IMPROVEMENTS IN SPEAKER ADAPTATION USING WEIGHTED TRAINING

Gyucheol Jang, Sooyoung Woo, Minho Jin and Chang D. Yoo

Department of Electrical Engineering and Computer Science
Korea Advanced Institute of Science and Technology
373-1, Guseong-dong, Yuseong-gu, Daejon, Republic of Korea, 305-701
tupp@mail.kaist.ac.kr, woodung@eeinfo.kaist.ac.kr, jinmho@mail.kaist.ac.kr, cdyoo@ee.kaist.ac.kr

ABSTRACT

Regardless of the distribution of the adaptation data in the testing environment, model-based adaptation methods that have so far been reported in various literature incorporate the adaptation data undiscriminatingly in reducing the mismatch between the training and testing environments. When the amount of data is small and the parameter tying is extensive, adaptation based on outlier data can be detrimental to the performance of the recognizer. The distribution of the adaptation data plays a critical role on the adaptation performance. In order to maximally improve the recognition rate in the testing environment using only a small number of adaptation data, supervised weighted training is applied to the structural maximum *a posterior* (SMAP) algorithm. We evaluate the performance of the proposed weighted SMAP (WSMAP) and SMAP on TIDIGITS corpus. The proposed WSMAP has been found to perform better for a small amount of data. The general idea of incorporating the distribution of the adaptation data is applicable to other adaptation algorithms.

1. INTRODUCTION

The performance of an automatic speech recognizer (ASR) degrades when there is a mismatch between the training and testing environments. To compensate for this mismatch, many methods have been proposed. These can be classified into two categories : feature compensation [1][8] which compensates for the observation in the process of feature extraction, and model adaptation[2]-[7] which estimates the new model parameters using only a small amount of adaptation data. This paper focuses on model adaptation.

Early model- adaptation methods can be categorized as either direct or indirect adaptation. Direct adaptation is based on Bayesian estimation[2]. Although it will approximately converge to the maximum likelihood estimator (MLE)- speaker dependent system- as the amount of adaptation data is increased, for a small amount of adaptation data, the improvement in recognition rate is limited. The difficulties associated with the determination of the prior density and the slow convergence with large number of hidden Markov model (HMM) parameters are also characteristics of this approach. Indirect model adaptation is an approach based on parameter transformation[3][4], which does not guarantee the convergence to the speaker dependent system. In this approach, the number of free parameters are small and thus the model can be adapted to the testing environment (or new speaker) with only a small amount of data. However, this approach does not take the full advantage of a large amount of data.

To overcome some of the demerits of each approach, recent methods have combined the two [5][6][7] so that a large improvement for a small amount of data and an approximate convergence to the MLE for a large amount of data can be achieved. One such method is the structural maximum *a posterior* (SMAP) algorithm which is a transformation-based maximum *a posteriori* (MAP) algorithm. With all its good qualities, the performance of SMAP is highly dependent on the adaptation data, and thus an outlier in the test environment can be detrimental to its performance. To reduce this dependency on the adaptation data, supervised weighted training is applied. Here each adaptation token is weighted by its confidence measure.

The organization of the paper is as follows. Section 2 describes weighted adaptation. Section 3 describes the proposed WSMAP. Section 4 discusses some experimental results. Section 5 finally concludes.

2. WEIGHTED ADAPTATION

The objective of a speaker adaptation system is to maximally improve its recognition rate using only a small number of adaptation data and to converge to the MLE as the amount of data increases. In order to achieve this, various transformation-based MAP algorithms have been proposed [2]-[7]. In all these methods, the effect of each adaptation data on the recognition rate in the testing environment is magnified with a decreasing amount of adaptation data.

This work was supported by grant No. R01-2000-000-00259-0(2002) from the Korea Science & Engineering Foundation.

Adaptation based on an outlier can degrade the performance of the recognizer.

Fig. 1 shows that adaptation based on an outlier data x_1 of the testing environment can give rise to an adapted model $\lambda_{x_1}^{adapt}$ that can be very different from the model λ_A^{test} of the testing environment, and the adaptation based on data x_2 that is representative of the testing environment can give rise to an adapted model $\lambda_{x_2}^{adapt}$ that is close to λ_A^{test}. For this reason, each adaptation token is given a weight that indicates its likelihood in the testing environment. Rather than discarding outliers, all tokens are incorporated in the adaptation procedure so that the adapted model can converge to the MLE of testing environment as the number of data increases as long as the weight satisfies certain constraints.

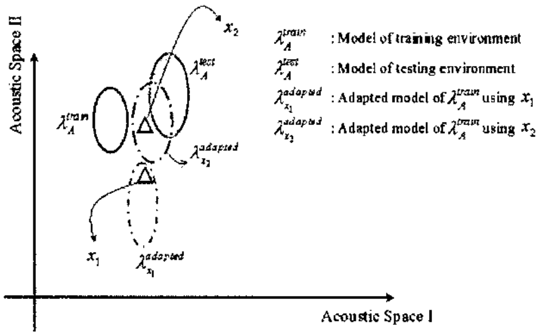
x_2
λ_A^{test}
λ_A^{train}
$\lambda_{x_2}^{adapted}$
x_1
$\lambda_{x_1}^{adapted}$
Acoustic Space II
Acoustic Space I
λ_A^{train} : Model of training environment
λ_A^{test} : Model of testing environment
$\lambda_{x_1}^{adapted}$: Adapted model of λ_A^{train} using x_1
$\lambda_{x_2}^{adapted}$: Adapted model of λ_A^{train} using x_2

Fig. 1. Influence of adaptation data on adaptation model

The degree of mismatch of the data is represented by the confidence measure of each data. The confidence measure of each token is used to weight each token[9]. The weighting places preference on data that is close to the training environment.

2.1. Convergence with confidence weighting

In the proposed adaptation procedure, each adaptation token is weighted by its confidence measure. In order to verify whether the expectation- maximization (EM) algorithm can be applied to the weighted adaptation data, the following argument is considered. From the following relationship[10]

$$\log\frac{P(\mathbf{X}|\lambda')}{P(\mathbf{X}|\lambda)} \geq Q(\lambda,\lambda') - Q(\lambda,\lambda) \ , \qquad (1)$$

where auxiliary function in obtaining the MLE [10]

$$Q(\lambda,\lambda') = \frac{1}{P(\mathbf{X}|\lambda)}\sum_{allS} P(\mathbf{X},S|\lambda)\ \log P(\mathbf{X},S|\lambda') \qquad (2)$$

and λ, λ', S and $\mathbf{X}$ are the previous HMM parameter, the updated HMM parameter, a particular state sequence through the HMM, and an observation sequence or an adaptation token respectively, it can be easily seen that $Q(\lambda,\lambda') \geq Q(\lambda,\lambda)$ gives $P(\mathbf{X}|\lambda') > P(\mathbf{X}|\lambda)$. For N observations, inequality (1) can be represented by the following expression given by Arslan[9]:

$$\frac{1}{N}\sum_{r=1}^{N}\log\frac{P(\mathbf{X}_r|\lambda')}{P(\mathbf{X}_r|\lambda)} \geq Q(\lambda,\lambda') - Q(\lambda,\lambda) \ . \qquad (3)$$

Both the previous and updated weights of the nth adaptation token $w_n = w(\mathbf{X}_n,\lambda)$ and $w_n' = w(\mathbf{X}_n,\lambda')$ are incorporated into the inequality given in (3). This gives the inequality shown by

$$\begin{aligned}&\frac{1}{N}\sum_{r=1}^{N}\log\frac{w_r' P(\mathbf{X}_r|\lambda')}{w_r P(\mathbf{X}_r|\lambda)}\\ &= \frac{1}{N}\sum_{r=1}^{N}\log\frac{P(\mathbf{X}_r|\lambda')}{P(\mathbf{X}_r|\lambda)} + \frac{1}{N}\sum_{r=1}^{N}\log\frac{w_r'}{w_r}\\ &\geq Q(\lambda,\lambda') - Q(\lambda,\lambda) \ . \qquad (4)\end{aligned}$$

From the above formulation, when $\frac{1}{N}\sum_{r=1}^{N}\log\frac{w_r'}{w_r} \leq 0$ then $P(\mathbf{X}|\lambda') > P(\mathbf{X}|\lambda)$ for $Q(\lambda,\lambda') > Q(\lambda,\lambda)$. In this paper, we do not consider this sufficient condition for the convergence property for each confidence weight. Through the experiments, we will show that the weight in equation (6) satisfies the convergence property.

2.2. Confidence weight

The likelihood ratio, which is a measure of the confidence on each token, can be formulated as

$$C_n^{(i)} = \frac{P(\mathbf{X}_n^{(i)}|\lambda_i)}{(\prod_{j=1,j\neq i}^{N} P(\mathbf{X}_n^{(i)}|\lambda_j))^{1/(N-1)}} \qquad (5)$$

where $\mathbf{X}_n^{(i)}$ is the nth training token of the ith word. There are many possible ways to formulate a measure based on the above confidence measure. In this paper, we employ the following weight to the token:

$$w_n^{(i)} = \alpha + \exp(-|\ln(P(\mathbf{X}_n^{(i)}|\lambda_i)) - \ln(P(\mathbf{X}_n^{(i)}|\lambda_j)) + \gamma|) \quad (6)$$

, where λ_j is the model with the largest likelihood given training token $\mathbf{X}_n^{(i)}$. In the above equation, α sets a floor on the minimum possible weight on each training token, and γ controls the level of adaptation data emphasis. In our experiments we used a value of 0.2 for α, a value of 1 for γ. The above expression is similar to the measure used in Arslan[9] and Juang[11].

3. WEIGHTED STRUCTURAL BAYES ADAPTATION(WSMAP)

As mentioned above, early direct adaptation algorithms show little improvement in recognition rate for a small amount of data. And early indirect adaptation algorithm cannot guarantee the convergence to speaker-dependent model for a large amount of data. To eliminate these degradations, an algorithm using the hierarchical tree structure was proposed by Shinoda[7].

The SMAP algorithm presents an effective method for deciding the *a priori* probability and estimating the mismatch between groups of Gaussian mixtures in HMM.

3.1. Tree Structure

To incorporate the benefits of indirect model adaptation when the amount of adaptation data is small, parameters are clustered into nodes and then the adaptation is applied. For continuous density HMM, Gaussian mixtures are used as the parameters. To make up the tree of the Gaussian mixtures, we defined distance between two Gaussian components as the sum of Kullback-Leibler divergence and used K-means algorithm in a top-down manner, as shown by Shinoda[7].

3.2. Weighted Structural Bayes Adaptation

3.2.1. Normalization of Gaussian distributions

For adaptation using a tree structure, we generate the normalized observation vectors and find the normalized Gaussian distribution using the normalized vectors. The tth observation x_{nt} of $\mathbf{X}_n$ is transformed into the vector y_{nmt} for each mixture component m and time t with the parameter θ_m of mixture m, as shown by

$$y_{nmt} = \Sigma_m^{-1/2}(x_{nt} - \mu_m) \tag{7}$$

Then the mismatch between the training environment(θ_m) and the testing environment(θ_l) can be found using the distribution of $\mathbf{Y_{nm}} = \{y_{nm1}, ..., y_{nmT}\}$. When the mismatch does not exist, x_t follows the distribution of θ_m and the normalized observation vector Y_{nm} follows the standard normal distribution $N(Y|, \vec{0}, I)$. When the mismatch does exist, Y_{nm} follows the distribution of $N(Y|\nu, \eta)$, where ν and η represent the shift and rotation of mixture components due to the mismatch, respectively. Therefore, we can represent the overall mismatch in a node using the parameter (ν, η).
For a set of M_k Gaussian mixture components
$G_k = \{g_1, ..., g_m, ..., g_{M_k}\}$ at the kth node, the MLE of the $(\bar{\nu}_k, \bar{\eta}_k)$ using the weight w_n is given by

$$\bar{\nu}_k = \frac{\Sigma_{n=1}^{N} w_n \Sigma_{t=1}^{T} \Sigma_{m=1}^{M_k} \gamma_{nmt} y_{nmt}}{\Sigma_{n=1}^{N} w_n \Sigma_{t=1}^{T} \Sigma_{m=1}^{M_k} \gamma_{nmt}} , \tag{8}$$

$$\bar{\eta}_k = \frac{\Sigma_{n=1}^{N} w_n \Sigma_{t=1}^{T} \Sigma_{m=1}^{M_k} \gamma_{nmt} (y_{nmt} - \bar{\nu})(y_{nmt} - \bar{\nu})^t}{\Sigma_{n=1}^{N} w_n \Sigma_{t=1}^{T} \Sigma_{m=1}^{M_k} \gamma_{nmt}} \tag{9}$$

where N is the number of adaptation data and $\gamma_{nmt} = P(m_t = m|\mathbf{X}_n, \lambda)$.

3.2.2. MAP estimator using Hierarchical Tree Structure

One of the difficulties of using MAP-based adaptation is the determination of the *a priori* probabilities of the parameters. The *a priori* must represent the characteristics of HMM parameters, which it may not. However using a hierarchical tree structure can alleviate the difficulty of determining *a priori* probability. That is, a child node inherits *a priori* probability of a parent node and makes use of it as a parameter for the child node's *a priori* probability[7].

The kth-level MAP estimate $(\hat{\nu}_k, \hat{\eta}_k)$ can be calculated from the $(k-1)$th-node estimates $(\hat{\nu}_{(k-1)}, \hat{\eta}_{(k-1)})$ as shown by

$$\hat{\nu}_k = \frac{\Gamma_k \tilde{\nu}_k + \tau_k \hat{\nu}_{k-1}}{\Gamma_k + \tau_k} , \tag{10}$$

$$\hat{\eta}_k = \frac{\hat{\eta}_{k-1} + \Gamma_k \tilde{\eta}_k + \frac{\tau_k \Gamma_k}{\tau_k + \Gamma_k} (\tilde{\nu}_k - \hat{\nu}_{k-1})^t (\tilde{\nu}_k - \hat{\nu}_{k-1})}{\Gamma_k + \xi_k} , \tag{11}$$

where Γ_k is defined as $\Gamma_k = \sum_{n=1}^{N} w_n \sum_{t=1}^{T} \sum_{m \in G_k} \gamma_{nmt}$ and $(\tilde{\nu_k}, \tilde{\eta_k})$ is a ML estimate of (ν_k, η_k). τ_k and ξ_k are hyperparameters to define the prior distributions of HMM parameters[7]. In this paper, these hyperparameters are not varied in all tree layer. In this equation, $\hat{\nu_0} = \vec{0}$ and $\hat{\eta_0} = I$ are assumed. Finally the mth MAP estimates of the Gaussian parameters $\hat{\mu}_m$ and $\hat{\Sigma}_m$ at each leaf (assume tree structure has K levels) can be calculated from $(\hat{\nu}_K, \hat{\eta}_K)$ by the following

$$\hat{\mu}_m = \bar{\mu}_m + (\bar{\Sigma}_m)^{1/2} \hat{\nu}_K \tag{12}$$

$$\hat{\Sigma}_m = \bar{\Sigma}_m^{1/2} \hat{\eta}_K (\bar{\Sigma}_m^{1/2})^t. \tag{13}$$

where $\bar{\Sigma}_m$ and $\bar{\mu}_m$ are the covariance and the mean for the mixture component $g_m(\cdot)$ respectively.

4. EXPERIMENTAL RESULTS AND DISCUSSION

We used TIDIGITS[12] to show the performance of WSMAP proposed in this paper. We trained a model for women with 1254 utterances and tested it with 1232 men's utterances. The feature was 13th MFCC calculated using 30ms frame with 10ms shift window. The recognition rate of men's test using women's model was 80.38% and that of men's test using men's model was 98.46%.

The adaptation result using TREE[7], SMAP and WSMAP are presented in and Table 1. As the number of adaptation data increased, the recognition rate of both SMAP and WSMAP converged to 98.46% which is the recognition rate of the speaker dependent system. Although both SMAP and WSMAP converged approximately to the same limit, Table 1 shows that WSMAP performed on average 3-5% better than SMAP. This can be attributed to WSMAP's efficient use of the adaptation data. The above result was based on

Table 1. Recognition rate obtained with supervised adaptation done with TREE, SMAP and WSMAP

Number of Adaptation Data	TREE	SMAP	WSMAP
Baseline	80.38	80.38	80.38
20	83.57	85.08	86.04
40	86.68	90.27	91.54
90	91.39	94.49	94.90
300	91.39	97.29	97.21

a four level and three node tree structure. However, the authors have supervised weighted training to be effective for different tree structures.

5. CONCLUSION

The performance of the proposed WSMAP algorithm is found to be better than the SMAP for a small amount of adaptation data, and as the amount of data is increased both methods converged to speaker dependent model. This shows the effectiveness of the supervised weighted training in adaptation. The general idea of incorporating the distribution of the adaptation data is applicable to other adaptation algorithms.

6. REFERENCES

[1] F.H. Liu, A. Acero, and R. Stern, "Efficient Joint Compensation of Speech For the Effects of Additive Noise and Linear Filtering," *IEEE International Conference on Acoustics, Speech, and Signal Processing,* pp. I-257 - I-260, March, 1992

[2] J.L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of markov chains," *IEEE Trans. Speech Audio Processing,* vol. 2, pp.291-298, 1994

[3] S. Furui, "Unsupervised speaker adaptation method based on hierarchical spectral clustering," *IEEE Trans. Acoust., Speech, Signal Processing,* vol. 37, pp. 1923.1930, December, 1989

[4] C. J. Leggetter and P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous-density hidden markov models," *Comput, Speech Lang.,* vol. 9, pp.171-185, 1995

[5] J.-I. Takahashi and S. Sagayama, "Vector-field-smoothed Bayesian learning for incremental speaker adaptation," *in Proc. ICASSP-95, Detroit, MI,* 1995, pp. 696.699

[6] O. Siohan, C. Chesta, and C.-H. Lee, "Hidden Markov model adaptation using maximum a posteriori linear regression," *in Proc.Workshop Robust Methods for Speech Recognition in Adverse Conditions, Tampere, Finland,* 1999, pp. 147.150

[7] K. Shinoda and C.-H. Lee, "Structural MAP speaker adaptation using hierarchical priors," *in Proc. IEEE Workshop Speech Recognition Understanding,* vol. 2, pp.291-298, 1994

[8] C.-H. Lee, "On stochastic feature and model compensation approaches to robust speech recognition," *Speech Commun.,* vol 25, pp.29-47, 1998

[9] Levent M. Arslan and John H. L. Hansen, "Selective training for hidden makov models with applications to speech classification," *IEEE Trans. Speech and Audio Processing,* vol. 7, No. 1 pp. 46-54 January 1999

[10] L. E. Baum and J. A. Eagon, "An inequility with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology," *Bull. Amer. Math. Soc.,* vol. 73,pp. 360-363, 1967

[11] B. -H. Juang and S.Katagiri, "Discriminative learning for minimum error classification," *IEEE Trans. Signal Processing,* vol. 40 ,pp. 3043-3054, 1992

[12] R. G. Leonard, "A Database for Speaker Independent Digit Recognition," *in ICASSP, San Diego California,* 1984, vol. 3,p.42.11