Maximum Likelihood Training and Adaptation of Embedded Speech Recognizers for Mobile Environments

Youngkyu Cho and Dongsuk Yook

For the acoustic models of embedded speech recognition systems, hidden Markov models (HMMs) are usually quantized and the original full space distributions are represented by combinations of a few quantized distribution prototypes. We propose a maximum likelihood objective function to train the quantized distribution prototypes. The experimental results show that the new training algorithm and the link structure adaptation scheme for the quantized HMMs reduce the word recognition error rate by 20.0%.

Keywords: Embedded speech recognition, maximum likelihood distribution clustering (MLDC), quantized HMM.

I. Introduction

Spoken language interfaces for mobile applications are important for the convenient operation of small mobile devices such as cellular phones and personal digital assistants (PDAs). Hidden Markov models (HMMs), of which states are represented as a mixture of Gaussian distributions, have been widely used for automatic speech recognition. Since HMMs require many Gaussians to obtain high recognition accuracy, they have a major disadvantage in the deployment of embedded speech recognition systems for mobile devices with low processing power. The memory requirement and the computation time of embedded speech recognizers need to be reduced without loss of recognition accuracy.

Typically, the recognition accuracy of a speech recognizer is increased by environment adaptation techniques such as maximum likelihood linear regression (MLLR) [1]. However, because mobile devices are used in unknown environments where the acoustic conditions change often, it is not always possible to collect enough adaptation data in advance. Therefore, we need an adaptation method which requires very little adaptation data, such as a few seconds of speech data.

We aim to design a speech recognition system that can run on mobile devices with low processing power and a small amount of memory without sacrificing recognition accuracy and that can be adapted to a new mobile environment with very little adaptation data. We propose a maximum likelihood distribution clustering (MLDC) algorithm to train the embedded speech recognition systems and study several adaptation methods for the embedded speech recognizers in the framework of MLLR. The experimental results show that the proposed MLDC algorithm combined with the link structure adaptation method decreases the word error rate (WER) of the recognizer by 20.0% compared to a conventional method.

II. Maximum Likelihood Distribution Clustering

One common approach that reduces both the size of the models and the computation time is to use parameter tying techniques, such as subspace distribution clustering HMMs (SDCHMMs) [2]. SDCHMMs divide the input feature space into multiple subspaces and cluster similar distributions within each subspace to produce a set of distribution prototypes for

Manuscript received June 8, 2009; revised July 20, 2009; accepted Aug. 19, 2009.

This work was supported by the MKE (Ministry of Knowledge Economy), Korea, under the ITRC (Information Technology Research Center) support program supervised by the NIPA (National IT Industry Promotion Agency) (NIPA-2009-C1090-0902-0007). It was also supported by the Korea University Grant.

Youngkyu Cho (phone: +82 2 3290 3641, email: ccameo@voice.korea.ac.kr) and Dongsuk Yook (corresponding author, email: yook@voice.korea.ac.kr) are with the Speech Information Processing Laboratory, Department of Computer and Communication Engineering, Korea University, Seoul, Rep. of Korea.

doi:10.4218/etrij.10.0209.0242

each subspace. The original full space distributions are represented as combinations of a few subspace distribution prototypes. Therefore, obtaining proper subspace Gaussian distribution prototypes is an important issue for SDCHMMs. In [2], subspace distributions are clustered by the *k*-means clustering algorithm using a distance measure such as Euclidean distance or Bhattacharyya distance. It minimizes the quantization error between the original continuous density HMMs (CDHMMs) and the SDCHMMs. However, the quantization error may not be directly related to the likelihood of the training data. For example, Euclidean distance does not make use of the variance of the training data. Thus, minimizing the quantization error does not necessarily maximize the likelihood of the training data, which is a commonly used training criterion for speech recognition systems. In this letter, we propose a novel training procedure using the maximum likelihood criterion to produce high-accuracy SDCHMMs.

The likelihood of the training data, given a set of *k*-th subspace distribution prototypes, can be defined as

$$\Lambda_k = \prod_{t=1}^{T} \sum_{l=1}^{L} \gamma_{kl}^t P(o_k^t \mid \mathcal{N}_{kl}), \tag{1}$$

where T is the number of training vectors; L is the number of distribution prototypes, which is the same as the number of clusters; γ_{kl}^t is the probability of training vector o_k^t coming from the l-th cluster, that is, occupation probability; and $P(o_k^t \mid \mathcal{N}_{kl})$ is the likelihood of observation vector o_k^t, given the l-th cluster prototype $\mathcal{N}_{kl}$. The likelihood of the training data in (1) can be increased by maximizing Baum's auxiliary function [3]. It can be simplified as follows when a Gaussian distribution is used for each state of the SDCHMMs [4]:

$$Q_k = \sum_{t=1}^{T} \sum_{l=1}^{L} \log\left(\frac{1}{\sqrt{(2\pi)^n \left|\Sigma_k^{\mathrm{CB}}[l]\right|}} e^{-\frac{1}{2}(o_k^t - \mu_k^{\mathrm{CB}}[l])'(\Sigma_k^{\mathrm{CB}}[l])^{-1}(o_k^t - \mu_k^{\mathrm{CB}}[l])} \right) \gamma_{kl}^t$$

$$\propto -\sum_{l=1}^{L} \frac{1}{2} \log\left|\Sigma_k^{\mathrm{CB}}[l]\right| \sum_{t=1}^{T} \gamma_{kl}^t, \tag{2}$$

where μ_k^{CB} and Σ_k^{CB} represent the codebooks for the *k*-th subspace mean vectors and covariance matrices, respectively, of which codewords are the distribution prototypes. Therefore, $\mu_k^{\mathrm{CB}}[l]$ and $\Sigma_k^{\mathrm{CB}}[l]$ are the l-th elements of the codebooks, which are the mean vector and the covariance matrix of prototype $\mathcal{N}_{kl}$, respectively.

By using (2) as an objective function when clustering the subspace Gaussian distributions to produce a set of prototypes for the *k*-th subspace, a set of clusters which maximizes the likelihood of the training data for the *k*-th subspace can be obtained. After finding the maximum likelihood clusters, a representative subspace Gaussian distribution for each cluster, that is, $\mu_k^{\mathrm{CB}}[l]$ and $\Sigma_k^{\mathrm{CB}}[l]$, can be obtained, which becomes the subspace distribution prototype for the cluster. Note that $\mu_k^{\mathrm{CB}}[l]$ and $\Sigma_k^{\mathrm{CB}}[l]$ are not calculated directly from the data. Instead, they are computed from the statistics of the Baum-Welch algorithm using a method similar to that described in [4].

Algorithm. MLDC
Project each Gaussian of the original CDHMMs on to the K subspaces.
for each subspace *k*
- Build a cluster that contains all subspace Gaussian distributions which originate from the original CDHMMs.
- Set the number of clusters, L, equal to 1.

repeat
- Increase the number of clusters L by splitting the cluster that gives rise to the largest likelihood improvement using (2) when it is split.

repeat
- Find all subspace Gaussian distributions to move from one cluster to another which result in the largest likelihood improvement using (2).
- Update the subspace distribution prototype of each cluster as in [4].

until the likelihood change falls below a preset threshold.
until L is equal to the number of clusters required.
end

The MLDC algorithm assigns each subspace Gaussian distribution into the cluster which results in the largest likelihood improvement. After all subspace Gaussian distributions are assigned, the set of representative Gaussian distributions are re-estimated. These steps are repeated until the likelihood increase falls below a preset threshold in the inner most loop of the algorithm.

Another important issue for embedded speech recognition is the rapid adaptation of the recognizers. We are considering unknown and constantly changing mobile environments where the recognizers are used. Since there is not a large amount of adaptation data available for such environments, the adaptation methods for the recognizers have to improve the recognition accuracy with a limited amount of adaptation data. This is called rapid adaptation. Some adaptation schemes for the SDCHMMs can be found in [5]-[7], namely, full space adaptation, codeword adaptation, and link structure adaptation. Some experimental results using maximum *a posteriori* (MAP) estimation can be found in [5], [7]. However, MAP is not suitable for rapid adaptation because it requires a relatively large amount of adaptation data. In this letter, we use the adaptation methods in the framework of MLLR, which requires less adaptation data than MAP. In the following section, we show that the link structure adaptation, which adapts the parameter tying structure while keeping the codebooks unchanged, is suitable for rapid adaptation.

III. Experiments

To evaluate the performance of our method, we performed speaker independent continuous speech recognition experiments using the Resource Management corpus. The speech feature vector used in the experiments was composed of 12 mel-frequency cepstral coefficients, one normalized log energy, and their first- and second-order time derivatives. For the initial CDHMMs, we trained acoustic models which are word internal triphones and decision-tree-based tied-state HMMs. There were 1,439 states in the CDHMMs and 6 Gaussian distributions per state. The WER of the baseline system was 6.2%.

The baseline CDHMMs were converted into 39-subspace SDCHMMs. Figure 1 shows the WERs of the SDCHMMs obtained using the conventional *k*-means clustering method [2] and using the MLDC method for various numbers of clusters. It can be seen that the SDCHMMs obtained using the proposed method have consistently lower WERs than those obtained using the conventional method. The SDCHMMs using the MLDC method showed the lowest WER of 5.5% when the number of subspace Gaussian prototypes was 64. The WER of the SDCHMM system was lower than that of the CDHMM system because the parameters of the SDCHMMs were more robustly estimated by the parameter sharing effect of the SDCHMMs [2]. This effect was diminished when the number of prototypes was more than 64.

In the next experiment, we compared the performance of the three adaptation methods for the SDCHMMs with 64 clusters obtained using the MLDC method. We varied the number of adaptation words from 1 to 11 to analyze the effect of the amount of adaptation data. Figure 2 shows that the link structure adaptation method is more rapid than the other two. This is because it imposes a stronger restriction (more inductive bias) on the adaptation process than the other two methods. With 11 adaptation words, the link structure adaptation method achieved a WER of 5.0%.

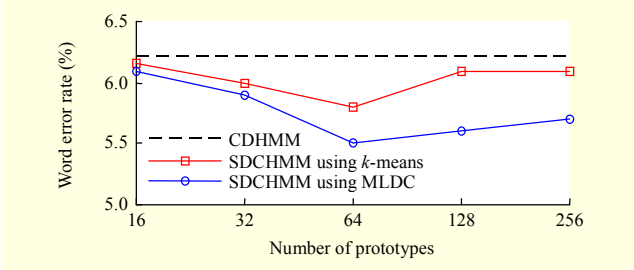

Fig. 1. Word error rates of the SDCHMMs obtained using the *k*-means clustering method and using the MLDC method.

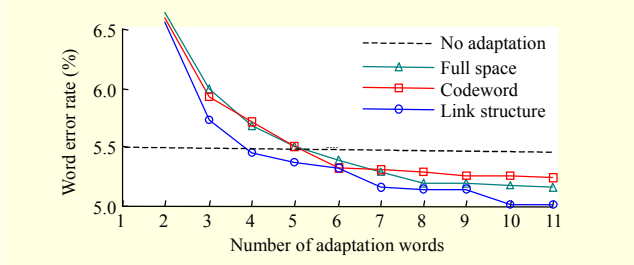

Fig. 2. Word error rates of the three adaptation schemes using SDCHMMs with MLDC in the framework of MLLR.

IV. Conclusion

We proposed a novel distribution clustering algorithm for SDCHMMs, which is based on the maximum likelihood criterion, and evaluated three adaptation schemes for the SDCHMMs in the framework of MLLR. The proposed MLDC method and the MLLR-based link structure adaptation scheme reduced the WER by 20.0% (from 6.2% to 5.0%). Link structure adaptation can be combined with recently developed faster approaches [8], [9].

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, 1995, pp. 171-185.

[2] E. Bocchieri and B. Mak, "Subspace Distribution Clustering Hidden Markov Model," *IEEE Trans. Speech Audio Process.*, vol. 9, 2001, pp. 264-276.

[3] L.E. Baum, "An Inequality and Associated Maximization Technique in Statistical Estimation of Probabilistic Functions of Markov Processes," *Inequalities*, vol. 3, 1972, pp. 1-8.

[4] J.J. Odell, *The Use of Context in Large Vocabulary Speech Recognition*, PhD Thesis, Cambridge University, 1995.

[5] K. Wong and B. Mak, "MAP Adaptation with Subspace Regression Classes and Tying," *IEEE Proc. Int. Conf. Acoust., Speech, Signal Process.*, vol. 3, 2000, pp. 1551-1554.

[6] K. Wong and B. Mak, "Rapid Speaker Adaptation Using MLLR and Subspace Regression Classes," *Proc. European Conf. Speech Commun. Technol.*, vol. 2, 2001, pp. 1253-1256.

[7] M. Zhang and J. Xu, "An Investigation into Subspace Rapid Speaker Adaptation," *IEEE Proc. Int. Symp. Chinese Spoken Language Process.*, 2004, pp. 273-276.

[8] D. Kim and D. Yook, "Linear Spectral Transformation for Robust Speech Recognition Using Maximum Mutual Information," *IEEE Signal Process. Lett.*, vol. 14, 2007, pp. 496-499.

[9] Y. Cho and D. Yook, "Rapid Adaptation Using Linear Spectral Transformation for Embedded Speech Recognizers," *IET Electron. Lett.*, vol. 44, no. 17, 2008, pp. 1040-1042.