INTERSPEECH 2009 BRIGHTON

A Study on Soft Margin Estimation of Linear Regression Parameters for Speaker Adaptation

Shigeki Matsuda[1], *Yu Tsao*[1], *Jinyu Li*[2], *Satoshi Nakamura*[1], *and Chin-Hui Lee*[3]

[1] Spoken Language Communication Group, National Institute of Information and Communications Technology, Kyoto, Japan
[2] Speech Component Group, Microsoft Corporation, Redmond, WA. USA
[3] School of Electrical and Computer Engineering, Georgia Institute of Technology, GA. USA

{shigeki.matsuda, yu.tsao, satoshi.nakamura}@nict.go.jp,
jinyuli@exchange.microsoft.com, chl@ece.gatech.edu

10.21437/Interspeech.2009-208

Abstract

We formulate a framework for soft margin estimation-based linear regression (SMELR) and apply it to supervised speaker adaptation. Enhanced separation capability and increased discriminative ability are two key properties in margin-based discriminative training. For the adaptation process to be able to flexibly utilize any amount of data, we also propose a novel interpolation scheme to linearly combine the speaker independent (SI) and speaker adaptive SMELR (SMELR/SA) models. The two proposed SMELR algorithms were evaluated on a Japanese large vocabulary continuous speech recognition task. Both the SMELR and interpolated SI+SMELR/SA techniques showed improved speech adaptation performance in comparison with the well-known maximum likelihood linear regression (MLLR) method. We also found that the interpolation framework works even more effectively than SMELR when the amount of adaptation data is relatively small.

Index Terms: speech recognition, speaker adaptation, discriminative training, soft margin estimation

1. Introduction

In recent years, discriminative training (DT) has been extensively studied to boost the performance of automatic speech recognition (ASR) systems, with acoustic models based on hidden Markov models (HMMs) with state observation densities characterized by Gaussian mixture models (GMMs). The most successful DT criteria used to estimate HMM parameters are maximum mutual information (MMI) [1], minimum classification error (MCE) [2], and minimum word/phone error (MWE/MPE) [3] in acoustic modeling. MMI training increases the distance between correct and competing candidates by maximizing the posterior probabilities of the observed speech utterances when correct transcriptions are provided. MCE minimizes the loss function defined to approximate string errors. Finally, MPE attempts to directly minimize an approximate word or phone error. If the acoustic conditions in the testing data closely match those in the training set, these DT algorithms usually achieve very good performance. However, such a good match cannot always be guaranteed for practical recognition situations, and therefore generalization capabilities of learning algorithms to unseen conditions are important research issues in DT.

[0]This study was conducted at Georgia Institute of Technology while the first author was a visiting scholar during January-March 2009

On another front, margin-based algorithms such as large margin estimation (LME) [4], large margin GMM/HMM (LM-GMM [5], LM-HMM [6]) and soft margin estimation (SME) [7] have recently been proposed to reduce the degree of model overfitting to the training data. In contrast to the above conventional DT methods, these margin-based techniques deal with the generalization issue from the perspective of statistical learning theory [9]. Among them, SME attempts to make a direct use of soft margins in support vector machines [10] to optimize a combination of a measure of generalization and an empirical risk in order to enhance model separation in classifier learning.

In addition to estimating HMM parameters directly from training data, one can also learn a set of global transformations to project these parameters to obtain transformed HMMs. Linear regression (LR) matrix transformations are the most commonly used indirect method, and they are often estimated with a maximum likelihood criterion, resulting in the popular MLLR [13] method. Because of its ability to transform all HMM parameters simultaneously, MLLR is often used for model adaptation, especially when the amount of training/adaptation data is relatively limited. To explore DT-based adaptation, MCE has been studied to adapt LR parameters [11, 12], and MCELR demonstrated a better performance than conventional MLLR in speaker adaptation.

When only a small amount of adaptation data is available, class boundaries between correct and competing candidates must be estimated carefully from the viewpoint of "overfitting" on adaptation data. The recently proposed SME framework seems a good candidate for this purpose because of its built-in frame selection mechanism and generalization capability. In this paper, we propose a soft margin estimation-based linear regression (SMELR) framework for speaker adaptation. To be able to flexibly use any amount of adaptation data, we also propose a model interpolation scheme that is based on using a weighted combination of the SMELR-based speaker adaptive (SMELR/SA) model and the speaker-independent (SI) model before adaptation. We call this combination strategy SI+SMELR/SA. The combination weight can be adjusted according to the amount of available adaptation data.

The two proposed SMELR algorithms were evaluated on a Japanese large vocabulary continuous speech recognition task. Both the SMELR and interpolated SI+SMELR/SA techniques showed improved speech adaptation performance in comparison with the well-known MLLR method. We also found that the interpolation framework works even more effectively than

SMELR when the amount of adaptation data is relatively small.

2. Soft Margin Estimation of Linear Regression

We now briefly review the theory of SME and formulate a framework involving soft margin estimation for linear regression.

2.1. Soft Margin Estimation

In statistical learning theory, a test risk is bounded by the summation of two terms: an empirical risk (i.e., the risk on the training set) and a generalization function. Hence, there are two targets for optimization. One is to minimize the empirical risk; the other, to maximize the margin. The generalization function is defined using a monotonic increasing function of Vapnik & Chervonenkis dimension (VC_{dim}). Usually a classifier generalizes better with a small VC_{dim}, then VC_{dim} can be reduced by increasing the margin. The objective function to be minimized is

$$L^{SME}(\Lambda) = \frac{\lambda}{\rho} + R_{emp}(\Lambda) \quad (1)$$

$$= \frac{\lambda}{\rho} + \frac{1}{N}\sum_{i=1}^{N} l(O_i, \Lambda) \quad (2)$$

where Λ denotes the set of model parameters; $l(O_i, \Lambda)$ is a loss function for a training sample O_i such as an utterance, word, phone, or frame; N is the number of training samples. ρ is a soft margin; and λ is a coefficient to balance soft margin maximization and empirical risk minimization. $R_{emp}(\Lambda)$ is the empirical risk of the set of model parameters Λ, and it can be calculated using the loss function. Posterior probabilities are used for measuring the margin in this paper. The empirical risk is calculated using the set of training samples that have smaller posterior probabilities than the margin (i.e. are difficult to recognize).

Only a small amount of data is usually available for speaker adaptation. We decided to employ frame selection [8] to efficiently extract key local discrimination information from individual frames:

$$l(O_i, \Lambda) = \sum_j l(O_{ij}, \Lambda) \quad (3)$$

$$l(O_{ij}, \Lambda) = \begin{cases} \rho - d(O_{ij}, \Lambda), & if\ \tau < p(S_i|O_{ij}) < \rho \\ 0, & otherwise \end{cases} \quad (4)$$

where O_{ij} denotes the j-th frame of the i-th training utterance. $d(O_{ij}, \Lambda)$ is a separation measure between the correct and competing candidates for O_{ij}. $p(S_i|O_{ij})$ is the posterior probability of the correct transcription S_i at the j-th frame. The hinge loss function has an additional threshold τ. A training sample whose posterior probability is too small will not contribute to the loss computation because it could be an outlier. $p(S_i|O_{ij})$ is calculated from a lattice obtained by decoding as follows:

$$p(S_i|O_{ij}) = \sum_{S \in H_{ij}} \frac{P_\Lambda(O_i|S)P(S)}{\sum_{\hat{S} \in G_i} P_\Lambda(O_i|\hat{S})P(\hat{S})} \quad (5)$$

where G_i is the set of all word sequences in a lattice obtained by decoding the i-th training utterance O_i. H_{ij} denotes all word sequences that contain correct words passing the j-th frame. $p(S_i|O_{ij})$ can be calculated by applying a forward-backward algorithm on the lattice.

2.2. Linear Regression Based on SME

Model parameters of continuous density HMMs are transformed by means of two linear transformations as follows:

$$\hat{\mu}_{m_r} = \hat{W}_m \xi_{m_r} \quad (6)$$

$$\hat{\Sigma}_{m_r} = B_{m_r}^{Tr} \hat{H}_m B_{m_r} \quad (7)$$

where $\hat{W}_m$ and $\hat{H}_m$ are the respective linear transformations for a mean vector and a covariance matrix, and m and r denote the respective class and Gaussian density indices. Each class consists of R similar Gaussian components. ξ_{m_r} is the extended vector of a mean vector μ_{m_r}. Σ_{m_r} is a covariance matrix, and B_{m_r} is the inverse covariance matrix.

These transformations are estimated using the SME. A separation measure between correct and competing candidates is defined as follows:

$$d(O_{ij}, \Lambda) = -g(O_{ij}, \Lambda) + \bar{g}(O_{ij}, \Lambda) \quad (8)$$

where $g(O_{ij}, \Lambda)$ and $\bar{g}(O_{ij}, \Lambda)$ are the likelihoods of correct and competing candidates, respectively. The derivative of an SME loss function is calculated as:

$$\frac{\partial l(O_{ij}, \Lambda)}{\partial \Lambda} = \frac{\partial(\rho - d(O_{ij}, \Lambda))}{\partial \Lambda} \quad (9)$$

$$= \frac{\partial g(O_{ij}, \Lambda)}{\partial \Lambda} - \frac{\partial \bar{g}(O_{ij}, \Lambda)}{\partial \Lambda} \quad (10)$$

Linear transformation parameters are optimized by applying a generalized probabilistic descent algorithm [14]:

$$\Lambda_{k+1}^m = \Lambda_k^m - \epsilon_k \sum_{O_{ij} \in F_i} \frac{\partial l}{\partial \Lambda^m}\Big|_{\Lambda^m = \Lambda_k^m} \quad (11)$$

where F_i is the set of selected frames obtained with Eq. (4) We use Λ to generically denote the linear transformation, $\hat{W}$ or $\hat{H}$. Note that MCELR [12] can be implemented by replacing the summation in Eq. (11) with the derivative of a sigmoid function. The learning rate ϵ_k is calculated as:

$$\epsilon_{k+1} = \epsilon_k - \frac{\epsilon_0 T_k}{E \sum_{k=1}^{K} T_k} \quad (12)$$

where E and T_k are total number of epochs and number of frames in the k-th adaptation utterance, respectively.

2.3. Model Interpolation

Since SME was originally formulated to enhance the discriminative power of acoustic models and improve separation among competing models it may not be as effective when the amount

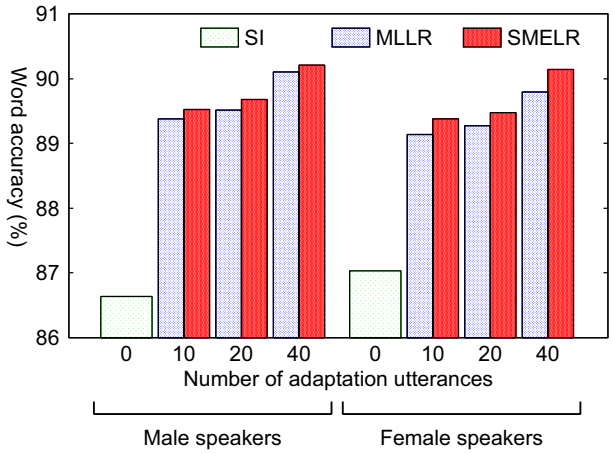

Figure 1: *Performance comparison with respect to the number of adaptation utterances: MLLR versus SMELR.*

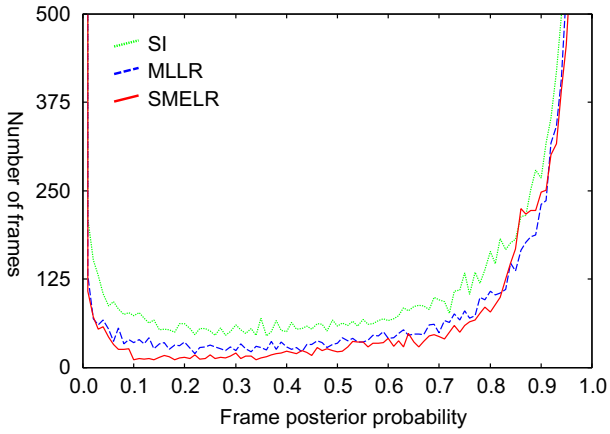

Figure 2: *Histograms of frame posterior probabilities of SI, MLLR-adapted, SMELR-adapted models.*

of adaptation data is relatively small. To flexibly utilize SMELR for any amount of data we also propose a model interpolation scheme that is based on using a weighted combination of the SMELR-based speaker adaptive (SMELR/SA) model and the speaker-independent (SI) model before adaptation. Interpolated mean vectors $\hat{\mu}$ are obtained using the mean vectors μ^{SI} of the available SI model and the mean vectors $\mu^{SMELR/SA}$ of the SMELR-adapted model.

$$\hat{\mu}_{m_r} = (w-1)\mu^{SI}_{m_r} + w\mu^{SMELR/SA}_{m_r} \tag{13}$$

where w is the weight assigned to the interpolation; it can be adjusted according to the amount of available data. When only a limited set is provided, this weight should be a small value so that we can rely more on prior information in the SI model. On the other hand, if more data is available, we expect the weight to approach 1, i.e., only the SMELR/SA model is used to reflect the dominance of posterior information. This is similar to the weighting mechanism used in MAPLR [15] but our proposed approach avoids using complex prior densities, which may be difficult to obtain.

3. Adaptation Experiments

3.1. System Configurations

We tested the proposed SMELR techniques on Japanese large vocabulary continuous speech recognition experiments. The Japanese Newspaper Article Sentences (JNAS) corpus [16] was used in our evaluations. We first used HTK to build the baseline speaker-independent (SI) HMMs with maximum likelihood (ML) training using a total of 25848 utterances from 250 speakers (125 male and 125 female). There were 3,000 tied-states, and each state had four Gaussian mixture components with diagonal covariance matrices. The acoustic feature comprised 12 MFCCs, 12 ΔMFCCs, and a Δpow, extracted with a 10-ms frame shift and a 20-ms frame length. The ATR speech recognition engine was used for decoding. Our ASR system uses word uni-, bi-, and tri-gram language models, that were estimated using the Mainichi Newspaper Corpus of 510M words spanning an 11-year period. The vocabulary size was 60,000.

A group comprising 23 male and 23 female speakers in the JNAS corpus was tested. We used 50 utterances for testing and 10, 20, and 40 sentences from each speaker for adaptation. MLLR adaptation was performed on a regression tree with 64 leaf nodes. The leaf occupation count threshold was set to 500. Full transformation matrices were used. The proposed SMELR commenced the adaptation process using transformation matrices estimated by MLLR. The initial learning rate ϵ_0 was fixed at 3×10^{-6}. The total number of training epochs was set to 20. Margin ρ and threshold τ were set to 0.83 and 0.10, respectively. The word lattices used for SMELR adaptation were obtained using the MLLR-adapted models.

3.2. Experimental Results

Fig. 1 shows the word accuracies provided by SI, MLLR, and SMELR. It is noted that SMELR can achieve better performance than MLLR. SMELR obtained slight relative error rate reductions of 1.4% and 2.5% from the MLLR baseline for the male and female speakers, respectively. In Fig. 2, we plot the histogram of the frame posterior probabilities of SI, MLLR, and SMELR on the adaptation set. When compared with the SI and MLLR models, SMELR can efficiently reduce the training samples that have posterior probabilities smaller than τ. It is clear that SMELR achieved a better separation than either the ML-estimated SI or MLLR-adapted models.

For model interpolation between SI and SMELR-adapted models with different combination weights (from 0.0 to 1.0), the resulting word accuracies are shown in Fig. 3. The best interpolation weight apparently depends on the amount of data used for adaptation. With 20 adaptation utterances (the two middle curves in Fig. 3), the optimal weight was 0.7 and the interpolation approach yielded an average 4.7% word error rate reduction over SMELR for both genders. With only 10 adaptation utterances (the two lower curves in Fig. 3), the optimal weight was 0.6 and the interpolation approach still achieved relative word error rate reductions of about 2.6% and 4.5% for male and female speakers, respectively. With 40 adaptation utterances the optimal weight for SMELR was 1.0. Thus, this set of experiments verifies that in comparison with the MLLR-adapted and SMELR-adapted models, model interpolation with appropriate combination weights can provide improved performance or at least the same level of performance.

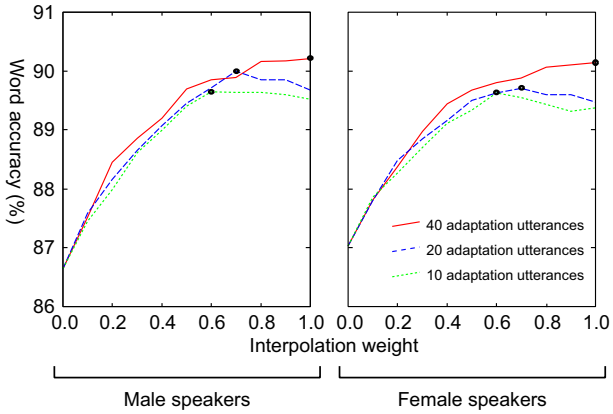

Figure 3: *Performance comparison for model interpolation with several weights on 10, 20, and 40 adaptation utterances. Each small circle on the curves indicates the best performance for each adaptation case.*

4. Summary and Future Work

We presented a model adaptation formulation of linear regression matrix parameters based on soft margin estimation, called SMELR, to improve the effectiveness of adaptation when the amount of training/adaptation data is relatively small. To be able to flexibly use any amount of data, we also proposed a novel model interpolation scheme between SI and SMELR-adapted models. Experimental results showed that SMELR can achieve better model separation than the MLLR-adapted models, and the model interpolation can further reduce recognition errors of the SMELR-adapted models by using an appropriate combination weight, depending on the amount of available adaptation data. In future work, we plan to analyze speaker-dependency properties of the soft margin. We expect that SMELR with an optimized speaker-dependent soft margin will achieve even higher accuracies.

5. References

[1] L. R. Bahl, P. F. Brown, P. V. de Souza, and R. L. Mercer, "Maximum mutual information estimation of hidden Markov model parameters for speech recognition," Proc. ICASSP, vol. 1, pp. 49–52, 1986.

[2] B. -H. Juang, W. Chou, and C. -H. Lee, "Minimum classification error rate methods for speech recognition," IEEE Trans. on Speech and Audio Proc., vol. 5, no. 3, pp. 257–265, 1997.

[3] D. Povey, and P. Woodland, "Minimum phone error and i-smoothing for improved discriminative training," Proc. ICASSP, vol. 1, pp. 105–108, 2002.

[4] X. Li, H. Jiang, and C. Liu, "Large margin HMMs for speech recognition," Proc. ICASSP, pp. V513–V516, 2005.

[5] F. Sha and L. K. Saul, "Large margin Gaussian mixture modeling for phonetic classification and recognition," Proc. ICASSP, pp. I265–I268, 2006.

[6] F. Sha and L. K. Saul, "Large margin hidden Markov models for automatic speech recognition," in Advances in Neural Information Processing Systems 19, B. Scholkopf, J. C. Platt, and T. Hofmann, Eds., MIT Press, 2007.

[7] J. Li, M. Yuang, and C. -H. Lee, "Approximate test risk Bound minimization through soft margin estimation," IEEE Trans. on Speech and Audio Proc., vol. 15, no. 8, pp. 2393–2404, 2007.

[8] J. Li, Z. -J. Yan, C. -H. Lee, and R. -H. Wang, "A study on soft margin estimation for LVCSR," Proc. ASRU, pp. 268–271, 2007.

[9] V. Vapnik, "The Nature of Statistical Learning Theory," Springer-Verlag, New York, 1995.

[10] C. Burges, "A tutorial on support vector machines for pattern recognition," Data Mining and Knowledge Discovery, vol. 2, no. 2, pp. 121–167, 1998.

[11] X. -D. He and W. Chou, "Minimum classification error linear regression for acoustic model adaptation of continuous density HMMs," Proc. ICASSP, pp. I556–I559, 2001.

[12] J. Wu and Q. Huo, "A study of minimum classification error (MCE) linear regression for supervised adaptation of MCE-trained continuous-density hidden Markov models," IEEE Trans. on Speech and Audio Proc., vol. 15, no. 2, pp. 478–489, 2007.

[13] 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.

[14] S. Katagiri, B. -H. Juang, and C. -H. Lee, "Pattern recognition using a family of design algorithms based upon the generalized probabilistic descent method," Proc. IEEE, vol. 86, no. 11, pp. 2345–2373, 1998.

[15] O. Siohan, C. Chesta and C. -H. Lee, "Joint Maximum a Posteriori Adaptation of Transformation and HMM Parameters," IEEE Trans. on Speech and Audio Proc., vol. 9, no. 4, pp. 417-428, May 2001.

[16] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, and S. Itahashi, "The design of the newspaper-based Japanese large vocabulary continuous speech recognition corpus," Proc. ICSLP, vol. 7, pp. 3261–3264, 1998.