Speaker Adaptation Based on Confidence-Weighted Training

Gyucheol Jang, Minho Jin, and Chang D. Yoo

Department of Electrical Engineering and Computer Science
Korea Advanced Institute of Science and Technology
tupp@mail.kaist.ac.kr, jinmho@mail.kaist.ac.kr,
cdyoo@ee.kaist.ac.kr

Abstract

This paper presents a novel method to enhance the performance of traditional speaker adaptation algorithm using discriminative adaptation procedure based on a novel confidence measure and non-linear weighting. Regardless of the distribution of the adaptation data, traditional model adaptation methods incorporate the adaptation data undiscriminatingly. When the data size is small and the parameter tying is extensive, adaptation based on outliers can be detrimental. A way to discriminate the contribution of each data in the adaptation is to incorporate a confidence measure based on likelihood. We evaluate and compare the performances of the proposed weighted SMAP (WSMAP) which controls the contribution of each data by sigmoid weighting using a novel confidence measure. The effectiveness of the proposed algorithm is experimentally verified by adapting native speaker models to nonnative speaker environment using TIDIGIT.

1. Introduction

Environmental mismatch between training and testing degrades the performance of an automatic speech recognizer (ASR). To compensate for the 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 adapts existing parameters to new environment using a small amount of adaptation data. This paper focuses on the model adaptation.

Early model based adaptation methods can be categorized as either direct or indirect adaptation. Direct adaptation is based on Bayesian estimation [2]. Although the adapted model approximately converges to the maximum likelihood estimator (MLE) 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 direct adaptation.

Indirect model adaptation is an approach based on parameter transformation [3] [4], which does not guarantee the convergence to the speaker dependent system. In indirect adaptation, 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, it does not take full advantage of all the information that the data carries when data size is large.

To overcome some of the limitations of each approach, recent methods have combined the two approaches [5] [6] [7] so that a large improvement for a small data size and an approximate convergence to the MLE for a large data size can be achieved. One such method is the structural maximum *a posterior* (SMAP) algorithm which is a transformation-based maximum *a posteriori* (MAP) algorithm. The performance of SMAP like many of the traditional method is highly dependent on the adaptation data, and thus an outlier can be detrimental to its performance. To reduce this dependency on the adaptation data, this paper, the continuation of our previous work [13], proposes a novel training method based on supervised non-linear weighting using novel confidence measure.

This paper is organized as follows. Section 2 describes adaptation with sample weighting based on confidence measure. Section 3 describes the proposed WSMAP. Section 4 presents experimental results.

2. Adaptation Based On Token Weighting Using Confidence Measure

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 and MAP algorithms have been proposed [2]- [7]. All these methods efficiently compensate for the mismatch between training and testing environment. In all these methods, the effect of each adaptation data on the recognition rate in the testing environment is magnified with a decreasing adaptation data size. Outliers can degrade the performance of the recognizer. Fig. 1 shows that 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 adap-

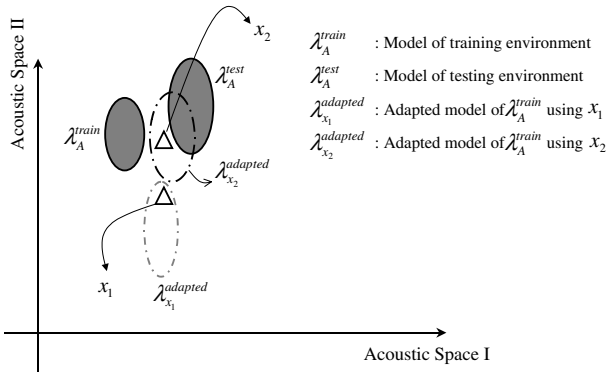

Figure 1: *Influence of adaptation data on adaptation model*

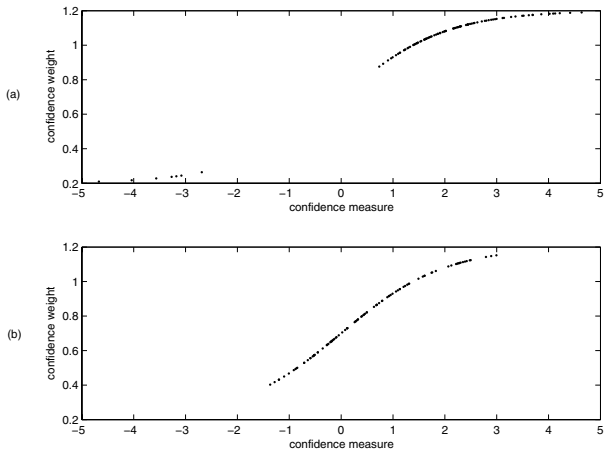

Figure 2: *The weight distribution versus confidence measure based on native speaker model: 100 data spoken by (a) native speaker and (b) non-native speaker.*

tation 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. 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. Confidence weight

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

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

where $\mathbf{X}_n^{(i)}$ and λ_j is the nth training token of the ith word and model of jth word. The above confidence measure can be calculated for each frame of the token, but we have chosen to incorporate confidence measure based on token. In this paper, normalized logarithm of confidence measure of each token expressed as

$$C_n^{\lambda_i} = 1 - \frac{1}{(N-1)} \frac{\sum_{j=1,j\neq i}^{N} \ln P(\mathbf{X}_n^{(i)}|\lambda_j)}{|\ln P(\mathbf{X}_n^{(i)}|\lambda_i)|} \quad (2)$$

is used to measure the confidence degree in the experiments. To simplify the process of the adaptation, the minimum difference of likelihoods of tokens was adopted as

$$\hat{C}_n^{\lambda_i} = \frac{\ln P(\mathbf{X}_n^{(i)}|\lambda_i) - \max_{j,j\neq i} \ln P(\mathbf{X}_n^{(i)}|\lambda_j)}{|\ln P(\mathbf{X}_n^{(i)}|\lambda_i)|}. \quad (3)$$

There are many possible ways to formulate a measure based on the above confidence measure. In this paper, we employ the following weight function to each token

$$\begin{aligned} w_n^{(i)} &= w_n(\hat{C}_n^{\lambda_i}) \\ &= \alpha + \frac{1}{1+\exp(-\beta(\hat{C}_n^{(i)} - \gamma)} \end{aligned} \quad (4)$$

where α, β and γ are constants shaping the weighting function: α is a floor value on the minimum possible weight on each training token, β represents the emphasis or de-emphasis of each token according to confidence degree, and γ represents the level of confidence measure. In this paper, we sets $\alpha = 0.2$, $\beta = 1$ and $\gamma = 0$. This weight function is somewhat different with that used in [13]: we have found the proposed nonlinear confidence weight to be more effective that that proposed in [13].

2.2. Confidence measure in speaker adaptation

It is difficult to assess the impact or contribution of an adaptation token on the overall adaptation. However, assuming that tokens of equal content but different confidence measures have different impact on adaptation and the confidence measure of a token in the testing environment can be approximated by the confidence measure in the training environment, we can weight each token based on its confidence measure in the adaptation process.

Data that may seem to be equal in terms of perceptual quality can have different confidence measure; however, their confidence measures are generally clustered closed to each together. We looked at the confidence weights as functions of confidence measure, defined by Equation (3) and (4) with constant values given in Section 2.1, of 100 randomly chosen data from TIDIGIT corpus. This data are from native speakers. As shown in Fig.2 (a), the clustering of the confidence measure is perceived and α, β and γ are set so that the weight of these data have value close to 1 with the exception of few that have confidence measure away from the cluster. We collect data with the same contents to that of 100 randomly chosen data from a non-native speaker. The weight distribution of the non-native speaker versus confidence measure based on native speaker model is obtained and shown in Fig.2 (b).

To understand the effect of confidence measure on adaptation, adaption of native speaker model to non-native speaker environment using SMAP [7] is considered. Since higher confidence measure means lower mismatch with training environment and lower valued confidence measure means higher mismatch, data for adaptation can be separated into two groups: a group of data having confidence measures lower than 0 and the

other group of data having confidence measures higher than 0. The word error rate (WER) of the recognition system adapted with the former group of 38 data is 25.56%, and 18.89% with the other group with the same number of data. The baseline system, tested non-native data on the native model, has WER of 34.55%. This results indicate that there were more outliers, respect to testing environment (non-native), in the former group than the other group. Therefore, the confidence measure based on likelihood can represent the impact or contribution of the adaptation data on the adapted model. We can vary the weights of the adaptation data to improve the recognition performance of the adapted model.

3. Weighted Structural Bayes Adaptation(WSMAP)

Before presenting the weighted speaker-adapted training algorithm, we should prove whether the given weights satisfy the property of convergence on HMM model parameters if we want to adapt the weights as the training iteration progresses. However, in our previous work [13], we have proven that there is a sufficient condition of weights and have shown experimentally that the given weights satisfy the property of convergence. In this paper, we will show experimentally that the given in 2.1 maintain the property of convergence of HMM parameters.

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]. In this chapter, WSMAP which is a weighed adaptation method based on SMAP will be presented.

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 a tree of 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

$$y_{nmt} = \Sigma_m^{-1/2}(x_{nt} - \mu_m) \quad (5)$$

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 $(\tilde{\nu}_k, \tilde{\eta}_k)$ using the weight w_n is given by

$$\tilde{\nu}_k = \frac{\Sigma_{n=1}^{N} w_n \Sigma_{t=1}^{T} \Sigma_{m=1}^{M_k} \gamma_{nmt} y_{nmt}}{\Sigma_{n=1}^{N} w_n \Sigma_{t=1}^{T} \Sigma_{m=1}^{M_k} \gamma_{nmt}} , \quad (6)$$

$$\tilde{\eta}_k = \frac{\Sigma_{n=1}^{N} w_n \Sigma_{t=1}^{T} \Sigma_{m=1}^{M_k} \gamma_{nmt} (y_{nmt} - \tilde{\nu})(y_{nmt} - \tilde{\nu})^t}{\Sigma_{n=1}^{N} w_n \Sigma_{t=1}^{T} \Sigma_{m=1}^{M_k} \gamma_{nmt}} \quad (7)$$

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

$$\hat{\nu}_k = \frac{\Gamma_k \tilde{\nu}_k + \tau_k \hat{\nu}_{k-1}}{\Gamma_k + \tau_k} , \quad (8)$$

$$\hat{\eta}_k = \frac{\hat{\eta}_{k-1} + \Gamma_k \tilde{\eta}_k + \frac{\tau_k \Gamma_k}{\tau_k + \Gamma_k} (\tilde{\nu}_k - \hat{\nu}_{k-1})^t (\tilde{\nu}_k - \hat{\nu}_{k-1})}{\Gamma_k + \xi_k} , \quad (9)$$

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

$$\hat{\mu}_m = \bar{\mu}_m + (\bar{\Sigma}_m)^{1/2} \hat{\nu}_K \quad (10)$$

$$\hat{\Sigma}_m = \bar{\Sigma}_m^{1/2} \hat{\eta}_K (\bar{\Sigma}_m^{1/2})^t . \quad (11)$$

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 native speaker with 1254 women's utterances and 1232 men's utterances. The 13th MFCC feature is calculated using 30ms frame with 10ms shift window. The word error rate of a nonnative speaker's test using native model was 34.55%.

The adaptation results with a nonnative speaker using SMAP [7] and WSMAP are presented in Table 1.

Table 1: *Word error rate obtained with supervised adaptation done with SMAP and WSMAP*

Number of Adaptation Data	SMAP	WSMAP
(Baseline)	34.55	34.55
5	34	31.56
10	28.22	17.33
20	16.44	14.44
40	14	13.78
70	12	10
100	2.89	2.89

Although both SMAP and WSMAP converged approximately to the same limit, Table 1 shows that WSMAP performed maximally 15% or on average 5-7%better than SMAP. This can be attributed to WSMAP's efficient use of the adaptation data. The above result was based on a four level and three node tree structure. The authors have verfied effectiveness of supervised weighted training for different tree structures.

5. Conclusion

In this paper, we present a novel supervised adaptation method that is effective in actual testing environment. The adaptation performance degrades with increasing number of outlier data. We show that through supervised weighted training the effect of outliers on the adaptation performance can be greatly reduced. It is experimentally verified that the proposed method outperforms SMAP. This work is a continuation of our previous work [13]. Instead of using exponential weight function whose constants are difficult to determine, we use a sigmoid function to determine the weight of each data using a novel confidence measure. 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,* 1997

[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,* vol. 9, p.328-331, March 1984

[13] Gyucheol Jang, Sooyoung Woo, Minho Jin and Chang D. Yoo, "Improvements in speaker adaptation using weighted training," *in Proc.IEEE ICASSP 2003, Hong Kong China,* vol. 1, pp.548-551, April 2003