UNSUPERVISED ADAPTATION WITHOUT ESTIMATED TRANSRIPTIONS

Hyeopwoo Lee and Dongsuk Yook

Speech Information Processing Laboratory
Department of Computer and Communication Engineering
Korea University, Korea

ABSTRACT

To estimate the unknown distortion parameters from input test signals, estimated transcriptions are typically used for unsupervised adaptation. In a low signal to noise ratio (SNR) condition, the transcription estimated by a decoding procedure can be error prone because of the high mismatch between the acoustic models and the input signal. As a result, it can cause performance degradation of the adapted systems. To account for this problem, we propose an unsupervised adaptation method that can adapt the acoustic models without the estimated transcription. Instead, Gaussian mixture models (GMM) and pseudo phoneme models (PPM) are used. Using these models the unknown distortion parameters are estimated based on the vector Taylor series (VTS) model adaptation scheme. On the Aurora2 task, we obtained relative reduction of 5.4% in word error rate (WER).

Index Terms — Unsupervised adaptation, vector Taylor series, robust speech recognition

1. INTRODUCTION

In general, automatic speech recognition performance is affected by environmental differences due to noises, channels, and speakers. Adaptation in automatic speech recognition aims to reduce these differences between training and testing conditions. For the adaptation, the normalized signal may be obtained from the noisy signal to make the testing condition similar to that of acoutic model training, i.e, feature adaptation [1]. Also, a method is available that modifies the pre-trained acoustic models to match the condition of the current signal, i.e., model adaptation [2].

The adaptation methods can be divided into two categories depending on the existence of transcription of the input data [3]. The supervised adaptation methods use the true transcription to estimate the adaptation information. Whereas, the unsupervised adaptation schemes typically estimate the transcription using current models. For most of the cases in automatic speech recognition, the transcription of the input signal is obtained by a decoding process with the pre-trained acoustic models [4][5][6]. Using these transcriptions, the distortion parameters are estimated to adapt the acoustic models or the feature vectors. After the adaptation, the recognition result may be generated with the adapted acoustic models or feature vectors. In this paper, the unsupervised model adaptation process is discussed.

Once the transcription is available, the unsupervised adaptation can be performed with any adaptation algorithm such as maximum likelihood linear regression (MLLR) and the vector Taylor series (VTS) methods [8][9][10]. However, the estimated transcription obtained by the conventional decoding procedure can be problematic. Especially in noisy conditions, the estimated transcriptions can be error prone because of large environmnetal differences between the training data for the acoustic models and the current input signal. So, the inaccurate estimation of the environmental distortion parameters can be caused by the unreliable estimation of the posterior probability of the hidden Markov model (HMM) states. Thus, it can lead to the performance degradation of the adaptated system.

To solve this problem, we consider two approaches that adapt the acoustic models without the hypothesis transcriptions. We assume that only the current test data is available for the adaptation. The first method uses Gaussian mixture models (GMM) to estimte the distortion parameters. In the second method, the pseudo phoneme model (PPM) sequence is used instead of the estimated transcriptions. Both approaches are applied to the VTS model adaptation scheme which shows good recognition performance in noisy conditions.

The remainder of this paper is organized as follows. Section 2 briefly reviews the VTS model adaptation scheme. The proposed methods, GMM and PPM based methods, are explained in Section 3. Section 4 provides some experimental results on the Aurora2 task. Finally, we conclude the paper with some future work in Section 5.

2. VTS MODEL ADAPTATION

For noisy speech recognition, we utilize the well-known VTS compensation scheme which shows good performance in noisy conditions [7]. To compensate the noise and channel effects, the distorted signal is modeled with

nonlinear formation. In the cepstral domain, the noisy speech signal can be expressed as

$$y = x + h + C\log(1+\exp(C^{-1}(n-x-h))) \quad , \tag{1}$$

where y, x, n, and h are the noisy speech, clean speech, additive noise, and channel distortion, respectively. C and C^{-1} represent the discrete cosine transform matrix and its pseudo inverse matrix, respectively. The noise parameters, n and h, are assumed to follow the Gaussian distribution. The channel distortion parameters are assumed to be stationary. With the VTS expansion using μ_x, μ_h, and, μ_n which are the mean of x, h, and n, respectively, we have

$$\begin{aligned} y &\approx \mu_x + \mu_h + g(\mu_x, \mu_h, \mu_n) \\ &+ G(x-\mu_x) + G(h-\mu_h) + (I-G)(n-\mu_n) \quad , \end{aligned} \tag{2}$$

where $g(\mu_x, \mu_h, \mu_n) = C\log(1+\exp(C^{-1}(\mu_n - \mu_x - \mu_h)))$ and G denotes the partial derivatives of y with respect to x and evaluated at μ_x, μ_h, and, μ_n.

By taking the expectation on both sides of Eq. (2), the mean and covariance of the noisy speech for the k-th Gaussian in the j-th state of the HMM becomes

$$\begin{aligned} \mu_{y,jk} &\approx \mu_{x,jk} + \mu_h + g(\mu_{x,jk}, \mu_h, \mu_n) \\ \Sigma_{y,jk} &\approx G_{jk}\Sigma_{x,jk}G_{jk}^T + (I-G_{jk})\Sigma_n(I-G_{jk})^T \quad , \end{aligned} \tag{3}$$

where G_{jk} indicates the partial derivatives of y with μ_x replaced with $\mu_{x,jk}$. The dynamic parts of the acoustic models are expressed as

$$\begin{aligned} \mu_{\Delta y,jk} &\approx G_{jk}\mu_{\Delta x,jk} \\ \Sigma_{\Delta y,jk} &\approx G_{jk}\Sigma_{\Delta x,jk}G_{jk}^T + (I-G_{jk})\Sigma_{\Delta n}(I-G_{jk})^T \quad . \end{aligned} \tag{4}$$

To estimate noise and channel distortion parameters, λ, the expectation-maximization (EM) algorithm formulated in [6] can be used. The auxiliary Q function for an utterance is

$$Q(\lambda \mid \bar{\lambda}) = \sum_t \sum_j \sum_k \gamma_{jk,t} \log p(y_t \mid j,k,\bar{\lambda}) \quad , \tag{5}$$

where y_t denotes the noisy feature vector at time t and $\gamma_{jk,t}$ represents the posterior probability of Gaussian k in state j. By differentiating the auxiliary function with respect to μ_n and equating it to zero, the mean of the additive noise can be obtained as

$$\begin{aligned} \mu_n = \mu_{n,0} &+ \left\{ \sum_t \sum_j \sum_k \gamma_{jk,t} (I-G_{jk})^T \Sigma_{y,jk}^{-1} (I-G_{jk}) \right\}^{-1} \times \\ &\left\{ \sum_t \sum_j \sum_k \gamma_{jk,t} (I-G_{jk})^T \Sigma_{y,jk}^{-1} [y_t - \mu_{x,jk} - \mu_{h,0} - g(\mu_{x,jk}, \mu_{h,0}, \mu_{n,0})] \right\}, \end{aligned} \tag{6}$$

where $\mu_{n,0}$ and $\mu_{h,0}$ are the expansion point of μ_n and μ_h, respectively. The mean of the channel distortion is also obtained similarly. In this paper, updating the noise variance is not considered. The unsupervised adaptation procedure and detailed formulations of the VTS adaptation can be found in [9].

3. PROPOSED METHODS

Though the VTS adaptation approach shows good performance generally, the performance can be degraded when the estimated transcriptions in low signal-to-noise ratio (SNR) conditions are used. One of the reasons is the unreliably estimated transcriptions since the error prone transcriptions can cause the unreliable estimation of the posterior probabilities in the adaptation process. In order to overcome this problem, we consider two simple methods. The first one uses GMM and the second one uses PPM to estimate the distortion parameters.

3-1. GMM based Method

In speaker adaptive training, use of a GMM or a single state HMM has been investigated [11][12]. A simple target model trained on the data without any adaptation (i.e., unnormalized data) is used to estimate the speaker specific transformation. Similar idea has been applied to vocal tract length normalization, where fast warping factor calculation was guided by GMM [13]. These methods demonstrated better performance than the conventional HMM based method. In this paper, we utilize a simple GMM for VTS based unsupervised adaptation. The VTS using this method will be referred to as VTS-GMM.

3-2. PPM based Method

Using a GMM, a sequence of pseudo phoneme models is generated to represent the (pseudo) phonetic information of the input data. PPM was originally proposed for unsupervised keyword and speaker model training [14]. The algorithm is summarized as follows.

1) For each input vector calculate the log likelihood of each Gaussian of the GMM.
2) Select the top Z Gaussian distributions for each frame.
3) Cluster the column vectors of the Gaussian index table.
4) For each cluster, select top Z unique Gaussians and assign them to the state of a PPM. We aim to build a one-state PPM for each pseudo phoneme unit.
5) The PPMs generated in step 4 are concatenated to form a pseudo phoneme sequence which constitutes a hypothesis phonetic transcription.

In a conventional unsupervised adaptation method using HMMs, the phonetic transcription is estimated first and used in the state occupancy estimation phase. Then, an environmental distortion function is applied to the HMMs in the model transformation phase. However, in the proposed method using PPM, the noise and channel effects are applied to the GMM first and then the pseudo phonetic information is generated based on the noise adapted GMM. The VTS using this method will be referred to as VTS-PPM.

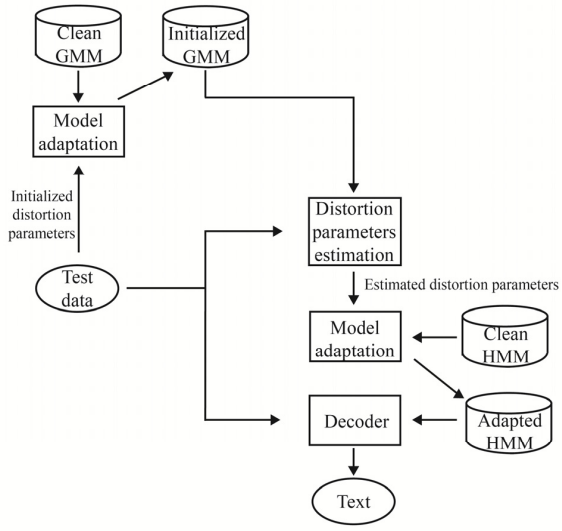

Fig. 1. Unsupervised VTS using GMM

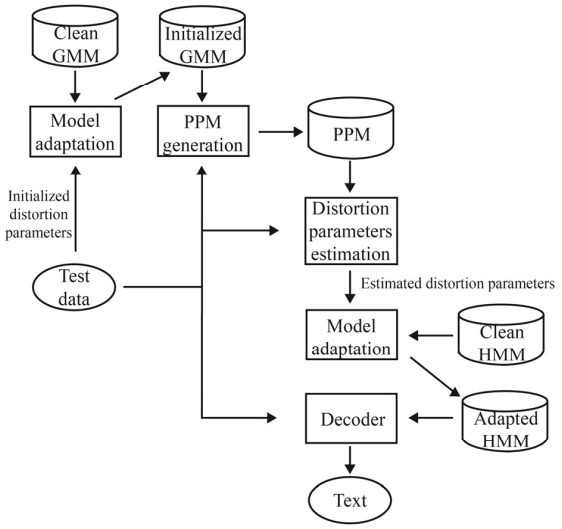

Fig. 2. Unsupervised VTS using PPM

3-3. Unsupervised Adaptation and Decoding Procedure

Unlike the conventional unsupervised adaptation methods, the VTS-GMM and VTS-PPM do not need lexical information such as language models. The overall procedure of VTS model adaptation using PPM is described as follows (Fig. 2);

1) Read a noisy test speech utterance and initialize the noise mean and variance using the first and last N frames and set zero for the channel parameter.
2) Adapt the clean GMM with the initial distortion parameters.
3) Make PPMs using the initialized GMM.
4) Compute the posterior probability of states and re-estimate the distortion parameters using the PPM.
5) Adapt the clean HMMs using the re-estimated distortion parameters.
6) Decode the utterance using the adapted HMMs.

For VTS-GMM, step 3) is omitted and PPM is replaced by the initialized GMM in step 4) (Fig. 1).

4. EXPERIMENTAL RESULTS

The connected digit data of Aurora2 [15] were used to verify the effectiveness of the proposed approaches. The 39-dimensional feature vectors were used; 13 mel-frequency cepstral coefficients and their first and second order time derivatives. 8,440 clean utterances of speaker independent training data were used for training the acoustic models. 11 word HMMs (‘zero’ to ‘nine’, and ‘oh’), silent, and a short pause models are trained. Each word HMM has 16 emitting states. Each state is modeled by 3 Gaussian mixture components. The silence model is modeled by 3 states, and each state has 6 Gaussian mixture components. The short pause model has 1 state, and it is modeled by 6 Gaussian mixture components. The diagonal covariance matrices were used. The test data is composed of three sets and each set has 7 different noise levels (clean and 20dB to -5dB). Test sets A and B are comprised of four different noise conditions. For the test set C, the convolutional noise was added to two different noise conditions. A GMM with 546 mixture components was trained using the same training data for VTS-GMM and VTS-PPM. The distortion parameters were estimated to compensate the acoustic models. For initialization of the noise mean, the first and the last 20 frames of the input signal were used. The initial channel mean was set to zero. The mean and covariance of the HMMs were adapted (including dynamic parts) for all experiments.

Table I summarizes the performance comparisons among the baseline system without any adaptation (Baseline), VTS, VTS-oracle, VTS-GMM, and VTS-PPM systems. VTS-oracle indicates a system that uses the true transcription to estimate the noise and channel distortion parameters. For 3rd column in Table I, the recognition

Table I
Performance (word error rate) of the baseline system, VTS, VTS-oracle, VTS-GMM, and VTS-PPM methods.

Test set	Baseline	VTS	VTS-oracle	VTS-GMM	VTS-PPM
Set A	43.8	21.0	17.0	19.9	19.8
Set B	45.7	19.3	15.7	19.0	18.8
Set C	37.7	21.8	17.2	19.9	19.8
Avg	43.4	20.5	16.5	19.5	19.4

Table II
Performance comparison for various SNRs

SNR (dB)	Baseline	VTS	VTS-oracle	VTS-GMM	VTS-PPM
Clean	0.9	1.0	0.8	0.9	0.9
20	5.3	2.0	1.5	2.1	2.1
15	15.3	3.4	2.6	3.4	3.3
10	37.7	6.7	5.1	6.7	6.7
5	66.6	15.8	12.2	15.4	15.4
0	85.8	40.4	30.9	37.5	37.3
-5	92.0	74.5	62.7	70.8	70.0
Avg	43.4	20.5	16.5	19.5	19.4

performance is significantly improved by the VTS model adaptation compared to the baseline system. As we expected, VTS-oracle utilizing the true transcription shows the best performance. On the average, the word error rate (WER) is reduced by 19.5% compared to VTS. It can be interpreted that the accuracy of the transcription has an important role to estimate the distortion parameters. As shown in the 5th column of Table I, use of GMM (VTS-GMM) shows good performance compared to VTS. The VTS-PPM method, in the last column of Table I, reduces the WER by 5.4% relatively compared to the VTS and 0.5% further compared to the VTS-GMM. The major improvement of the proposed methods come from the results of the Set C condition which exhibits both the additive noise and channel distortion simultaneously.

Table II presents the experimental results according to various SNRs. The performance is averaged over all three test sets. In all conditions except 20dB case, the performance of the VTS-GMM and VTS-PPM are slightly better than the VTS method. Especially, in low SNR conditions (below 5dB), the WER is reduced significantly. The VTS-PPM adaptation system reduces the WER by 10.0% at clean condition, 2.5% at 5dB, 7.7% at 0dB, and 6.0% at -5dB relatively compared to the conventional VTS adaptation system. Due to the absence of the one decoding step compared with conventional VTS adaptation system, the proposed methods have lower computational complexity.

It is also interesting to compare the performance with the conventional adaptation method. The unsupervised MLLR was conducted for comparison. The average WER of a full matrix transformation case is 43.0%, that of a block diagonal matrix transformation case is 37.4%, and that of a diagonal matrix transformation case is 29.7%, which is much higher than the proposed methods.

5. CONCLUSION

In this paper, we proposed the GMM and PPM based methods to improve the performance of the unsupervised adaptation. The effectiveness of these algorithms has been shown in an experimental study on the Aurora2 task. The estimated transcription and the language model are not needed to estimate the environmental distortion by the proposed methods. Due to the low computational complexity of these methods, it can be utilized for feature transformation for unsupervised training of huge training data.

6. ACKNOWLEDGMENTS

This work was supported by the Korea Research Foundation (KRF) grant funded by the Korea government (MEST) (No. 2011-0002906).

7. REFERENCES

[1] M. J. F. Gales, "Maximum likelihood linear transformation for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, pp. 75-98, 1998.

[2] M. J. F. Gales, "Model-based techniques for noise robust speech recognition," Ph.D. Thesis, Cambridge University, 1995.

[3] R. O. Duda, P. E. Hart, and D. G. Stork, Pattern Classification, Wiley, 2001.

[4] C. Mokbel, "Online adaptation of HMMs to real-life conditions : a unified framework," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 4, pp. 342-357, 2001.

[5] J. –T. Chien, "Quasi-Bayes linear regression for sequential learning of hidden Markov models," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 5, pp. 268-278, 2002.

[6] D. Yook, "Unsupervised incremental online adaptation to unknown environment and speaker," in Proc. *ICASSP*, pp. 616-620, 2002.

[7] C. Leggetter and P. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density HMMs," *Computer Speech and Language*, vol. 9, pp. 171-185, 1995.

[8] P. J. Moreno, "Speech Recognition in Noisy Environments," Ph.D. thesis, CMU, 1996.

[9] J. Li, L. Deng, D. Yu, Y. Gong, and A. Acero, "High-performance HMM adaptation with joint compensation of additive and convolutive distortions via vector Taylor series," in Proc. *ASRU*, pp. 65-70, 2007.

[10] J. Li, L. Deng, D. Yu, Y. Gong, and A. Acero, "A unified framework of HMM adaptation with joint compensation of additive and convolutive distortions," *Computer Speech and Language*, vol. 23, pp. 389-405, 2009.

[11] G. Stemmer, F. Brugnara, and D. Giuliani, "Adaptive training using simple target models," in Proc. *ICASSP*, pp. 997-1000, 2005.

[12] Q. Huo and B. Ma, "Irrelevant variability normalization in learning HMM state tying from data based on phonetic decision-tree," in Proc. *ICASSP*, pp. 577-600, 1999.

[13] L. Welling, H. Ney, and S. Kanthak, "Speaker adaptive modeling by vocal tract normalization," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 6, pp. 415-426, 2002.

[14] H. Lee, S. Chang, D. Yook, and Y. Kim, "A voice trigger system using keyword and speaker recognition for mobile devices," *IEEE Consumer Electronics*, vol. 55, no. 4, pp. 2377-2384, 2009.

[15] H. G. Hirsch and D. Pearce, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in Proc. *ISCA ITRW ASR*, pp. 181-188, 2000.