. The same utterance is recognized with the four models separately. Each process produces a lattice with a pruning threshold of 380. The four lattices are processed to add word confidences using Frank Wessel's approach [16]. After that, the four lattices are combined with ROVER, and the 1-best result is obtained.

6.2. Language model and lexicon

In this work, the 5k trigram backoff language model (tcb05cnp) that comes with the WSJ data is used. During recognition, language model weight is set to 16.

Lexicon (one pronunciation per word) is obtained with a trainable grapheme-to-phoneme converter (Sequitur G2P) that was trained with an in-house 58k dictionary. In the produced lexicon, each word has one pronunciation variant.

6.3. Speaker adaptive training (SAT)

The proposed adaptation approach is also evaluated against a general two-pass SAT approach. After training of a general acoustic model, a CMLLR transformation matrix is estimated for each one of the 283 speakers. After that, each utterance's feature is projected using its corresponding CMLLR matrix, and then a SAT model is trained with new transformed features.

During decoding, the first-pass results are used to train an unsupervised CMLLR matrix to SAT model. The unsupervised scheme is applied to each one of the eight speakers in the test data. When the clustering approach in Section 3 is applied to the test data, the same eight speaker label data can be obtained. After the CMLLR estimation, a second pass recognition is applied.

6.4. Results

Recognition WERs in percentage are displayed in Table 1 for each of the eight speakers and across the speakers in the WSJ0 test data. The baseline (BL) refers to the recognition with the general AM without adaptation. Results for the four switched models, top-CMLLR-model-selection (C1), second-CMLLR-model-selection (C2), top-CMLLR-MLLR-model-selection (M1), and second-CMLLR-MLLR-model-selection (M2) and their ROVER (ROV) combination are displayed in the third, fourth, fifth, sixth, and seventh column, respectively. To compare the proposed approach with popular adaptation approach, the SAT adaptation results are also reported in the eighth column.

Table 1.*Word error rates in % on the WSJ0 test set using the proposed approach.*

Speaker	BL	C1	C2	M1	M2	ROV	SAT
440	2.61	2.61	2.30	2.30	2.76	1.69	1.84
441	7.22	5.49	6.75	4.08	5.65	3.61	4.87
442	5.68	4.85	5.40	5.12	6.93	4.29	4.43
443	1.82	1.52	1.67	2.43	2.43	1.22	0.91
444	5.68	3.51	4.32	3.24	5.54	2.70	3.38
445	4.49	3.99	3.00	3.49	3.49	2.83	3.99
446	2.62	2.47	2.91	2.77	2.47	2.04	1.89
447	5.33	5.48	5.33	5.18	5.78	3.81	5.18
Overall	4.45	3.74	3.98	3.59	4.43	2.78	3.31

The baseline results are comparable to those in [17, 18]. For both CMLLR and CMLLR-MLLR models, the top selected model outperforms the second selected model. Interestingly, the top selected CMLLR-MLLR model outperforms the top selected CMLLR model, which is predicted; while the second selected CMLLR model outperforms the second selected CMLLR-MLLR model. This is probably due to the fact that the top two models are selected only based on the CMLLR models. The overall WER reduction for the top selected CMLLR-MLLR model is

19.3% relative, compared to the baseline. When the four systems are combined, the performance is further improved. That is, all eight speakers show a performance improvement, and the overall WER reduction is 37.5% relative.

The CMLLR-based SAT also improved the recognition significantly over the baseline by 25.6%. However, the proposed approach can further improve over the SAT approach, for which the formal is performed on each utterance, while the latter is required to perform on each speaker. As a comparison, when the CMLLR adaptation for SAT is applied to each utterance, the WER is deteriorated to 7.08%.

6.5. Computational complexity

With the current setting, the first-pass decoding has a Real-Time-Factor (RTF) of around 0.5. The segment classification is completed with a RTF of around 0.1 and can be further speed-up with GPU computing. Compared to the SAT approach, the current approach will have three additional recognition processes, which add to the computational complexity. However, this computational complexity can be alleviated using the lattice acoustic modeling rescoring, which is much faster, with a RTF of around 0.1 for each rescoring. At the same, the recognition quality is not affected or at least not that much. In the current case, the WER after ROVER with lattice rescoring is still 2.78%.

7. Discussion and Conclusions

In this paper, we present a novel adaptation scheme that takes advantage of the offline, supervised, high-quality adaptation and the online speech signal classification. The relative overall WER reduction is 37.5% with the WSJ0 training and test set, compared to the baseline. The proposed approach also outperforms the popular SAT adaptation approach. In addition to its improvement in the WER, the proposed approach can perform on each individual utterance in an unsupervised fashion, while SAT requires enough of speech data from the same speaker.

When a new utterance comes in, it might not match the defined cluster/speaker very well. Therefore, any individual model cannot represent the new data at its optimum. In [6, 7, 8], a new model is estimated through the projection in “eigenvoice” space that fits the speech data perfectly. Therefore, such approach works best with the limited amount of data in a supervised fashion, while the proposed approach doesn’t use any pre-defined sample data. In the work, the combination of sub-models to match the new utterance is achieved through the ROVER in the result space. The advantage is that the well trained sub-models are being selected and run in its optimum status. Another advantage is that such a framework can be easily extended to the deep neural networks acoustic models [14].

The proposed approach is very flexible in that new channel and/or speaker data can be added continuously without interfering with previously trained models. For example, during broadcast news speech recognition, when a new anchor’s data are added, a new cluster can be added to the system with the corresponding CMLLR and MLLR training, without affecting any existing models. With the increased number of clusters, the computation cost for the speech segment classification increases as well. However, the model selection time can be ignored here as it is very fast.

Also, the proposed system can track which models are selected more frequently than others, and correspondingly, more efforts (e.g., fine-tuning) can be spent to enhance/update those more frequently selected models. The system can also integrate recognition lattice outputs from different views of the data. The types of views (e.g., accent, vocal tract length, etc.) can be expanded based on the acoustic analysis of the training data.

The current approach has some similarities to the discriminative training [19, 20] in that the adapted model for each cluster/speaker results in reduced phoneme error rate for the training data. Furthermore, each adapted model is more discriminative for its corresponding cluster. However, the advantage of the proposed approach is that each speech segment is recognized with an "optimized" adapted model but not a general discriminatively trained AM.

One immediate application of the proposed method is for broadcast news transcription for which the speech is segmented, each segment is classified, and the "best" models from multiple views are then used to recognize that segment.

Given that there are too many factors (accents, background noise, reverberation, etc.) that cause mismatches between training and test speech, it would be tremendous work to train a cluster-based adaptation for each condition. Nevertheless, any techniques that deal with robustness can be incorporated into this framework.

8. References

[1] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language,* vol. 9, no. 2, p. 171–185, 1995.

[2] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language,* vol. 12, no. 2, pp. 75-98, 1998.

[3] G. Saon, H. Soltau, U. Chaudhari, S. Chu, B. Kingsbury, H.-K. Kuo, L. Mangu and D. Povey, "THE IBM 2008 GALE Arabic speech transcription system," in *ICASSP*, 2010.

[4] L. Nguyen, B. Xiang, M. Afify, S. Abdou, S. Matsoukas, R. Schwartz and J. Makhoul, "The BBN RT04 English broadcast news transcription system," in *Interspeech*, 2005.

[5] M. Graciarena, H. Franco, J. Zheng, D. Vergyri and A. Stolcke, "Voicing feature integration in SRI's decipher LVCSR system," in *ICASSP*, 2004.

[6] K. T. Chen, W. W. Liau, H. M. Wang and L. S. Lee, "Fast speaker adaptation using eigenspace-based maximum likelihood linear regression," in *Interspeech*, 2000.

[7] M. J. F. Gales, "Cluster adaptive training of hidden Markov models," vol. 8, no. 4, pp. 417-428, 2000.

[8] R. Kuhn, P. Nguyen, J. C. Junqua, L. Goldwasser, N. Niedzielski, S. Fincke, K. Field and M. Contolini, "Eigenvoices for speaker adaptation," in *ICSLP*, 1998.

[9] M. Padmanabhan, L. Bahl, D. Nahamoo and M. Picheny, "Speaker clustering and transformation for speaker adaptation in speech recognition systems," *IEEE Transactions on Speech and Audio Processing,* vol. 6, no. 1, pp. 71-77, 1998.

[10] D. Rybach, C. Gollan, G. Heigold, B. Hoffmeister, J. Lööf, R. Schlüter and H. Ney, "The RWTH Aachen University Open Source Speech Recognition System," in *Interspeech*, 2009.

[11] S. Wiesler, A. Richard, P. Golik, R. Schluter and H. Ney, "RASR/NN: The RWTH neural network toolkit for speech recognition," in *ICASSP*, 2014.

[12] S. S. Chen and P. S. Gopalakrishnan, "Clustering via the Bayesian information criterion with applications in speech recognition," in *ICASSP*, 1998.

[13] J. G. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer Output Voting Error Reduction (ROVER)," in *Automatic Speech Recognition and Understanding*, 1997.

[14] G. Hinton, D. L., 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,* vol. 29, no. 6, pp. 82-97, 2012.

[15] 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 *ASRU*, 2011.

[16] F. Wessel, R. Schlüter, K. Macherey and H. Ney, "Confidence measures for large vocabulary continuous speech recognition," *IEEE Transactions on Speech and Audio Processing,* vol. 9, no. 3, pp. 288-298, 2001.

[17] J. L. Gauvain, L. F. Lamel, G. Adda and M. Adda-Decker, "The LIMSI continuous speech dictation system: Evaluation on the ARPA Wall Street Journal task," in *ICASSP*, 1994.

[18] H. Ney, L. Welling, S. Ortmanns, K. Beulen and F. Wessel, "The RWTH large vocabulary continuous speech recognition system," in *ICASSP*, 1998.

[19] H. Jiang, "Discriminativetraining of HMMs for automatic speech recognition: A survey," *Computer Speech and Language,* vol. 24, no. 4, p. 589–608, 2010.

[20] X. He and L. Deng, "Speech Recognition, Machine Translation, and Speech Translation—A Unified Discriminative Learning Paradigm," *IEEE Signal Processing Magazine,* vol. 28, no. 5, pp. 126-133, 2011.