Front. Electr. Electron. Eng. China (2006) 2: 130–134
DOI 10.1007/s 11460-006-0025-0

RESEARCH ARTICLE

LUO Jun, OU Zhi-jian, WANG Zuo-ying

Eigenvoice-based MAP adaptation within correlation subspace

Abstract In recent years, the eigenvoice approach has proven to be an efficient method for rapid speaker adaptation, which directs the adaptation according to the analysis of full speaker vector space. In this article, we developed a new algorithm for eigenspace-based adaptation restricting eigenvoices in clustered subspaces, and maximum likelihood (ML) criterion was replaced with maximum aposteriori (MAP) criterion for better parameter estimation. Experiments show that even with one sentence adaptation data this algorithm would result in 6.45 % error ratio reduction relatively, which overcomes the instability of maximum likelihood linear regression (MLLR) with limited data and is much faster than traditional MAP method. This algorithm is not highly-dependent on subspace number of division, thus it proved to be a robust adaptation algorithm.

Keywords Information processing, Fast speaker adaptation, Eigenvoices, Maximum likelihood, Maximum aposteriori, Correlation subspace

1 Introduction

Speaker adaptation is regarded as one of the important techniques in speech recognition. A speaker-dependent (SD) system always outperforms a speaker-independent (SI) one. However, it requires a large amount of speaker-specific data and the performance is not guaranteed when the data is lacking, so the key issue of adaptation is to robustly estimate model parameters with the limited data. As a result, a speaker-adapted (SA) model is obtained [1].

Dominant speaker adaptation techniques can be classified into three categories: the maximum a-posteriori (MAP) [1-2] adaptation approaches, transformation-based adaptation approaches such as maximum likelihood linear regression (MLLR) [3], and the approaches related to speaker clustering such as eigenvoice adaptation. Recently, the eigenvoice method has proven to be an efficient adaptation algorithm, which is not only successfully applied to isolated word recognition [4] but also used in large vocabulary continuous speech recognition (LVCSR) [5].

In the traditional eigenvoice method, the estimation is performed in full speaker vector space, while in this paper it is based on the analysis of independent subspaces. In addition, the traditional maximum likelihood criterion is replaced by maximum a-posterior criterion aims to make the extimation more robust. Experimental results show that even with a one sentence speech data it could gain considerable improvements, which outperforms both MAP and MLLR methods.

Translated from *Journal of Tsinghua University* (*Science and Technology*), 2004, 44(6):829-832 (in Chinese)

LUO Jun(✉), OU Zhi-jian, WANG Zuo-ying
Department of Electronic Engineering, Tsinghua University, Beijing 100084, China
E-mail: luojun@thsp.ee.tsinghua.edu.cn

2 ML adaptation with eigenvoices

The eigenvoice approach begins with a reference set of well-trained SD models. For each of the SD model, a supervector is formed by concatenating all of the mean vector parameters (supposed to be D-dimensional). After that, a dimensionality reduction technique called principal component analysis (PCA) [6] is used to find the eigenvectors. The supervector for a new speaker is assumed to be a linear combination of selected eigenvectors.

Denote $\boldsymbol{C}$ as the sample covariance matrix of the supervectors, then by PCA method, C can be expressed as:

$$\boldsymbol{C} = [\boldsymbol{e}(1), \cdots, \boldsymbol{e}(D)]\ \mathrm{diag}\ (\lambda_1, \cdots, \lambda_D)\ [\boldsymbol{e}(1), \cdots, \boldsymbol{e}(D)]^{\mathrm{T}} \quad (1)$$

where $\boldsymbol{e}(i)$, $i = 1, 2, \cdots, D$ are eigenvectors and λ_i are the corresponding eigenvalues of $\boldsymbol{C}$ in descending order ($\lambda_1 \geq \lambda_2 \cdots \geq \lambda_D$), which also represent their contributions to speaker variation. The top K eigenvectors, named as "eigenvoices", are selected (where $K \ll D$). They account for most of the variation in the reference speakers.

Let $\boldsymbol{e}(0)$ be the mean supervector[1], then for a new speaker, the supervector $\boldsymbol{S}$ could be represented as follows:

$$\boldsymbol{S} = \boldsymbol{e}(0) + \sum_{k=1}^{K} (\boldsymbol{x}(k) \times \boldsymbol{e}(k)) \tag{2}$$

Given the adaptation data from a new speaker, only the coefficient vector $\boldsymbol{x}$=(x(1), x(2), ⋯, $x(k)$)$^{\mathrm{T}}$ of size K needs to be estimated.

Suppose that the single Gaussian state-output distribution is used. Given the adaptation data $O = \boldsymbol{o}_1^T \triangleq \boldsymbol{o}_1, \cdots, \boldsymbol{o}_T$ with the corresponding transcription $W = w_1^N \triangleq w_1, \cdots, w_N$, let:

n : the dimension of the feature vector;

$\boldsymbol{o}_t$: the feature vector at time t;

$\boldsymbol{\mu}_s$: the mean vector for state s;

C_s : the covariance for state s;

$\boldsymbol{e}_s(j)$: the subvector of eigenvoice j corresponding to state s;

We also define $\gamma_s(t)$ as the occupation probability for state s at the time t given sentence transcription, i.e.,

$$\gamma_s(t) = P(s_t = s \mid W, O) \tag{3}$$

The maximum likelihood estimation method, called maximum likelihood eigen-decomposition (MLED) in Ref. [4] aim to maximize the likelihood $P(O \mid W, x)$ with respect to the unknown parameters $\boldsymbol{x}$. This is done by iteratively maximizing an auxiliary function $Q(x', x)$ as follows:

$$Q(x', x) = P(W, O \mid x') \times \sum_s \sum_t \gamma_s(t) [\log P(\boldsymbol{o}_t \mid s, \boldsymbol{x})] \tag{4}$$

where x' is the current model, and $\boldsymbol{x}$ is the model to be estimated.

$$\log P(o_t \mid s, \boldsymbol{x}) = -0.5n \log(2\pi) - 0.5 \log |C_s| - 0.5(\boldsymbol{o}_t - \boldsymbol{\mu}_s)^{\mathrm{T}} C_s^{-1} (\boldsymbol{o}_t - \boldsymbol{\mu}_s) \tag{4}$$

According to Eq. (2), we get:

$$\boldsymbol{\mu}_s = \boldsymbol{e}_s(0) + \sum_{k=1}^{K} (\boldsymbol{x}(k) \boldsymbol{e}_s(k)) \tag{5}$$

Denote $\hat{\boldsymbol{x}}$ as the re-estimation of coefficient vector which maximizes $Q(\boldsymbol{x}', \boldsymbol{x})$ over $\boldsymbol{x}$. Let $\partial Q / \partial \boldsymbol{x}(j) = 0$, j = 1, ⋯, K

We obtain the update equation for each $j, j = 1, \cdots, K$:

$$\sum_s \sum_t \gamma_s(t) (\boldsymbol{e}_s(j))^{\mathrm{T}} C_s^{-1} (o_t - \boldsymbol{e}_s(0)) = \sum_s \sum_t \gamma_s(t) \times \sum_{k=1}^{K} \hat{\boldsymbol{x}}(k) (\boldsymbol{e}_s(k))^{\mathrm{T}} C_s^{-1} \boldsymbol{e}_s(j) \tag{6}$$

3 MAP adaptation with eigenvoices

Note that the use of eigenvoices imposes a strong constraint on mean vectors; it is beneficial to explore the prior information in the model estimated. Maximum aposteriori criterion could be used for this purpose. The MAP estimation of acoustic model parameter θ is:

$$\theta = \arg\theta \max [P(O \mid W, \theta) P_0(\theta)] \tag{7}$$

where $P_0(\theta)$ denotes the prior probability of the known parameters θ . The estimation procedure under MAP criterion is called maximum a posteriori eigen-decomposition (MAPED).

Since $x(k)$, $k = 1, \cdots, K$ are the projections of the speaker supervector to the eigenvectors, the prior probability can be derived directly from the eigen-analysis of covariance matrix $\boldsymbol{C}$. Multi-dimensional Gaussian distribution is used to model supervector $\boldsymbol{S}$ with the mean $\boldsymbol{e}(0)$ and the covariance $\boldsymbol{C}$, then the prior probability is given by:

$$P_0(\boldsymbol{x}) = P_0(\boldsymbol{S}) = \boldsymbol{N} \ (\boldsymbol{S} \mid \boldsymbol{e}(0), \ \boldsymbol{C}) \tag{8}$$

Substitute the eigenvoice expression of the speaker supervector from Eq. (2) into Eq. (9), and rewrite $\boldsymbol{C}$ as in Eq. (1), then the log prior probability of coefficient vector $\boldsymbol{x}$ is given by:

$$\log P_0(\boldsymbol{x}) = \text{const} - \frac{1}{2} \left(\sum_{k=1}^{K} \boldsymbol{x}(k) \boldsymbol{e}(k) \right)^{\mathrm{T}} (\boldsymbol{e}(1), \cdots, \boldsymbol{e}(D)) \times \text{diag}\,(\lambda_1^{-1}, \cdots, \lambda_D^{-1}) ((\boldsymbol{e}(1), \cdots, \boldsymbol{e}(D))^{\mathrm{T}} (\sum_{k=1}^{K} \boldsymbol{x}(k) \boldsymbol{e}(k))) \tag{9}$$

Since $\boldsymbol{e}(i)^{\mathrm{T}} \boldsymbol{e}(j) = 0$, $i \neq j$ and $\boldsymbol{e}(i)^{\mathrm{T}} \boldsymbol{e}(\mathrm{i}) = 1$,[2] Equation. (10) can be rewritten as[3]:

$$\log P_0(\boldsymbol{x}) = \text{const} - \frac{1}{2} \sum_{k=1}^{K} \frac{\boldsymbol{x}(k)^2}{\lambda_k} \tag{10}$$

Equation (11) gives that the mean of each coefficient $\boldsymbol{x}(k)$ is equal to zero, and the variance is equal to the corresponding eigenvalue λ_k .

Redefine auxiliary function as:

$$Q'(\boldsymbol{x}', \boldsymbol{x}) = P(W, O \mid \boldsymbol{x}') \left\{ \sum_s \sum_t \lambda_s(t) [\log P(\boldsymbol{o}_t \mid s, \boldsymbol{x})] + \log P_0(x) \right\} \tag{11}$$

Substitute $P_0(x)$ and let $\partial Q' / \partial \boldsymbol{x}(j) = 0, j = 1, \cdots, K$ we get:

$$\sum_s \sum_t \gamma_s(t) (\boldsymbol{e}_s(j))^{\mathrm{T}} C_s^{-1} (\boldsymbol{o}_t - \boldsymbol{e}_s(0)) = \sum_{k=1}^{K} \boldsymbol{x}(k) \left[\sum_s \sum_t \gamma_s(t) (\boldsymbol{e}_s(k))^{\mathrm{T}} C_s^{-1} \boldsymbol{e}_s(j) + \frac{\delta_{k,j}}{\lambda_j} \right] \tag{12}$$

for each $j, j = 1, 2, \cdots, K$. Here:

$$\delta_{k,j} = \begin{cases} 1, & k = j \\ 0, & \text{otherwise} \end{cases} \tag{13}$$

[1] In implementation, a supervector constructed from SI model is used instead of the mean supervector, since the robust estimation of mean vector will require speech data from large volumes of speakers, which is difficult in practice.

[2] It's the property of normalized eigenvectors.

[3] Though it's only derived for the case of full-ranked covariance C, it can also be adapted to other cases since the equation is only associated with the top K eigenvalues.

4 Eigenvoice-based MAP adaptation within correlation subspace

When applied to LVCSR, the speaker vectors are always high dimensional even if only single-Gaussian state output distribution is used. In contrast, the free parameters to be estimated are much fewer, which implies strict limits between components, while some states are not so highly correlated.

One of the methods for solving this problem is to use the speaker-specific transform matrixes to take the place of speaker vectors [7], while in this paper independent subspaces analysis is used. Firstly, all the states are clustered into several classes according to some similarity measures[4], thus the high-dimensional speaker vector space is divided into several low-dimensional subspaces, which we will call as "correlation subspaces". Secondly, eigenvoices are obtained via the PCA method in each correlation subspace respectively, which means the correlations between different classes are ignored. Assume total speech components are clustered into H correlation subspaces, and for each correlation subspace a super vector containing all the mean vectors of all the components in this subspace is constructed. Use h as the indication of a correlation subspace, then the supervector $s^{(h)}$ in this subspace is expressed as Eq. (15):

$$\boldsymbol{S}^{(h)} = \boldsymbol{e}^{(h)}(0) + \sum_{k=1}^{K^{(h)}} (\boldsymbol{x}^{(h)}(k)\ \boldsymbol{e}^{(h)}(k)) \tag{14}$$

Here $\boldsymbol{e}^{(h)}(i), i = 0, \cdots, K^{(h)}$ denote eigenvoices of size $K^{(h)}$ in the hth subspace, chosen from $\boldsymbol{D}^{(h)}$ eigenvectors. The update equation Eq. (13) is modified as Eq. (16) for each correlation subspace, $h = 1, \cdots, H$:

$$\begin{aligned} &\sum_{s\in\Phi(h)} \sum_{t} \gamma_s(t)(\boldsymbol{e}_s^{(h)}(j))^{\mathrm{T}} C_s^{-1} (\boldsymbol{o}_t - \boldsymbol{e}_s^{(h)}(0)) \\ &= \sum_{k=1}^{K^{(h)}} \boldsymbol{x}^{(h)}(k) \left[\sum_{s\in\Phi(h)} \sum_{t} \gamma_s(t)(\boldsymbol{e}_s^{(h)}(k))^{\mathrm{T}} C_s^{-1} \boldsymbol{e}_s^{(h)}(j) + \frac{\delta_{k,j}}{\lambda_j^{(h)}} \right] \end{aligned} \tag{15}$$

Here $j = 1, 2, \cdots, K^{(h)}$, $\lambda_j^{(h)}$ are eigenvalues in descent order, $\boldsymbol{x}^{(h)}(k), k = 1, \cdots, K^{(h)}$ are coefficients, and $\Phi(h)$ is the congregation of components[5] in this subspace.

5 Experimental results

5.1 Database and experimental approach

To evaluate the performance of MAPED within correlation subspaces[6], experiments were carried out on a Chinese LVCSR task using speech database for "China 863 Assessment". The training data was the utterances from 83 speakers, each with 650 sentences. Thus there were 83 SD models constructed. Other 5 persons' data were used for evaluation, each of them contributed 120 sentences, which comprised 2 940 syllables and lasted from 10 to 15 minutes.

All the Chinese characters are pronounced as one of the total 408 un-toned Chinese syllables in CV structure. A left-to-right HMM syllable model was used, with 2 states for consonant and 4 states for the vowel part. The total number of states is 857, and the acoustic features were 45-dimensional formed by 14 MFCCs along with normalized log-energy and their first and second order differentials. Single Gaussian model was used for state-output probability density with full covariance matrix.

Here we focused on the acoustic part. The speech was decoded into free syllable strings without any grammar constraints, and the result was organized into syllable-lattices. No language model was used. Only the syllable error rate (SER) was reported for performance comparisons.

Adaptation was carried out in supervised mode. The first 60 sentences were used to estimate the weight coefficients, and other 60 sentences reserved for evaluation. Adaptation sentences were gradually increased from 1 to 60, and the SERs with different adaptation data were reported compared with MAP method and MLLR method.

Performance with various eigenvoices number was compared. Since each eigenvalue represented the variation of the projection in the corresponding eigenvector direction, they could be regarded as important indications for each eigenvoice. Eigenvoices with small eigenvalue were viewed as the result of noise or unexpected variations, so they could be ignored without affecting recognition performance. Denote $E_{\mathrm{rest}}^{(h)}$ as the summation of $\lambda_i^{(h)}$, $i = K^{(h)}+1, \cdots, D^{(h)}$, which represents the energy of rest eigenvectors in the subspace, and $E_{\mathrm{all}}^{(h)} = \sum_{i=1}^{D^{(h)}} \lambda_{(i)}^{(h)}$, which represents the total energy. Threshold t_E is used to restrict the ratio of $E_{\mathrm{rest}}^{(h)}$ to $E_{\mathrm{all}}^{(h)}$ is no more than t_E, $h = 1, 2, \cdots, H$.

Different choices of subspaces were also compared. In experiment different clustering thresholds were used to construct different subspaces, and the results with different subspaces would be reported.

5.2 Experimental results

The first experiment was carried out with H fixed to be 52 and t_E = 0.01. MAP and MLLR methods were also performed to give a comparison. Results are listed in Table 1, which shows MAPED always outperformed both MAP and MLLR according to gradually added data. The SER decreases from 30.96 % to 21.81% with MAPED method, which gains 29.55 % relative error rate reduction. It's also observed that the advance of MAP is not so notable, and when adaptation data is too limited MLLR would even cause unexpected performance decrease.

[4] Euclidean distance is used in this paper.

[5] A "component" is same as a state in single Gaussian model.

[6] To simplify the description, MAPED will be used to indicate the MAPED within correlation subspaces in the following text.

Table 1 SER Comparison

nSent	MAP/%	MLLR/%	MAPED/%
0	30.96	30.96	30.96
1	30.90	103.22	28.96
2	30.75	73.26	28.64
3	30.72	49.03	28.12
4	30.57	40.84	27.99
5	30.55	33.64	27.46
10	30.01	28.34	26.83
20	28.89	26.33	24.97
30	27.42	24.61	23.95
40	26.07	25.49	23.28
50	24.92	24.00	22.84
60	24.01	23.83	21.81

Comparison of SER for MAP, MLLR and MAPED. The adaptation data was gradually added from 1 to 60 sentences. $K^{(h)} = 60,\ h = 1, \cdots, H$

Table 2 gives the SERs in recognized syllable graph, which is with 5 candidates at each node and thus is profit post processing. MAPED also outperforms MAP and MLLR method consistently in this experiment.

Table 2 SER comparison in syllable graph

nSent	MAP/%	MLLR/%	MAPED/%
0	9.17	9.17	9.17
1	9.06	101.01	8.81
2	9.05	46.78	8.47
3	8.99	20.44	8.47
4	8.98	14.61	8.31
5	8.94	10.74	8.34
10	8.74	8.38	7.60
20	8.51	7.55	6.86
30	8.03	7.16	6.47
40	7.69	7.50	6.38
50	7.51	6.99	6.39
60	7.20	7.00	6.11

Comparison of SER in syllable graph for MAP, MLLR and MAPED. The adaptation data was gradually added from 1 to 60 sentences. $K^{(h)} = 60, h = 1, \cdots, H$

Figure 1 shows the affection of t_E. In this experiment, H is fixed to be 52. MAPED-1 gives the result with $t_E = 0.1$ and MAPED-2 with t_E = 0.01. There are 3 720 eigenvoices in MAPED-2, much more than 1 976 eigenvoices in MAPED-1, whereas the performance is very similar, which indicates that PCA is well adapted to reduce redundancy.

Figure 2 gives the result with different subspace number, and $t_E = 0.01$. $H = 39$ in MAPED-1, $H = 52$ in MAPED-2, and $H = 89$ in MAPED-3. They all gave similar SERs,

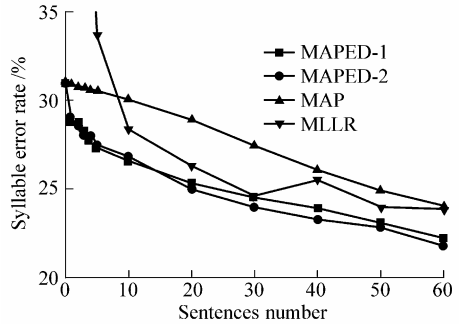

Fig. 1 Syllable error rate (SER) with different eigenvoice number, the results of MAP and MLLR are also drawn to give a comparison, H = 52 . $t_E = 0.1$ in MAPED-1 and $t_E = 0.01$ in MAPED-2.

which show that the results are not strongly dependent on the cluster procedure of components, and prove that weak correlations have little effect on the performance. Thus the subspace method would help to perform adaptation when the full space is too large to handle. We also expect it would help to depict the model more accurately.

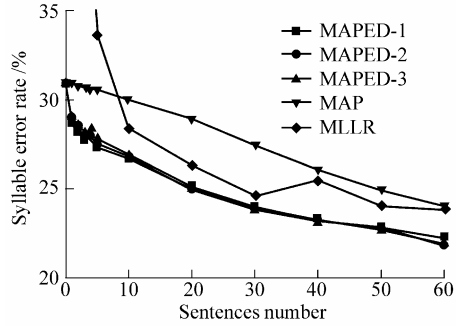

Fig. 2 SER with different clustering thresholds, $t_E = 0.01$. H = 39 in MAPED-1, H = 52 in MAPED-2 and H = 89 in MAPED-3

6 Conclusions

Eigenspace-based MAP adaptation within correlation subspaces is an efficient and robust adaptation method. Remarkable improvements are observed even with very limited adaptation data, which outperforms both MAP and MLLR. Another notable benefit of this method is that only a few eigenvectors are needed to update all the parameters, which greatly reduces the complexity of computation. In addition, the subspace method helps to depict the model more accurately, and makes the implementation of the eigenvoice method on high-dimensional full space practical. Applying this method to Gauss mixture model will be a promising technique for real applications.

Acknowledgements This work was supported by NSFC(No. 60402029).

References

1. Lee C., Lin C., Juang B., A study on speaker adaptation of parameters of continuous density hidden markov models, IEEE Trans. Signal Process., 1991, 39(4): 806–814
2. C. R, Deng L., A maximum a posteriori approach to speaker adaptation using the trended hidden markov model, IEEE Trans. Speech Audio Process., 2001, 9(5): 549–557
3. Lee C., Lin C., Juang B., Speaker adaptation of continuous density hmms using linear regression, Proceedings of ICSLP, 1994, 451–454
4. K. R, J. J-C, N. P., et al., Rapid speaker adaptation in eigenvoice space, IEEE Trans. Speech Audio Process., 2000, 8(6): 695–707
5. B. H, Very fast adaptation for large vocabulary continuous speech recognition using eigenvoices, Proceedings of ICSLP, 2000, 354–357
6. Jolliffe I. T., Principal Component Analysis. Springer, 1986
7. Kuan-ting C., Wen-wei L., Hsin-min W., Fast speaker adaptation using eigenspace-based maximum likelihood linear regression, Proceedings of ICSLP, 2000, 742–745