A Closed-Form Solution of Linear Spectral Transformation for Robust Speech Recognition

Donghyun Kim and Dongsuk Yook

ABSTRACT—The maximum likelihood linear spectral transformation (ML-LST) using a numerical iteration method has been previously proposed for robust speech recognition. The numerical iteration method is not appropriate for real-time applications due to its computational complexity. In order to reduce the computational cost, the objective function of the ML-LST is approximated and a closed-form solution is proposed in this paper. It is shown experimentally that the proposed closed-form solution for the ML-LST can provide rapid speaker and environment adaptation for robust speech recognition.

Keywords—Speech recognition, environment adaptation, linear spectral transformation, closed-form solution.

I. Introduction

Though speech recognition systems using hidden Markov models (HMMs) are trained with a large number of speech corpora, a well-trained acoustic model for various speakers and environments is difficult to obtain. A practical speech recognizer for real environment applications has to cope with noisy conditions and high computational cost. Therefore, many acoustic model adaptation techniques have been proposed to reduce the gap between the training and testing conditions without the necessity for retraining. Previously, the maximum likelihood linear spectral transformation (ML-LST) method was proposed for rapid adaptation [1], [2]. It is useful for handling additive and convolutional noise in order to provide rapid acoustic model adaptation. This method is a transformation-based adaptation technique like the maximum likelihood linear regression (MLLR) [3], but its transformation function parameters directly represent the additive and convolutional noise components in the linear spectral domain. ML-LST can adapt a clean acoustic model to a noisy environment, which contains additive and convolutional noise, more effectively than MLLR. Unlike the parallel model combination method [4], it estimates only the transformation function parameters without extracting the noise spectra or making noise models. However, since the ML-LST requires complicated domain conversion between the cepstra and linear spectra, it uses a numerical iteration method to estimate the optimal transformation function parameters. Because this procedure incurs a large amount of computation time, it is not appropriate for many real-time applications.

In this letter, we propose a closed-form ML-LST (C-ML-LST) method as an approximate solution for the iterative ML-LST. We found that the characteristics of the transformation function parameters, estimated using a linear spectral domain objective function, are similar to those obtained using the cepstral domain objective function. Therefore, the C-ML-LST directly estimates the transformation function parameters using an approximate linear spectral domain objective function without any numerical iteration procedure. The proposed method enables both rapid adaptation with a small amount of data and the real-time execution of the adaptation.

This letter is organized as follows. Section II explains the closed-form solution of the maximum likelihood linear spectral transformation. Section III evaluates the proposed method. Some conclusions are given in section IV.

Manuscript received Jan. 12, 2009; revised Apr. 12, 2009; accepted Apr. 21, 2009.

This work was supported by the Korea Research Foundation Grant funded by the Korean Government (KRF-2006-311-D00822). It was also supported by the MKE (Ministry of Knowledge Economy), Rep. of Korea, under the ITRC (Information Technology Research Center) support program supervised by the IITA (Institute for Information Technology Advancement) (IITA-2008-C1090-0803-0006).

Donghyun Kim (phone: +82 2 3290 3641, email: kaizer@voice.korea.ac.kr) and Dongsuk Yook (phone: +82 2 3290 3202, email: yook@voice.korea.ac.kr) are with Speech Information Processing Laboratory, Department of Computer and Communication Engineering, Korea University, Seoul, Rep. of Korea.

doi:10.4218/etrij.09.0209.0012

II. Closed-Form ML-LST

As discussed in [1], [2], and [5], the linear spectral noisy mean vector, $\tilde{\boldsymbol{\mu}}^s$, affected by some additive noise, such as background sounds, and convolutional noise, such as microphone differences, can be approximated by

$$\tilde{\boldsymbol{\mu}}^s = \mathbf{A}(\boldsymbol{\mu}^s + \mathbf{b}), \tag{1}$$

where $\mathbf{A}$ is a diagonal matrix representing the convolutional noise, $\boldsymbol{\mu}^s$ is the corresponding clean mean vector, and $\mathbf{b}$ is a vector representing the additive noise, all in the linear spectral domain. Once $\mathbf{A}$ and $\mathbf{b}$ are estimated, the adapted model parameters in the cepstral domain can be computed easily [1], [2], [5]. However, as speech recognition systems typically use acoustic models represented in the cepstral domain, we need to perform a domain conversion of the model parameters before we can make use of (1) to adapt the model parameters.

It is difficult to find a closed-form solution for $\mathbf{A}$ and $\mathbf{b}$ since ML-LST requires the complicated domain conversion that involves logarithmic and exponential operations. Therefore, instead of Baum's auxiliary function, which is designed to increase the likelihood in the cepstral domain, we propose a modified objective function whose purpose is to reduce the distance between the observation vectors and the model parameters in the linear spectral domain. This will help to avoid the complicated domain conversion and directly estimate the transformation function parameters $\mathbf{A}$ and $\mathbf{b}$. A modified version of Baum's auxiliary function (showing only optimization related terms) is

$$F = -\sum_t \sum_g \gamma_g^t [\log|\tilde{\boldsymbol{\Sigma}}_g^s| + (\mathbf{o}^{s,t} - \tilde{\boldsymbol{\mu}}_g^s)^T \tilde{\boldsymbol{\Sigma}}_g^{s-1} (\mathbf{o}^{s,t} - \tilde{\boldsymbol{\mu}}_g^s)], \tag{2}$$

where γ_g^t is the posterior probability of the Gaussian probability density function (PDF) g being used at time t. This modified version of Baum's auxiliary function uses the linear spectral observation vector $\mathbf{o}^s$ (adaptation data), linear spectral noisy mean vector $\tilde{\boldsymbol{\mu}}^s$, and linear spectral noisy covariance matrix $\tilde{\boldsymbol{\Sigma}}^s$, instead of the cepstral observations and cepstral model parameters. When (2) is optimized, the distance between the linear spectral noisy observations and the linear spectral noisy models is reduced, leading to an increase in the value of the original Baum's auxiliary function. To estimate the transformation function parameters using (2), we replace the mean vector by (1) and replace the covariance matrix by

$$\tilde{\boldsymbol{\Sigma}}^s = \mathbf{A}\boldsymbol{\Sigma}^s\mathbf{A}^T. \tag{3}$$

Then, the modified version of Baum's auxiliary function can be simplified as

$$\begin{aligned} F = -\sum_t \sum_g \gamma_g^t [&\log|\boldsymbol{\Sigma}_g^s| - 2\log|\mathbf{A}^{-1}| \\ &+ (\mathbf{A}^{-1}\mathbf{o}^{s,t} - \mathbf{b} - \boldsymbol{\mu}_g^s)^T \boldsymbol{\Sigma}_g^{s-1} (\mathbf{A}^{-1}\mathbf{o}^{s,t} - \mathbf{b} - \boldsymbol{\mu}_g^s)]. \end{aligned} \tag{4}$$

To obtain the optimal parameters, the derivative of F with respect to A_{kk}^{-1}, which is the k-th diagonal element of $\mathbf{A}^{-1}$, is obtained by

$$\frac{\partial F}{\partial A_{kk}^{-1}} = \sum_t \sum_g \gamma_g^t \left[\frac{-1}{A_{kk}^{-1}} + \frac{(A_{kk}^{-1} o_k^{s,t} - b_k - \mu_{g,k}^s) o_k^{s,t}}{\Sigma_{g,kk}^s} \right], \tag{5}$$

where $\Sigma_{g,kk}^s$ is the k-th diagonal element of $\boldsymbol{\Sigma}_g^s$. Similarly, the derivative of F with respect to b_k is obtained by

$$\frac{\partial F}{\partial b_k} = \sum_t \sum_g \gamma_g^t \left[\frac{(A_{kk}^{-1} o_k^{s,t} - b_k - \mu_{g,k}^s)}{\Sigma_{g,kk}^s} \right]. \tag{6}$$

By setting (5) and (6) to zero and solving A_{kk}^{-1} and b_k, the optimal transformation function parameters can be found as follows:

$$\begin{aligned} A_{kk}^{-1} = &\frac{1}{2} \frac{\Gamma_k(\sigma)\Gamma_k(\sigma\mu o) - \Gamma_k(\sigma\mu)\Gamma_k(\sigma o)}{\Gamma_k(\sigma o^2)\Gamma_k(\sigma) - \Gamma_k^2(\sigma o)} \\ &\pm \sqrt{\left(\frac{1}{2} \frac{\Gamma_k(\sigma)\Gamma_k(\sigma\mu o) - \Gamma_k(\sigma\mu)\Gamma_k(\sigma o)}{\Gamma_k(\sigma o^2)\Gamma_k(\sigma) - \Gamma_k^2(\sigma o)} \right)^2 + \frac{\Gamma_k(\sigma)\Gamma(1)}{\Gamma_k(\sigma o^2)\Gamma_k(\sigma) - \Gamma_k^2(\sigma o)}}, \end{aligned} \tag{7}$$

$$b_k = \frac{A_{kk}^{-1} \cdot \Gamma_k(\sigma o) - \Gamma_k(\sigma\mu)}{\Gamma_k(\sigma)}, \tag{8}$$

where

$$\Gamma(1) = \sum_t \sum_g \gamma_g^t, \tag{9}$$

$$\Gamma_k(\sigma) = \sum_t \sum_g \gamma_g^t \frac{1}{\Sigma_{g,kk}^s}, \tag{10}$$

$$\Gamma_k(\sigma\mu) = \sum_t \sum_g \gamma_g^t \frac{1}{\Sigma_{g,kk}^s} \mu_{g,k}^s, \tag{11}$$

$$\Gamma_k(\sigma o) = \sum_t \sum_g \gamma_g^t \frac{1}{\Sigma_{g,kk}^s} o_k^{s,t}, \tag{12}$$

$$\Gamma_k(\sigma o^2) = \sum_t \sum_g \gamma_g^t \frac{1}{\Sigma_{g,kk}^s} o_k^{s,t} o_k^{s,t}, \tag{13}$$

$$\Gamma_k(\sigma\mu o) = \sum_t \sum_g \frac{\gamma_g^t}{\Sigma_{g,kk}^s} \mu_{g,k}^s o_k^{s,t}. \tag{14}$$

Since A_{kk}^{-1} is positive, only the addition of the two terms in (7) yields a valid value.

III. Evaluation of the Closed-Form ML-LST

The baseline acoustic models used in the experiments were speaker-independent, tied-state, cross-word triphone HMMs. We used 41 monophone models including two silence models

to generate 15,571 cross-word triphone models. The decision-tree-based top-down state clustering method produced 850 tied states shared by the left-to-right 3-state cross-word triphone models. We used 10 Gaussian distributions per state, resulting in 8,500 Gaussian distributions in the baseline speech recognition system. We used 3,696 training utterances from the TIMIT corpus for training and 1,296 testing utterances from a noisy version of the far-field microphone TIMIT (FFMTIMIT) data for testing. When the training and testing environments were the same, the phone error rate (PER) was 26.4%. The PER increased to 47.0% when testing was performed with the noisy FFMTIMIT data without any adaptation.

While the 39-dimension mel-frequency cepstral coefficient (MFCC) feature was used for training and testing, only the first 13 dimensions were used for the adaptation in order to reduce the computational complexity. Figure 1 shows a comparison of the supervised adaptations via the ML-LST using a numerical iteration method, C-ML-LST, and MLLR, for increasing amounts of adaptation data. C-ML-LST does not perform as well as ML-LST using a numerical iteration method because it is an approximation of the latter. However, as the amount of the adaptation data is increased, the performance of the approximated ML-LST approaches that of the ML-LST using a numerical iteration method. When we compare C-ML-LST and MLLR, both of which can be run in real-time, C-ML-LST outperforms MLLR.

Since the conventional ML-LST requires several hundred numerical iterations for the convergence of the algorithm, it is not appropriate for real-time adaptation. In contrast, the proposed closed-form solution has only polynomial time complexity. Table 1 analyzes time complexity using big-O notation, where k is the number of dimensions for the observation vector, t is the number of observation vectors, and

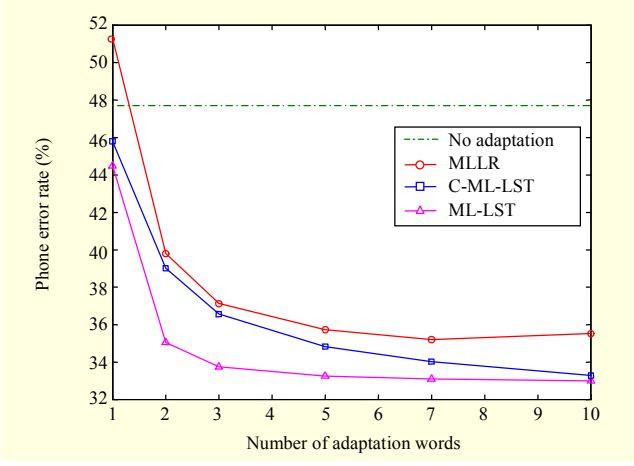

Fig. 1. Phone error rates (%) of the ML-LST using a numerical iteration method, C-ML-LST, and MLLR vs. the number of adaptation words.

Table 1. Computational time complexity.

	ML-LST	C-ML-LST	MLLR
Occupancy computation	$O(ktg)$	$O(ktg)$	$O(ktg)$
Parameter estimation	$O(nktg)$	$O(ktg)$	$O(ktg+k^2g+k^3)$
Model update	$O(nk^2g)$	$O(k^2g)$	$O(k^2g)$
Total	$O(nktg)$	$O(ktg)$	$\approx O(ktg)$
Real time factor	4.59	0.11	0.09

g is the number of Gaussian PDFs. Assuming that k is much smaller than both t and g, the closed-form solution of ML-LST and MLLR has the time complexity of $O(ktg)$, while the iterative ML-LST is n times slower than the closed-form solution, where n is the number of numerical iterations. The last row of the table shows the real-time factor of the adaptation methods. The closed-form ML-LST runs in 0.11 times of the real time on a 2 GHz Pentium machine.

IV. Conclusion

In this paper, we proposed a closed-form solution of the ML-LST adaptation algorithm, which uses an approximated form of Baum's auxiliary function. A closed-form parameter estimation formula was derived for real-time applications. It was shown experimentally that the C-ML-LST technique outperforms conventional adaptation methods such as MLLR.

References

[1] D. Kim and D. Yook, "Fast Channel Adaptation for Continuous Density HMMs Using Maximum Likelihood Spectral Transform," *IEE Electron. Lett.*, vol. 40, no. 10, 2004, pp. 632-633.

[2] D. Kim and D. Yook, "Robust Model Adaptation Using Mean and Variance Transformation in Linear Spectral Domain," *Lecture Notes in Computer Science*, vol. 3578, 2005, pp. 149-154.

[3] C. Leggetter and P. Woodland, "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models," *Computer Speech and Language*, vol. 9, 1995, pp. 171-185.

[4] M. Gales, *Model-Based Techniques for Noise Robust Speech Recognition*, Ph.D. Thesis, Cambridge University, 1995.

[5] D. Kim and D. Yook, "Linear Spectral Transformation for Robust Speech Recognition Using Maximum Mutual Information," *IEEE Signal Process. Lett.*, vol. 14, no. 7, 2007, pp. 496-499.