LETTER

Speaker Adaptation Based on PARAFAC2 of Transformation Matrices for Continuous Speech Recognition

Yongwon JEONG[†a)], ***Member***, **Sangjun LIM**[†], **Young Kuk KIM**[††], ***and*** **Hyung Soon KIM**[†], ***Nonmembers***

SUMMARY We present an acoustic model adaptation method where the transformation matrix for a new speaker is given by the product of bases and a weight matrix. The bases are built from the parallel factor analysis 2 (PARAFAC2) of training speakers' transformation matrices. We perform continuous speech recognition experiments using the WSJ0 corpus.

key words: *maximum likelihood linear regression, parallel factor analysis, PARAFAC2, speaker adaptation, speech recognition*

1. Introduction

In hidden Markov model (HMM) based speech recognition [1], speaker adaptation techniques are used to update an HMM system such that the updated model better captures the acoustic characteristics of target speakers. One class of speaker adaptation techniques is a transformation-based technique such as the maximum likelihood linear regression (MLLR) adaptation [2]. In the MLLR adaptation, the model for a new speaker is expressed as a linear transformation of a speaker-independent (SI) model, and the transformation is estimated by maximizing the likelihood of adaptation data. In the eigenspace-based MLLR (EMLLR) adaptation [3], the set of MLLR transformation matrices of training speakers is decomposed by principal component analysis (PCA) to build bases, and the transformation matrix for a new speaker is expressed as a linear combination of bases. Thus, the EMLLR adaptation is an application of the eigenvoice adaptation [4] in the transformation space. In the eigenvoice adaptation, training acoustic models are decomposed by PCA to obtain bases and the model for a new speaker is represented as a linearly weighted sum of bases. Our approach is closely related to the EMLLR adaptation, but bases are built from the parallel factor analysis 2 (PARAFAC2) [5], [6] of MLLR transformation matrices of training speakers. In our approach, the transformation matrix for a new speaker is expressed as a product of bases and a weight matrix. We derive the weight in a maximum likelihood (ML) criterion. We evaluate the performance of the proposed method on large vocabulary continuous speech recognition (LVCSR) experiments. In this letter, the adaptation of acoustic models to a new speaker is performed by updating the Gaussian mean parameters of output distributions among continuous density HMM (CDHMM) parameters.

The rest of this letter is organized as follows. Section 2 explains the MLLR adaptation and Sect. 3 explains the EMLLR adaptation. In Sect. 4, we present the proposed speaker adaptation method using PARAFAC2. Section 5 presents experiments and Sect. 6 concludes this work.

2. MLLR Adaptation

In the MLLR adaptation, the updated HMM mean vector for mixture component r ($r = 1, \cdots, R$) is given by

$$\hat{\boldsymbol{\mu}}_r = \mathbf{W}_r \boldsymbol{\xi}_r \tag{1}$$

where $\mathbf{W}_r$ denotes the transformation matrix, and $\boldsymbol{\xi}_r = [\omega\, \mu_1 \cdots \mu_D]^T$ the extended mean vector of an SI HMM corresponding to mixture component r (ω is the bias offset term: $\omega = 1$ to include an offset or $\omega = 0$ otherwise). Here, we use a single transformation matrix for all mixture components, thus we drop out the index r in the transformation matrix. Given adaptation data $\mathbf{O} = \{\mathbf{o}_1, \cdots, \mathbf{o}_T\}$, the $D \times (D + 1)$ transformation matrix is estimated in an ML criterion:

$$\hat{\mathbf{W}} = \arg\max_{\mathbf{W}} \sum_r \log p(\mathbf{O}|\mathbf{W}\boldsymbol{\xi}_r). \tag{2}$$

In finding the transformation matrix using the expectation-maximization (EM) algorithm [7], the auxiliary Q-function to be optimized is given as (discarding the terms that are independent of $\mathbf{W}$)

$$\begin{aligned} &Q(\mathbf{W}) \qquad\qquad (3)\\ &= -\frac{1}{2}\sum_{t=1}^{T}\sum_{r=1}^{R}\gamma_r(t)\,(\mathbf{o}_t - \mathbf{s}_r(\mathbf{W}))^T\boldsymbol{\Sigma}_r^{-1}(\mathbf{o}_t - \mathbf{s}_r(\mathbf{W})) \end{aligned}$$

where $\gamma_r(t)$ denotes the *a posteriori* probability of occupying mixture component r at t given $\mathbf{O}$, $\boldsymbol{\Sigma}_r$ the covariance matrix for Gaussian mixture component r of an SI HMM (which is a diagonal matrix in this letter), and $\mathbf{s}_r(\mathbf{W}) = \mathbf{W}\boldsymbol{\xi}_r$. Setting $\partial Q(\mathbf{W})/\partial \mathbf{W} = 0$ produces

$$\sum_{t=1}^{T}\sum_{r=1}^{R}\gamma_r(t)\,\boldsymbol{\Sigma}_r^{-1}\mathbf{o}_t\,\boldsymbol{\xi}_r^T = \sum_{t=1}^{T}\sum_{r=1}^{R}\gamma_r(t)\,\boldsymbol{\Sigma}_r^{-1}\mathbf{W}\,\boldsymbol{\xi}_r\boldsymbol{\xi}_r^T. \tag{4}$$

The above equation can be solved for $\mathbf{W}$ in the row-by-row fashion as [2].

Manuscript received April 12, 2013.

[†] The authors are with the School of Electrical Engineering, Pusan National University, Busan 609–735, Korea.

[††] The author is with CTO Group, AirPlug, Seoul 135–920, Korea.

a) E-mail: jeongy@pusan.ac.kr

DOI: 10.1587/transinf.E96.D.2152

3. Eigenspace-Based MLLR (EMLLR) Adaptation

The EMLLR [3] adaptation is closely related to our approach. In the EMLLR adaptation, PCA is applied to MLLR transformation matrices. Let $\{\mathbf{W}_1, \cdots, \mathbf{W}_S\}$ be transformation matrices of S training speakers. The transformation matrices are converted to vectors and let $\{\mathbf{w}_1, \cdots, \mathbf{w}_S\}$ be the set of vectorized transformation of training speakers. The set of training vectors is decomposed by PCA. The sample covariance matrix is given by

$$\mathbf{C}_\mathbf{w} = \frac{1}{S-1}\sum_{s=1}^{S}(\mathbf{w}_s - \bar{\mathbf{w}})(\mathbf{w}_s - \bar{\mathbf{w}})^T \tag{5}$$

$$\text{where} \quad \bar{\mathbf{w}} = \frac{1}{S}\sum_{s=1}^{S}\mathbf{w}_s.$$

The K dominant eigenvectors ($\boldsymbol{\phi}$'s) of the sample covariance matrix are obtained as the basis vectors and the transformation for a new speaker is expressed as

$$\hat{\mathbf{w}} = \mathbf{\Phi}_K\mathbf{x} + \bar{\mathbf{w}} \tag{6}$$

$$\text{where} \quad \mathbf{\Phi}_K = [\boldsymbol{\phi}_1 \cdots \boldsymbol{\phi}_K].$$

So, the updated model for mixture component r is given by

$$\hat{\boldsymbol{\mu}}_r = \mathrm{mat}(\mathbf{\Phi}_K\mathbf{x} + \bar{\mathbf{w}})\,\boldsymbol{\xi}_r \tag{7}$$

where $\mathrm{mat}(\cdot)$ denotes the matricization of a vector (i.e., conversion from $D \cdot (D+1) \times 1$ vector to $D \times (D+1)$ matrix). Given adaptation data $\mathbf{O}$, the $K \times 1$ weight $\mathbf{x}$ is estimated in an ML criterion:

$$\hat{\mathbf{x}} = \arg\max_{\mathbf{x}} \sum_r \log p(\mathbf{O}|\mathrm{mat}(\mathbf{\Phi}_K\mathbf{x} + \bar{\mathbf{w}})\,\boldsymbol{\xi}_r). \tag{8}$$

The weight can be computed by the EM algorithm. The auxiliary Q-function is given as

$$Q(\mathbf{x}) \tag{9}$$
$$= -\frac{1}{2}\sum_{t=1}^{T}\sum_{r=1}^{R}\gamma_r(t)\,(\mathbf{o}_t - \mathbf{s}_r(\mathbf{x}))^T\mathbf{\Sigma}_r^{-1}(\mathbf{o}_t - \mathbf{s}_r(\mathbf{x}))$$

where $\mathbf{s}_r(\mathbf{x}) = \mathrm{mat}(\mathbf{\Phi}_K\mathbf{x} + \bar{\mathbf{w}})\,\boldsymbol{\xi}_r$. Setting $\partial Q(\mathbf{x})/\partial\mathbf{x} = 0$ and solving for $\mathbf{x}$ yields

$$\hat{\mathbf{x}} = \left[\sum_{t=1}^{T}\sum_{r=1}^{R}\gamma_r(t)\,\mathbf{G}_r^T\mathbf{\Sigma}_r^{-1}\mathbf{G}_r\right]^{-1} \tag{10}$$
$$\times\left[\sum_{t=1}^{T}\sum_{r=1}^{R}\gamma_r(t)\,\mathbf{G}_r^T\mathbf{\Sigma}_r^{-1}(\mathbf{o}_t - \mathrm{mat}(\bar{\mathbf{w}})\,\boldsymbol{\xi}_r)\right]$$

where

$$\mathbf{G}_r(:,k) = \mathrm{mat}(\boldsymbol{\phi}_k)\,\boldsymbol{\xi}_r, \quad k = 1, \cdots, K \tag{11}$$

and $\mathbf{G}_r(:,k)$ denotes the kth column vector of $\mathbf{G}_r \in \mathbb{R}^{D \times K}$.

4. Speaker Adaptation Using the PARAFAC2 Model

In the previous section, transformation matrices are converted to vectors before decomposition. In our approach, transformation matrices are decomposed in their matrix forms by PARAFAC2. Given MLLR transformation matrices of training speakers, $\{\mathbf{W}_1, \cdots, \mathbf{W}_S\}$, the collection of centered transformation matrices $\{\tilde{\mathbf{W}}_s\}_{s=1}^S = \{\mathbf{W}_s - \bar{\mathbf{W}}\}_{s=1}^S$ where $\bar{\mathbf{W}} = (1/S)\sum_s \mathbf{W}_s$ is expressed as follows by PARAFAC2:

$$\tilde{\mathbf{W}}_s = \mathbf{F}_s\mathbf{H}_s\mathbf{A}^T + \mathbf{R}_s \tag{12}$$
$$\text{such that} \;\; \mathbf{F}_s^T\mathbf{F}_s = \mathbf{\Phi} \text{ (invariant matrix)}, \quad s = 1, \cdots, S$$

where $\mathbf{F}_s$ is a $D \times K$ matrix of factor scores for the row units, $\mathbf{H}_s$ a $K \times K$ weight matrix for $\tilde{\mathbf{W}}_s$, $\mathbf{A}$ a $(D+1) \times K$ matrix of weights for the column units, and $\mathbf{R}_s$ the residual matrix. The PARAFAC2 model is depicted in Fig. 1. If $\mathbf{F}_s = \mathbf{F}$ in Eq. (12), the PARAFAC2 model becomes the PARAFAC model [8]. In a least squares criterion, the model can be fitted to $\{\tilde{\mathbf{W}}_s\}$ by minimizing

$$\text{Error} = \sum_{s=1}^{S}\|\tilde{\mathbf{W}}_s - \mathbf{F}_s\mathbf{H}_s\mathbf{A}^T\|^2. \tag{13}$$

A representative algorithm for finding the components that minimize the squared error is the alternating least squares (ALS) [6]. The basic idea is to minimize Eq. (13) over each of $\mathbf{F}_s$, $\mathbf{H}_s$, and $\mathbf{A}$ alternatingly while the rest of components are fixed. When all components except one are fixed, the problem becomes a linear modeling problem so the component can be found by using singular value decomposition (SVD); please refer to [6] for more details.

We modify the PARAFAC2 model to our application by setting $\mathbf{B}_s = \mathbf{F}_s\mathbf{H}_s$ in Eq. (12):

$$\tilde{\mathbf{W}}_s = \mathbf{B}_s\mathbf{A}^T + \mathbf{R}_s \tag{14}$$

so that $\mathbf{B}_s$ becomes the speaker-dependent weight "matrix" and $\mathbf{A}$ becomes the matrix of bases which is common across training speakers. Based on Eq. (14), the transformation matrix for a new speaker is expressed as

$$\mathbf{W}_{\text{new}} = \mathbf{B}\mathbf{A}^T + \bar{\mathbf{W}} \tag{15}$$

and the updated model for mixture component r is given by

$$\hat{\boldsymbol{\mu}}_r = \mathbf{W}_{\text{new}}\,\boldsymbol{\xi}_r \tag{16}$$
$$= (\mathbf{B}\mathbf{A}^T + \bar{\mathbf{W}})\,\boldsymbol{\xi}_r$$

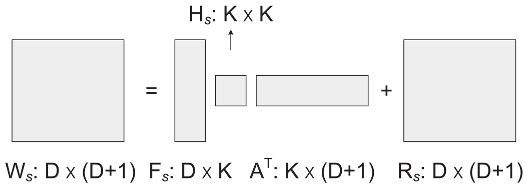

Fig. 1 PARAFAC2 model.

$$= \mathbf{B}\,\underbrace{\mathbf{A}^T\boldsymbol{\xi}_r}_{=\boldsymbol{\xi}_r^{\text{reduced}}} + \bar{\mathbf{W}}\,\boldsymbol{\xi}_r.$$

Given adaptation data **O**, we derive the $D \times K$ weight matrix **B** for a new speaker in an ML criterion. The weight can be obtained by the EM algorithm; the auxiliary function is given as (discarding the terms that are independent of **B**):

$$Q(\mathbf{B}) \tag{17}$$

$$= -\frac{1}{2}\sum_{t=1}^{T}\sum_{r=1}^{R}\gamma_r(t)\,(\mathbf{o}_t - \mathbf{s}_r(\mathbf{B}))^T\boldsymbol{\Sigma}_r^{-1}(\mathbf{o}_t - \mathbf{s}_r(\mathbf{B}))$$

where $\mathbf{s}_r(\mathbf{B}) = (\mathbf{B}\mathbf{A}^T + \bar{\mathbf{W}})\,\boldsymbol{\xi}_r$. Setting $\partial Q(\mathbf{B})/\partial\mathbf{B} = 0$ yields

$$\sum_{t=1}^{T}\sum_{r=1}^{R}\gamma_r(t)\,\boldsymbol{\Sigma}_r^{-1}(\mathbf{o}_t - \bar{\mathbf{W}}\boldsymbol{\xi}_r)\,\boldsymbol{\xi}_r^{\text{reduced}} \tag{18}$$

$$= \sum_{t=1}^{T}\sum_{r=1}^{R}\gamma_r(t)\,\boldsymbol{\Sigma}_r^{-1}\mathbf{B}\,\boldsymbol{\xi}_r^{\text{reduced}}\boldsymbol{\xi}_r^{\text{reduced}^T}.$$

The above equation can be solved for **B** using a similar technique in [2]. We define the left-side term in Eq. (18) as

$$\mathbf{Z} = \sum_{t=1}^{T}\sum_{r=1}^{R}\gamma_r(t)\,\boldsymbol{\Sigma}_r^{-1}(\mathbf{o}_t - \bar{\mathbf{W}}\boldsymbol{\xi}_r)\,\boldsymbol{\xi}_r^{\text{reduced}} \tag{19}$$

and we also define

$$g_{(j)}(p,q) = \sum_{r=1}^{R} v_r(j,j)\,d_r(p,q) \tag{20}$$

where $g_{(j)}(p,q)$ denotes the (p,q) element of $\mathbf{G}_{(j)}$, $v_r(j,j)$ the (j,j) element of $\mathbf{V}_r$, and $d_r(p,q)$ the (p,q) element of $\mathbf{D}_r$:

$$\mathbf{V}_r = \sum_{t=1}^{T}\gamma_r(t)\,\boldsymbol{\Sigma}_r^{-1} \tag{21}$$

$$\mathbf{D}_r = \boldsymbol{\xi}_r^{\text{reduced}}\boldsymbol{\xi}_r^{\text{reduced}^T}.$$

Then, the weight can be computed by

$$\hat{\mathbf{b}}_{(j)}^T = \mathbf{G}_{(j)}^{-1}\,\mathbf{z}_{(j)}^T, \quad j = 1,\cdots,D \tag{22}$$

where $\hat{\mathbf{b}}_{(j)}$ and $\mathbf{z}_{(j)}$ denote the jth row vectors of $\hat{\mathbf{B}}$ and **Z**, respectively. The obtained weight $\hat{\mathbf{B}}$ is plugged into Eq. (16) and utterances from the new speaker are recognized by the updated model.

5. Experiments

In experiments, we used the Wall Street Journal corpus WSJ0 with 5k vocabulary [9]. As the acoustic feature vector, the 39-D vector consisting of 13 mel-frequency cepstral coefficients (MFCCs) including the 0th coefficient, Δ coefficients, and Δ-Δ coefficients, was extracted from waveforms with the 20-ms Hamming window with the frame sliding of 10 ms. In the training phase, we used 7,138 utterances of 83 speakers form the standard SI-84 training set (the total training utterances amounted to about 14 h). With HMM toolkit (HTK), we built a tied-state triphone (cross-word triphone) model with 3,120 tied states and a mixture of 8 Gaussians. So, the number of mixture components $R = 3{,}120 \times 8 = 24{,}960$. From the SI HMM, we obtained a transformation matrix for each training speaker by the MLLR technique with 32 regression classes followed by the maximum *a posteriori* (MAP) adaptation [10]. These 83 transformation matrices were used to build bases for the EMLLR and the PARAFAC2-based model.

For the adaptation test, we used the adaptation data of 8 testing speakers from the WSJ0 corpus, i.e., the November 92 NIST evaluation set [11]. We used 1 to 5 utterances from the adaptation set (an adaptation utterance was about 6 s in length). The adaptation was performed in a supervised mode. We performed recognition test on 330 utterances from the testing set using the WSJ 5K non-verbalized 5k closed-vocabulary set and the language model of WSJ standard 5K non-verbalized closed bigram. Approximately 40 utterances were tested by updated models for each testing speaker.

Figure 2 shows results. Good performance is obtained by the PARAFAC2-based method with $K = 32$ (the number of adaptation parameters is 39 × 32), the EMLLR technique with $K = 70$ (the number of adaptation parameters is 70), and the MLLR adaptation with a diagonal regression matrix (the number of adaptation parameters is 39 × 2). The PARAFAC2-based method exhibits better performance than the EMLLR and MLLR adaptation methods for adaptation sentences ≥ 4. The EMLLR technique shows the best performance for adaptation sentences ≤ 2. Using bases (a kind of prior information about training speakers) contributes to better performance of the PARAFAC2-based method (for adaptation sentences ≥ 2) and the EMLLR technique (for adaptation sentences ≥ 1) over the MLLR adaptation. Because the dimension of the PARAFAC2-based method is given by $D \times K$, the amount of adaptation data needed for reliable estimation of weight should be larger than the amount

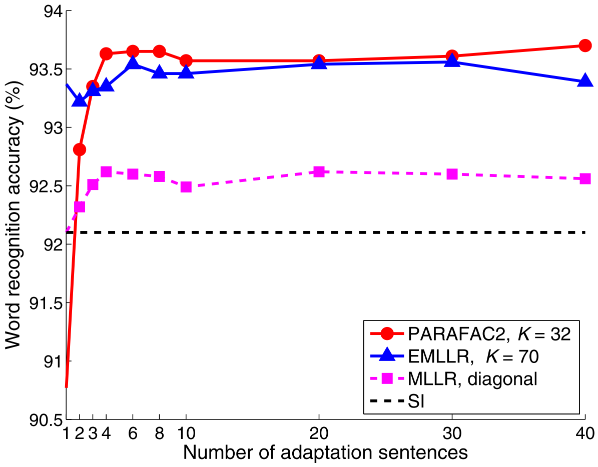

Fig. 2 Word recognition accuracy of adapted models. The word recognition accuracy of the SI HMM is 92.1%.

needed in the EMLLR technique where the dimension of the weight is given by K. We think this is the reason for which the PARAFAC2-based method exhibits worse performance than the EMLLR technique for adaptation sentences ≤ 2 and better performance than the EMLLR technique for adaptation sentences ≥ 4. Moreover, we think that the good performance of the PARAFAC2-based model for large amounts of adaptation data is due to its bases built by PARAFAC2 where the structure of transformation "matrix" is preserved during decomposition whereas the matrix structure is lost during decomposition by PCA in the EMLLR technique.

6. Conclusions

In this letter, we presented a basis-based speaker adaptation method in the MLLR framework. In our approach, the transformation matrices of training speakers in matrix form are decomposed by PARAFAC2 to build bases. In continuous speech recognition experiments, the proposed method outperforms the MLLR and EMLLR adaptation techniques for adaptation data ≥ 24 s.

References

[1] L.R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," Proc. IEEE, vol.77, no.2, pp.257–286, Feb. 1989.

[2] 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, no.2, pp.171–185, April 1995.

[3] K.T. Chen, W.W. Liau, H.M. Wang, and L.S. Lee, "Fast speaker adaptation using eigenspace-based maximum likelihood linear regression," Proc. ICSLP, vol.3, pp.742–745, 2000.

[4] R. Kuhn, J.-C. Junqua, P. Nguyen, and N. Niedzielski, "Rapid speaker adaptation in eigenvoice space," IEEE Trans. Speech Audio Process., vol.8, no.6, pp.695–707, Nov. 2000.

[5] R.A. Harshman, "PARAFAC2: Mathematical and technical notes," UCLA Working Papers in Phonetics, vol.22, pp.30–44, 1972.

[6] H.A.L. Kiers, J.M.F. ten Berge, and R. Bro, "PARAFAC2 - Part I. A direct fitting algorithm for the PARAFAC2 model," J. Chemometr., vol.13, no.3–4, pp.275–294, July 1999.

[7] A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," J. R. Stat. Soc. Ser. B-Stat. Methodol., vol.39, no.1, pp.1–38, 1977.

[8] R.A. Harshman, "Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multimodal factor analysis," UCLA Working Papers in Phonetics, vol.16, pp.1–84, 1970.

[9] D.B. Paul and J.M. Baker, "The design for the Wall Street Journal-based CSR corpus," Proc. DARPA Speech and Natural Language Workshop, pp.357–362, 1992.

[10] J.-L. Gauvain and C.-H. Lee, "Maximum *a posteriori* estimation for multivariate Gaussian mixture observations of Markov chains," IEEE Trans. Speech Audio Process., vol.2, no.2, pp.291–298, April 1994.

[11] D.S. Pallett, J.G. Fiscus, W.M. Fisher, J.S. Garofolo, B.A. Lund, and M.A. Przybocki, "1993 benchmark tests for the ARPA spoken language program," Proc. Workshop on Human Language Technology, Association for Computational Linguistics, pp.49–74, 1994.