Maximum Likelihood Smoothes and Predictions for Fast Speaker Adaptation

Guo-Qiang LI, Li-Min DU, Yan-Jun XU, Zi-Qiang HOU

Lab. for Interactive Information Systems

Institute of Acoustics, CAS, Beijing, 100080

Tel: (8610) 6262,7570 (O) Fax:(8610)6262,9250 E-mail: ligq@eudoramail.com

ABSTRACT

To realize fast speaker adaptation in the case of limited adaptation data, we propose a fast speaker adaptation approach, called maximum likelihood smoothes and predictions. It smoothes and predicts target mean vectors based on their source mean vector by maximizing the likelihood of the smoothed model generating the adaptation data. So it can make best use of the first few adaptation data to quicken adaptation process. It increases the model's prediction accuracy by off-line estimating regression matrices and on-line robustly estimating shift matrices. Moreover, it increases the model's predictive power at mean vector level to obtain the estimators of more bad-adapted and no-adapted model parameters even with a few of adaptation data.

1. INTRODUCTION

One problem faced by some speaker adaptation techniques is that only the parameters of those models, which are observed in the adaptation data, are updated. Hence, with small amounts of adaptation data most of the system parameters remain unchanged. So their convergence is slow. One approach [4] to this problem, called maximum likelihood linear regression (MLLR) trains a small number of regression matrices on the available adaptation data for a new speaker and transforms all the mean vectors in the system, i.e. from initial parameters of each mean vector to new estimators of the same mean vector, using one of these regression matrices. However, this technique is restricted to fairly broad adjustments to the parameter values, and requires several adaptation sentences before it starts to be effective.

Ahadi [1] has investigated the use of regression-based prediction to improve the performance of speech recognition systems. Ahadi discussed the application of predictive technique [2], such as regression-based model prediction (RMP) to a CDHMM-based large vocabulary continuous speech recognizer. In the RMP approach, given a set of well-adapted models (called source distributions), for example by MAP [3], and a set of regression parameters relating the sources and bad-adapted distributions (called target), the corresponding target parameters can be updated. But, in the RMP, the relationship between model parameters is at element level of state output Gaussian distribution mean vectors, that is, diagonal transformation matrices are used and vector elements are independent. It is desirable to consider the relationships between elements to obtain more precise regression coefficients.

We propose a fast speaker adaptation approach, called maximum likelihood smoothes and predictions (MLSP), that smoothes and predicts a target mean vector based on its source mean vector by maximizing the likelihood of the smoothed model generating the adaptation data, similar to MLLR. This approach makes best use of the first few adaptation data to quicken adaptation process. MLSP increases the model's prediction accuracy by off-line estimating regression matrices and on-line robustly estimating shift matrices. Moreover, MLSP increases the model's predictive power at mean vector level to obtain the estimators of more bad-adapted and no-adapted model parameters even with a few of adaptation data.

The main process of MLSP is as following. Firstly, the best-adapted distribution predicts no-adapted distributions and smoothes worse adapted distributions only if their correlation coefficients are high. Second, the smoothed distributions also predict no-adapted distributions and smooth worse adapted distributions only if their correlation coefficients are high. Thus, much more model parameters can be estimated only by the limited adaptation data and therefore the convergence of our new approach is faster.

But it is important how robustly regression matrices suited for the new speaker can be estimated only with a few of adaptation data, which are applied to source parameters to obtain estimators of target parameters, since the precision of the smoothes and predictions heavily depends on the regression matrices. To solve the above problem, we off-line estimate the regression matrices for each pair of better-correlated distributions. Moreover, we employ a simple shift matrix, which contains less parameters,

such as, a diagonal matrix, or a tri-diagonal matrix. Therefore, the shift matrix can be on-line robustly estimated even with a limited amount of adaptation data. The shift matrix adds up to the old regression matrix, which was estimated off-line by a number of SD HMM models, to form a new regression matrix suited for the new speaker. Robust regression matrix can then be estimated by maximizing the likelihood of smoothed distribution generating the adaptation data.

Moreover, predictions at mean vector level, instead of its element level, are multiple regression using full regression matrices. Full regression matrices are more effective than diagonal ones. In the multiple regression, the correlation coefficients among different elements vary sharply, even between the same pair of target and source distributions. It is useful to have multiple streams. The use of multiple streams also reduces calculation load and the storage of off-line estimated regression parameters, since it divides the long vector into multiple short vectors.

2. OFF-LINE ESTIMATION of RERESSION MATRICES and CORRELATION COEFFICIENTS

For two model parameters, e.g. output probability mean vectors of mixture components of different models, x (source), y (target), their linear relationship is assumed as such:

$$y = (a \quad A)\begin{pmatrix}1\\ x\end{pmatrix} + \varepsilon = \tilde{A}\tilde{x} + \varepsilon \qquad (1)$$

where $x, y, a,$ are D dimension vectors, $\tilde{x}$ is extended mean vector, A is DXD matrix, and a, A are regression parameters, ε is the error associated with this approximation. And its regression model in matrix notation is

$$\begin{pmatrix} y_1 & \cdots & y_s \end{pmatrix} = \tilde{A}\begin{pmatrix} \tilde{x}_1 & \cdots & \tilde{x}_s \end{pmatrix} + \begin{pmatrix} \varepsilon_1 & \cdots & \varepsilon_s \end{pmatrix} \quad (2)$$

or

$$Y = \tilde{A}\tilde{X} + \mathrm{E} \qquad (3)$$

where x_i, y_i are mean vectors of i speaker-specific HMM, s is the total number of speakers, (3) is equivalent to (2). The sample correlation coefficient for the i-th elements of the mean vector y and its predicted mean vector $\hat{y}$ [6]:

$$R_i^2 = 1 - \frac{\sum_{j=1}^{s}(y_{i,j} - \hat{y}_{i,j})^2}{\sum_{j=1}^{s}(y_{i,j} - \bar{y}_i)^2} \qquad (4)$$

The sample correlation coefficients for mean vectors $y, \hat{y}$ is defined as such:

$$R^2 = \sum_{i=1}^{D} R_i^2 \qquad (5)$$

If $(\tilde{X}\tilde{X}')$ is non-singular, (3) has a unique least squares estimate:

$$\tilde{A} = Y\tilde{X}'\left(\tilde{X}\tilde{X}'\right)^{-1} \qquad (6)$$

For every pair of potential source and target parameters, the above regression parameters and correlation coefficients are off-line calculated. In [1], all distributions are on-line divided into two groups of sources and targets based on a state or mixture occupation and then best source is selected for each target. Thus, it needs much on-line calculation time for all these. In our work, for each model parameter as source, its target parameters with better correlation coefficient above a threshold are selected and saved for on-line use with their regression parameters. So, it can dispose of much on-line calculation time at some sacrifice of storage. The threshold can be experimentally set. By introducing multiple streams and clustering techniques, the above calculation loads and storage can be significantly reduced.

3. ON-LINE MAXIMUM LIKELIHOOD SMOOTHES and PREDICTIONS

3.1 Linear Regression Prediction

For a particular source distribution, when its target distribution has no adaptation data, the target distribution can only be predicted by the linear regression parameters calculated in an off-line way and saved with the source distribution. In [1,2] only the model parameter relationships at an element level are considered. In our work, the model relationships at vector level are used for predictions. So more precise target parameters can be predicted. Moreover, all regression parameters are off-line calculated and saved with their source parameters for later use in on-line prediction. So this greatly reduces the on-line calculation time of regression parameters. For a well adapted mean vector x, its predicted vector value, $\hat{y}$ of target mean vector y is as such:

$$\hat{y} = \begin{pmatrix} a & A \end{pmatrix}\begin{pmatrix}1\\ x\end{pmatrix} = \tilde{A}\tilde{x} \qquad (7)$$

3.2 Maximum Likelihood Smoothes

For a particular source distribution, when its target distribution has a little adaptation data but not enough to robustly estimate the target parameter, the target parameters can be estimated by smoothing its regression parameters with the adaptation data corresponding to the target mean vector. These data stand for the little information of the target mean vector, but the information is not enough to estimate the target. So the regression parameters as general information are used to enforce the specific information for the target parameter.

The source distribution, s, is characterized by a

mean vector, x, and a diagonal covariance C_x, while its target distribution, t, is characterized by a mean vector, y, and a diagonal covariance C_y. Given a parameterized speech frame vector o, the probability density of that vector being generated by distribution t is $p_t(o)$:

$$p_t(o) = \frac{1}{(2\pi)^{D/2}\left|C_y\right|^{1/2}} e^{-1/2(o-y)'C_t^{-1}(o-y)} \quad (8)$$

The adaptation of the target mean vector is achieved by applying a transformation matrix $W_{x,y}$ to the extended mean vector $\tilde{x}$ to obtain an smoothed target mean vector $\hat{y}$:

$$\hat{y} = W_{x,y}\tilde{x} \quad (9)$$

where $W_{x,y} = \begin{pmatrix} a+b & A+B \end{pmatrix}$, a, A are regression parameters estimated off-line for the pair of better-correlated distributions, b is $D \times 1$ dimension vectors, B is $D \times D$ diagonal or tri-diagonal matrix or other simple matrix. For target distribution t, the pdf of the smoothed model generating the corresponding adaptation data, o, becomes:

$$p_t(o) = \frac{1}{(2\pi)^{D/2}\left|C_y\right|^{1/2}} e^{-1/2(o-W_{x,y}\tilde{x})'C_t^{-1}(o-W_{x,y}\tilde{x})} \quad (10)$$

The new approach MLSP estimates the new parameters b, B containing less parameters by maximizing the likelihood, containing (10), of the smoothed model generating the corresponding adaptation data.

3.3 Estimation of MLSP regression parameters b, B

Assume that the adaptation data, O, is a series of T observations $O = o_1 \cdots o_T$. Denote the current set of model parameters by λ and a re-estimated set of model parameters as $\bar{\lambda}$. Like [4], the likelihood of the smoothed models generating the adaptation data can be maximized by iteratively maximizing the following auxiliary function with smoothed parameters:

$$Q(\lambda, \bar{\lambda}) = \sum_{\theta \in \Theta} L(O, \theta|\lambda) \log\left(L(O, \theta|\bar{\lambda})\right) \quad (11)$$

where the set Θ denotes all possible state sequences of length T, $L(\cdot)$ is the above likelihood. The estimates of MLSP regression parameters, $\hat{b}$, $\hat{B}$ can be obtained like the derivation in [4].

The number and structure of regression parameters b, B, can be designed beforehand according to different amount of available adaptation data. Thus the number and structure can be automatically selected based on the amount of its adaptation data available. For example, the simplest form is as: $b = \begin{pmatrix} m & \cdots & m \end{pmatrix}^T$, $B = n \cdot I_D$, where I_D is identity matrix, m and n are scalar variables to be estimated.

4. ALGORITHM DEFSCRIPTION

The whole process of our approach is as following. See fig.1.

1) Off-line calculate regression parameters between pairs of better-correlated source and target mean vectors using all SD HMM parameters. For each source, save the regression parameters and its better-correlated target mean vectors if their correlation coefficients are above a threshold.
2) When adaptation data for a new speaker is available, make initial MAP adaptation for all mean vectors with adaptation data.
3) The best-adapted distribution predicts no-adapted distributions using formula (7) and smoothes worse adapted distributions by using formula derived from (11) only if their correlation coefficients are higher than a threshold.
4) The smoothed distributions also predict no-adapted distributions using formula (7) and smooth worse adapted distributions by using formula derived from (11) only if their correlation coefficients are higher than a threshold.
5) At last, make final MAP adaptation for the target mean vectors processed in the previous steps.

In order to improve the performance of the algorithm, speaker clustering can be made. The simplest example is speaker adaptation for only male speakers or female speakers.

5. EXPERIMENTAL EVALUATION

This section focuses on experimental setups. We use a speech database that consists of 123 speakers, female 57, male 66, each speaking about 500 to 600 sentences of continuous speech. The speech database is parameterized using 12 Mal frequency cepstral coefficient, normalized log energy and the first and second differentials of these parameters. The basic phone set consists of 47 phone symbols plus silence. Each phone is modeled by a single left-to-right five-state CDHMM. Each state has 4 stream and each stream has only a component. Each component is modeled by a diagonal covariance matrix. The system uses triphone acoustic model and language model of trigram.

SD models are trained for each speaker by ML or MAP [3] algorithms with 500 to 600 sentences as training data and the rest 20 to 50 sentences as testing

data. To evaluate MLSP algorithm, 16 speakers are selected as test speakers.

The initial experiments show that our algorithm is effective and has great improvements over MLLP and RMP. Detailed experimental results will be represented in the conference.

6. DISCUSSION

We propose a new fast speaker adaptation approach MLSP that makes best use of the first few adaptation data. The relationships between distributions can be obtained by off-line calculating linear regression parameters and correlation coefficients. These as general knowledge are saved for on-line speaker adaptation. When a few of adaptation data for a new speaker are available, model parameters with more adaptation data will be better adapted as compared with other parameters with less or no adaptation data. To quicken adaptation process, more model parameters must be estimated even with such a few of adaptation data. Thus, relationships between model distributions are used with the better-estimated model parameters as source to predict model parameters as target unseen in the adaptation data.

When target distribution has no adaptation data, its mean vector can be predicted by directly using regression parameters off-line calculated and source mean vector. When target distribution has a little adaptation data but not enough to robustly estimate its parameters, the regression parameters can be combined with better-adapted parameters to estimate the target mean vector. In order to make use of the adaptation data corresponding to target distribution and in consideration of robust estimation with the adaptation data of target distribution, simple shift matrix is employed. Moreover, likelihood of the target distribution generating its adaptation data is maximized to estimate the shift matrix.

7. REFERENCES

[1] Ahadi, S. M. & Woodland, P. C. (1997). "Combined Bayesian and predictive techniques for rapid speaker adaptation of continuous density hidden Markov models." *Computer Speech &Language* 11, 187-206.

[2] Cox, S.J. (1995). "Predictive speaker adaptation in speech recognition." *Computer Speech and Language* 9, 1-17.

[3] Gauvain, J. -L. & Lee, C. -H. (1994). "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains." *IEEE Transactions on Speech and Audio Processing* SAP-2(2), 291-298.

[4] Leggetter, C. J. & Woodland, P. C. (1995). "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models." *Computer Speech &Language* 9, 171-185.

[5] Padmanabhan, M., Bahl, L. R., Nahamoo, D., and Picheny, M. A., "Speaker clustering and transformation for speaker adaptation in large-vocabulary speech recognition systems." *Proceedings of ICASSP96.*

[6] Sen, A. & Srivastava, M. (1997). "Regression analysis: theory, methods, and applications." Springer-Verlag New York Inc.

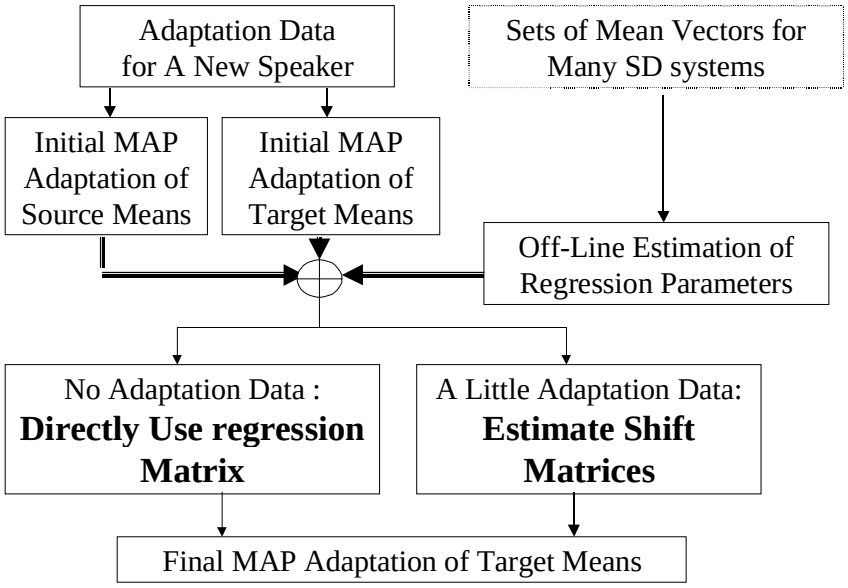

Fig.1 Diagram of our new adaptation approach