A NEW METHOD FOR SPEAKER ADAPTATION USING BILINEAR MODEL

Hwa Jeon Song, Yongwon Jeong, and Hyung Soon Kim

Department of Electronics Engineering
Pusan National University
Geumjeong-gu Jangjeon 2-dong, Busan 609-735, Korea.
E-mail: {hwajeon, jeongy, kimhs}@pusan.ac.kr

ABSTRACT

In this paper, a novel method for speaker adaptation using bilinear model is proposed. Bilinear model can express both characteristics of speakers (style) and phonemes across speakers (content) independently in a training database. The mapping from each speaker and phoneme space to observation space is carried out using bilinear mapping matrix which is independent of speaker and phoneme space. We apply the bilinear model to speaker adaption. Using adaptation data from a new speaker, speaker-adapted model is built by estimating the style(speaker)-specific matrix. Experimental results showed that the proposed method outperformed eigenvoice and MLLR. In vocabulary-independent isolated word recognition for speaker adaptation, bilinear model reduced word error rate by about 38% and about 10% compared to eigenvoice and MLLR respectively using 50 words for adaptation.

***Index Terms*—** Speaker adaptation, Bilinear model, Eigenvoice, maximum likelihood linear regression (MLLR)

1. INTRODUCTION

Speech is composed of phonetic units and these units are uttered differently depending on speakers. Therefore, we can naturally assume that speech has two independent factors if 'style' is defined as the way of speaking (i.e. how a person speaks) and 'content' is defined as phonetic unit (i.e. what is spoken). Same words spoken by different persons in various conditions can be perceived consistently by another person because human being can recognize learned words (i.e. contents) in various conditions (i.e. styles) such as speed and accent using obtained knowledge. This is an example of recognizing the contents in different styles. Likewise, recognizing a person's unique way of speaking regardless of what is spoken is an example of recognizing the style in different contents. Assuming that the style is consistent within a person, we obtain two factor problem of recognizing style and/or content.

Factoring out the two independent variations in a training database and expressing them into a model can be useful in various applications. For example, the performance of a speech recognizer can be improved when regularizing styles across different speakers. Another possible application is speaker identification/verification where the style of a speaker can be estimated using the model which captures the various styles of the speakers in a training database.

Even speech recognizer using speaker-independent (SI) model tries to minimize the difference across the speakers in a training database for better performance. For a specific speaker, speaker-dependent (SD) system performs better than SI system but SD model is usually impractical or requires users' training session which is tedious. Therefore, to improve the performance of an SI system, speaker adaptation method is used with relatively small amount of adaptation data from the new speaker. Speaker adaptation methods can generally be categorized into maximum a posteriori (MAP) [1], maximum likelihood linear regression (MLLR) [2], and speaker clustering [3]. These methods either use SI models to build adapted models for new speakers or build adapted models as a linear combination of clustered speakers in the training database. The two factors inherent in the training database - style and content - are not used in these methods.

In this paper, a novel approach to speaker adaptation is proposed using bilinear model which can capture these two inherent variations - within a speaker and across speakers. Bilinear model with two independent control parameters is built from a training database with two types of variation. Then, this generic model is adapted to a new speaker by estimating the style of the new speaker using small amount of adaptation data while maintaining the content factor. Using bilinear model, style and content of a speaker can be efficiently expressed in terms of style/content space and mapping matrix that transfers the style and content space into observation space.

This paper is organized as follows. In the next section, bilinear model is briefly described. The proposed method for speaker adaptation using bilinear model is introduced in Section 3. In Section 4, experiments for the performance evaluation and comparison to other methods are discussed and in Section 5, we concludes the results and future work is mentioned.

2. BILINEAR MODEL

Bilinear models can separate two variations in a set of observations and express these two independent factors. In a speech, the two factors can be the way of speaking and the spoken phonetic unit. The style and content are interchangeable but they have natural association in this case; the style is the way of speaking, i.e. characteristics of a speaker and the content is what is spoken. There are two types in bilinear models - symmetric and asymmetric. These two models are briefly explained in the following. Refer [4] for more details.

2.1. Symmetric bilinear model

A D-dimensional observation vector $\mathbf{o}^{sc}$ can be expressed by using bilinear model with style parameter $\mathbf{a}^{s}$ and content parameter $\mathbf{b}^{c}$ as

This research was performed for the Intelligent Robotics Development Program, one of the 21st Century Frontier R&D Programs funded by the Ministry of Knowledge Economy of Korea.

follows :

$$\mathbf{o}^{sc} = \sum_{i,j} \mathbf{w}_{ij} a_i^s b_j^c \tag{1}$$

where $\mathbf{a}^s$ is an I-dimensional vector whose elements are a_i^s, $\mathbf{b}^c$ is a J-dimensional vector whose elements are b_j^c, and $\mathbf{w}_{ij}$ is a D-dimensional vector composed of w_{ijd}. The d-th observation vector can be expressed as follows using w_{ijd}

$$o_d^{sc} = \sum_{i=1}^{I} \sum_{j=1}^{J} w_{ijd} a_i^s b_j^c, \ 1 \leq d \leq D. \tag{2}$$

(2) can be expressed in vector-matrix form as

$$o_d^{sc} = \mathbf{a}^{s^T} \mathbf{W}_d \mathbf{b}^c. \tag{3}$$

Here, there is $\mathbf{W}_d \in \mathbb{R}^{I \times J}$ for each dimension and there are D matrices. These matrices map the style space spanned by $\mathbf{a}^s$ and the content space spanned by $\mathbf{b}^c$ into D-dimensional observation space. $\mathbf{W}_d$ is independent of both $\mathbf{a}^s$ and $\mathbf{b}^c$ and expresses the interaction between the two factors.

2.2. Asymmetric bilinear model

Often, there are cases where a new style cannot be expressed accurately as a linear combination of trained basis styles. In these cases, asymmetric bilinear model can change the interaction term w_{ijd} according to a style and this is more flexible than symmetric bilinear model. (2) can be modified as follows to introduce style-dependent mapping matrices:

$$o_d^{sc} = \sum_{i=1}^{I} \sum_{j=1}^{J} w_{ijd}^s a_i^s b_j^c. \tag{4}$$

Defining a style-specific term $a_{jd}^s = \sum_{i=1}^{I} w_{ijd}^s a_i^s$, (4) becomes $o_d^{sc} = \sum_{j=1}^{J} a_{jd}^s b_j^c$ and can be expressed as

$$\mathbf{o}^{sc} = \sum_{j} \mathbf{a}_j^s b_j^c \tag{5}$$

where $\mathbf{a}_j^s$ is a D-dimensional vector whose elements are a_{jd}^s. In vector-matrix form,

$$\mathbf{o}^{sc} = \mathbf{A}^s \mathbf{b}^c. \tag{6}$$

Here, $\mathbf{A}^s$ is a $D \times J$ matrix whose elements are $\mathbf{a}_j^s$ and expresses the style-specific linear mapping from content space into observation space.

2.3. Bilinear model building

How a bilinear model is built from a training database is described in this section. The methods are different depending on the type of bilinear model used (symmetric or asymmetric).

2.3.1. Asymmetric bilinear model building

The model building of asymmetric bilinear model is simpler than symmetric bilinear model. Let $\mathbf{o}(t) \in \mathbb{R}^D$ be the t-th observation vector in a training set. There are T training vectors so the training set is $\{\mathbf{o}(1), \mathbf{o}(2), \cdots, \mathbf{o}(t), \cdots, \mathbf{o}(T)\}$. In model building, one tries to find the parameters of the model that minimize the following error:

$$E = \sum_{t=1}^{T} \sum_{s=1}^{S} \sum_{c=1}^{C} \gamma^{sc}(t) \| \mathbf{o}(t) - \mathbf{A}^s \mathbf{b}^c \|^2, \tag{7}$$

$$\gamma^{sc}(t) = \begin{cases} 1, & \text{when } \mathbf{o}(t) \in (s, c) \\ 0, & \text{otherwise.} \end{cases}$$

When training database has the same numbers of observations for each style and content classes, there exists a closed-form solution to the problem (7) using the singular value decomposition (SVD) [4]. In other cases, a direct minimization method can be used such as quasi-Newton method. Here, we explain the case where SVD-based method can be used. Mean observation vector for each style and content classes is defined as

$$\overline{\mathbf{m}}^{sc} = \frac{\sum_t \gamma^{sc}(t) \mathbf{o}(t)}{\sum_t \gamma^{sc}(t)} \tag{8}$$

To use a standard matrix algorithm, observation matrix is arranged as a $(SD) \times C$ matrix:

$$\overline{\mathbf{M}} = \begin{bmatrix} \overline{\mathbf{m}}^{11} & \cdots & \overline{\mathbf{m}}^{1C} \\ \vdots & \ddots & \vdots \\ \overline{\mathbf{m}}^{S1} & \cdots & \overline{\mathbf{m}}^{SC} \end{bmatrix}. \tag{9}$$

Matrix $\overline{\mathbf{M}}$ can be decomposed and expressed for asymmetric bilinear model as

$$\overline{\mathbf{M}} = \mathbf{AB}. \tag{10}$$

Here, the stacked style parameter $\mathbf{A} \in \mathbb{R}^{(SD) \times J}$ is defined as

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}^1 \\ \vdots \\ \mathbf{A}^s \\ \vdots \\ \mathbf{A}^S \end{bmatrix} \tag{11}$$

where $\mathbf{A}^s \in \mathbb{R}^{D \times J}$ denotes the s-th style-specific matrix. Also, stacked content parameter $\mathbf{B} \in \mathbb{R}^{J \times C}$ is defined as

$$\mathbf{B} = \left[\mathbf{b}^1 \cdots \mathbf{b}^C \right]. \tag{12}$$

To find the optimal style and content parameters, $\mathbf{A}$ and $\mathbf{B}$, SVD is applied on $\overline{\mathbf{M}}$ to produce $\overline{\mathbf{M}} = \mathbf{USV}^T$ where $\mathbf{S}$ is a diagonal matrix whose elements are singular values arranged in descending order. Then, the style parameter $\mathbf{A}$ is defined as the first J columns of $\mathbf{US}$ and the content parameter $\mathbf{B}$ is defined as the first J rows of $\mathbf{V}^T$.

2.3.2. Symmetric bilinear model building

The objective function to be minimized in building a symmetric bilinear model is defined as

$$E = \sum_{t=1}^{T} \sum_{s=1}^{S} \sum_{c=1}^{C} \sum_{d=1}^{D} \gamma^{sc}(t) \| o_d(t) - \mathbf{a}^{s^T} \mathbf{W}_d \mathbf{b}^c \|^2 \tag{13}$$

Like in the asymmetric bilinear model, optimal $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{W}$ are estimated from $(SD) \times C$ observation matrix $\overline{\mathbf{M}}$ obtained from the training database. But, there is no closed-form solution for (13) in symmetric bilinear model. So, the optimal model parameters are obtained in an iterative way. The basic idea is to express the symmetric bilinear model as a asymmetric bilinear model and switch the roles of style and content parameters in each iteration until $\mathbf{A}$ and $\mathbf{B}$ converge. Iterative procedure for estimating the optimal values of $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{W}$ proceeds as follows:

1. Find $\mathbf{B}$ using SVD as in the asymmetric bilinear model.
2. Apply SVD to obtain $\left[\overline{\mathbf{M}}\mathbf{B}^T\right]^{VT} = \mathbf{W}^{VT}\mathbf{A} = \mathbf{USV}^T$. Take $\mathbf{A}$ matrix as $\mathbf{V}^T$.
3. Apply SVD to obtain $\left[\overline{\mathbf{M}}^{VT}\mathbf{A}^T\right]^{VT} = \mathbf{WB} = \mathbf{USV}^T$. Take $\mathbf{B}$ as $\mathbf{V}^T$.
4. Iterate 2 and 3 until $\mathbf{A}$ and $\mathbf{B}$ converge.
5. Upon convergence, $\mathbf{W}$ is found as $\mathbf{W} = \left[[\overline{\mathbf{M}}\mathbf{B}^T]^{VT}\mathbf{A}^T\right]^{VT}$.

In the procedure, $[\cdot]^{VT}$ denotes vector transpose of a matrix. The convergence of above procedure is guaranteed [5]. For more detailed description of the bilinear model building, refer [4].

3. SPEAKER ADAPTATION USING BILINEAR MODEL

Now, how the bilinear model is applied to speaker adaptation is described. We adapt a generic bilinear model to a new speaker by estimating and updating the style factor of the new speaker based on content basis vector which is common across all speakers in the training database. A small amount of utterances from the new speaker is used for speaker adaptation. This is an 'extrapolation' problem in bilinear models [4]. In this paper, asymmetric bilinear model is used to estimate the style-specific matrix (i.e. speaker-adapted model) for a new speaker.

To build a bilinear model, the observation matrix has to be built form the training database composed of S speakers. Each SD model is built as C Gaussian mixtures whose dimension is D. Here, we take the mean vectors of the Gaussians from the SD models and these mean vectors constitute the observation matrix. For example, s-th element of the observation matrix is expressed as $\mathbf{M}^s = [\boldsymbol{\mu}_1^s \cdots \boldsymbol{\mu}_c^s \cdots \boldsymbol{\mu}_C^s], 1 \leq c \leq C$. The observation matrix (9) is then expressed as

$$\overline{\mathbf{M}}_A = \begin{bmatrix} \mathbf{M}^1 \\ \vdots \\ \mathbf{M}^s \\ \vdots \\ \mathbf{M}^S \end{bmatrix}, \; 1 \leq s \leq S. \tag{14}$$

The dimension of the matrix $\overline{\mathbf{M}}_A$ is $(SD) \times C$. Next, $\overline{\mathbf{M}}_A$ is decomposed as $\mathbf{USV}^T$ by SVD. And then, $\mathbf{A}$ and $\mathbf{B}$ are assigned in the same way as aforementioned in section 2.3.1, respectively.

In adaptation session, using the adaptation data from a new speaker, the speaker-adapted model of the new speaker can be expressed as

$$\hat{\boldsymbol{\mu}}_c^{\tilde{s}} = \hat{\mathbf{A}}^{\tilde{s}}\mathbf{b}^c \tag{15}$$

where $\hat{\mathbf{A}}^{\tilde{s}} \in \mathbb{R}^{D \times J}$ is the estimated style-specific matrix for the new speaker $\tilde{s}$ and $\mathbf{b}^c$ is a content basis vector in $\mathbf{B} = \left[\mathbf{b}^1 \cdots \mathbf{b}^C\right]$. Note that $\mathbf{b}^c$ is fixed during the adaptation. $\hat{\mathbf{A}}^{\tilde{s}}$ can be estimated using the adaptation data $\{\mathbf{o}'(1), \cdots, \mathbf{o}'(t), \cdots, \mathbf{o}'(T')\}$ ($\mathbf{o}'(t)$ is a D-dimensional observation vector). $\hat{\mathbf{A}}^{\tilde{s}}$ is estimated by minimizing the following error:

$$E^* = \sum_{t=1}^{T'}\sum_{c=1}^{C}\gamma^{\tilde{s}c}(t)\|\mathbf{o}'(t) - \mathbf{A}^{\tilde{s}}\mathbf{b}^c\|^2. \tag{16}$$

$\mathbf{A}^{\tilde{s}}$ that minimizes the total squared error for the given adaptation data can be found by setting

$$\partial E^*/\partial \mathbf{A}^{\tilde{s}} = 0. \tag{17}$$

After some manipulation, we get

$$\hat{\mathbf{A}}^{\tilde{s}} = \left[\sum_{t=1}^{T'}\sum_{c}\gamma^{\tilde{s}c}(t)\mathbf{o}'(t)\mathbf{b}^{c^T}\right]\left[\sum_{t=1}^{T'}\sum_{c}\gamma^{\tilde{s}c}(t)\mathbf{b}^c\mathbf{b}^{c^T}\right]^{-1}. \tag{18}$$

EM (expectation maximization) algorithm can also be used to estimate $\mathbf{A}^{\tilde{s}}$. In that case, $\mathbf{A}^{\tilde{s}}$ can be estimated as the same way to obtain transform matrix in MLLR [2]. The model adapted to the new speaker is built when applying (15) for all content basis vectors using the estimated $\mathbf{A}^{\tilde{s}}$.

Bilinear model has the similar approach as eigenvoice in that it builds eigenvectors from a training database and estimates weights of the eigenvectors when adapting to a new speaker. But, while each speaker is expressed as a point in the speaker space in eigenvoice, two spaces - style and content vector spaces - are used and these two spaces are connected through bilinear mapping function in bilinear model. This difference can be seen from the composition of the observation matrix $\overline{\mathbf{M}}_A$ in (14) that comes from the training database. Also, eigenvectors obtained from this observation matrix are different from those of eigenvoice. In eigenvoice, the observation matrix is arranged as $(CD) \times S$ and eigenvectors (which are called eigenvoices) are obtained. Therefore, when using the same number of eigenvectors in bilinear model, the size of content basis vectors is reduced by the factor of D compared to eigenvoice so it has an advantage over eigenvoice in memory requirement. Notably, asymmetric bilinear model can be viewed as a generalization of MLLR methods. That is, an asymmetric bilinear model becomes MLLR when D-dimensional (which is the dimension of an observation vector) SI model is replaced with J-dimensional content basis vector. Therefore, depending on the number of eigenvectors, bilinear models show the properties of either eigenvoice or MLLR.

4. EXPERIMENTS

4.1. Experimental setup

We applied the proposed method to vocabulary-independent isolated word recognition for speaker adaptation and compared the results with eigenvoice and MLLR. For the building of the models, training database from 40 male speakers in Korean phonetically optimized words (POW) database was used [6]. For feature vector, 36-dimensional vector composed of 12-dimensional Mel-frequency cepstral coefficients (MFCCs), their delta coefficients, and their delta-delta coefficients was obtained using 20ms Hamming window sliding every 10ms. For acoustic unit, 46 phoneme like units (PLUs) were used. Triphones were used as the basic phonetic unit using tree-based clustering (TBC) on state level, with 3 states per model, and 1 mixture per state. There were 4050 tied-states using TBC [7].

After building SI HMM model using the training database of 40 male speakers, MLLR + MAP was applied for each speaker, resulting in 40 SD models. Using the 40 SD models as elements of the observation matrix in (14), asymmetric bilinear model composed of (11) and (12) was built by using the SVD. Here, $S = 40$, $D = 36$ and $C = 4050$ in (14).

For speaker adaptation and evaluation, a part of Korean phonetically balanced words (PBW) set was used which contains 452 words [8]. The database is composed of different speakers and recorded under different environment from the training database. Data from 10 male speakers was used for adaptation and testing. For each speaker, 1 to 50 words were used for adaptation. Then, the rest 400 words were used for evaluation. The adaptation was carried out in supervised mode for the experiments [7].

4.2. Experimental results

The results for speaker adaptation using eigenvoice, MLLR, and bilinear model are shown in Figure 1. We used (18) to estimate a style-specific matrix of a new speaker in the proposed method while estimation formula of MLLR and eigenvoice were based on EM. We used the global transformation in MLLR. The word accuracy of SI model was 95.78% as the baseline. It can be seen from the figure that eigenvoice performs well when the amount of adaptation data is small since it has the smallest number of model parameters. As amount of adaptation data increases, the performances of MLLR and the proposed method improves. Especially, using fewer number of parameters, the proposed method shows better performance than MLLR.

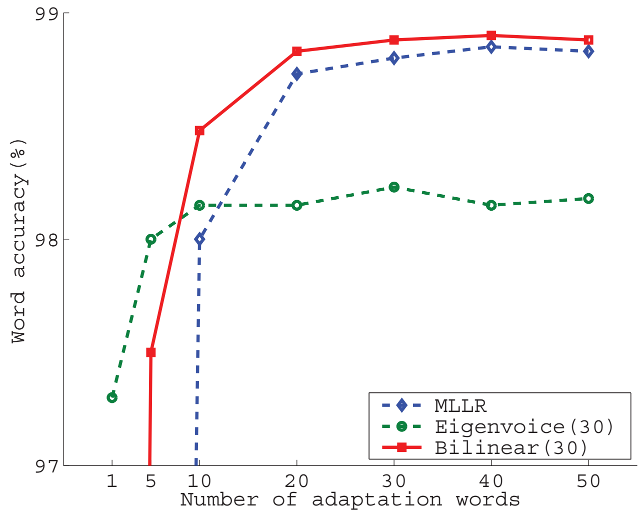

Fig. 1. Performance of the several adaptation methods. The number in the parenthesis indicates the number of eigenvectors used.

In Table 1, the performance of the bilinear model in speaker adaptation for various numbers of eigenvectors is shown. We can see from the table that if the number of eigenvectors is determined appropriately depending on the amount of adaptation data available, bilinear model can be efficiently used for rapid speaker adaptation. Therefore, bilinear model is more versatile than eigenvoice and MLLR.

Table 1. Word accuracy (%) of the bilinear model in speaker adaptation using different number of eigenvectors.

Number of adaptation words	Number of eigenvectors					
	1	5	10	20	30	40
1	97.43	96.80	95.53	89.78	82.05	77.75
5	97.45	98.20	98.25	98.23	97.50	96.90
10	97.58	98.10	98.33	98.63	98.48	98.38
20	97.60	98.25	98.58	98.78	98.83	98.95
30	97.68	98.23	98.55	98.75	98.88	98.95
40	97.70	98.15	98.65	98.83	98.90	98.90
50	97.65	98.15	98.65	98.88	98.88	98.95

The proposed method has the memory advantage over eigenvoice when same number of eigenvectors are used, by the factor of D due to the difference in the forms of supervectors in eigenvoice and bilinear model. Also, when only one eigenvector is used in the bilinear model, the number of parameters to be estimated becomes the dimension of an observation vector so style-specific matrix can be reliably estimated using small amount of adaptation data. The eigenvectors in a bilinear model can be taken as many as content basis vectors so more detailed model is possible by taking more eigenvectors when more adaptation data is available. In this view, the proposed method can be regarded as a combination of eigenvoice and MLLR.

5. CONCLUSION

In this paper, a new speaker adaptation method using bilinear model was proposed. By adjusting the number of basis vectors, it performs well for small amount of adaptation data and the performance improves as more adaptation data is available in the speaker adaptation experiments. It was shown to outperform both eigenvoice and MLLR from the experiments. The proposed model can be viewed as a generalization of MLLR and it has the advantage of eigenvoice that the parameters to be estimated can be adjusted depending on the size of adaptation data, which has a potential for rapid speaker adaptation. Future work includes the noise compensation using bilinear model with which the performance of a speech recognizer can be improved in a new noise environment.

6. REFERENCES

[1] C.-H. Lee, C.-H. Lin, and B.-H. Juang, "A study on speaker adaptation of the parameters of continuous density hidden Markov models," *IEEE Transactions on Signal Processing*, vol. 39, no. 4, pp. 806–814, April 1991.

[2] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, vol. 9, no. 2, pp. 171–185, April 1995.

[3] R. Kuhn, P. Nguyen, J.-C. Junqua, L. Goldwasser, N. Niedzielski, S. Fincke, K. Field, and M. Contolini, "Eigenvoices for speaker adaptation," in *Proceedings of the 5th International Conference on Spoken Language Processing*, 1998, vol. 5, pp. 1771–1774.

[4] J. B. Tenenbaum and W. T. Freeman, "Separating style and content with bilinear models," *Neural Computation*, vol. 12, no. 6, pp. 1247–1283, 2000.

[5] J. R. Magnus and H. Neudecker, *Matrix differential calculus with applications in statistics and econometrics*, John Wiley & Sons, 2nd edition, March 1999.

[6] Y. Lim and Y. Lee, "Implementation of the POW (phonetically optimized words) algorithm for speech database," in *International Conference on Acoustics, Speech, and Signal Processing*, May 1995, vol. 1, pp. 89–92.

[7] H. J. Song and H. S. Kim, "Simultaneous estimation of weights of eigenvoices and bias compensation vector for rapid speaker adaptation," in *Proceedings of the 8th International Conference on Spoken Language Processing*, 2004, pp. 2945–2948.

[8] Y.-J Lee, B.-W. Kim, J.-J Kim, O.-Y. Yang, and S.-Y. Lim, "Some considerations for construction of PBW set," in *Proceeding of the 12th Workshop on Speech Communications and Signal Processing*. Acoustical Society of Korea, June 1995, pp. 310–314.