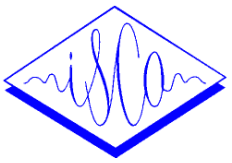

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

SPEAKER NORMALIZATION TRAINING AND ADAPTATION FOR SPEECH RECOGNITION

Lei HE, Ditang FANG, Wenhu WU
Center of Speech Technology, State Key Laboratory of Intelligent Technology and Systems
Department of Computer Science & Technology, Tsinghua University, Beijing, 100084, China
helei@sp.cs.tsinghua.edu.cn

ABSTRACT

This paper presents a speaker adaptation framework that combines the speaker normalization (SN) training. Because of the varieties among training speakers, more data are required in training and adaptation of speaker independent (SI) acoustic model. In this paper, a very simple but effective normalization method is presented, in which the distortions among different speakers are removed by subtracting the state-relative shift vectors between SI model and speaker dependent (SD) model. In adaptation stage, MAP estimation is used to update the models with adaptation data, and the interpolation of unseen models and smoothing of the final models are implemented by order-alterable weighted neighbor regression (WNR) method. In Mandarin syllable recognition task, with equal adaptation data, SN model as seed model makes a 5%-15% additional reduction in error rate comparing with SI model as seed model.

1. INTRODUCTION

The most common approach to train a speaker independent recognizer is to estimate the parameters of acoustic model with speech from a large population of speakers. Thus, the target model is a smoothing one of training corpus, in which there are both phonetic variation and inter-speaker variance. As a result, the performance of such SI model based system may degrade significantly when the testing speaker mismatches the training population. To alleviate such degradation, one approach is the speaker adaptation technique, which can effectively reduce the distortion between testing speaker and training set. However, because of the varieties among training speakers, the feature distribution often exhibits higher variances. Consequently, more parameters are required to describe the SI model, with great requirement of training and adaptation data. Furthermore, the higher overlap in feature space among different recognition units caused by the inter-speaker variance may be much more severe. Accordingly, reducing the inter-speaker variability with speaker normalization algorithm before model training becomes an attractive choice.

In past ten years, many efforts to reduce the inter-speaker variability have been done. The most common used method is cepstral mean normalization (or subtraction)[2]. Here, the average of cepstral is subtracted from each feature vector to remove the slow-variable factor in speech signal, such as change of transfer channels, speakers, and so on. Another popular method is vocal tract length normalization (VTLN) [3,4,6,8,10,11]. Here, the vocal tract length is regarded as the exclusive speaker specific characteristic. Then the inter-speaker variance can be removed with warping of frequency axis before computing cepstral coefficients. Usually, the warping factor is a global one for each speaker, which can be estimated with the relative vocal tract length comparing with a reference speaker, or the average of training population. Thus, the frequency warp function is equal to each recognition units. In [1], the speaker specific characters are modeled by linear regression transformation that maps the speaker independent mean vector to speaker dependent mean vector. The parameters of linear regressions can be estimated by MLLR [7] with the model classes clustered in term of a hierarchical structure, thus the transformations are model-class relative. In [9], the signal bias is used to reduce the cross unit overlaps with maximum likelihood estimation.

In this paper, we present a simple normalization method, in which the speaker specific attribute is described with the state-relative bias vector. The shift vector between SI model mean vector and SD model mean vector is used as the bias vector. In adaptation stage, the MAP estimation is used to update the models with adaptation data, and the model interpolation and smoothing are performed by the order-alterable weighted neighbor regression method [5].

In next section, the details of speaker normalization are presented. The order-alterable WNR algorithm is introduced in section 3. Section 4 is the experimental result with corresponding analysis. Conclusions are described in last section.

2. SPEAKER NORMALIZATION WITH DIRECT MEAN SHIFT

For each adaptation technique, the selection of an appropriate seed model is important. In speaker adaptation environment, the smoothed model trained with data from a large population of

speakers is the general choice. However, because those training data are collected from different speakers, the result model has rather broad distributions, with severe cross-unit overlap, which may not be sufficient specific to adaptation. Figure 1 is the sketch map of cross-unit overlap caused by inter-speaker variance. Here the speaker independent model has a broader distribution than speaker dependent model (1 or 2). Meanwhile, unit E and unit A are distinguishable either in speaker dependent model 1 or 2, but overlap each other in speaker independent model space. Obviously, more cross-unit overlaps here, more data required in adaptation.

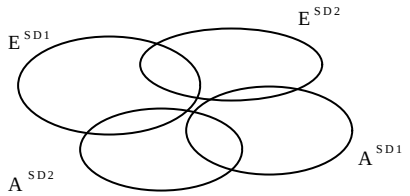

Figure 1. Cross-Unit overlaps caused by inter-speaker variance

To alleviate such underlying disadvantage, a normalization process is combined into the adaptation framework, to form a better seed model. In this process, the speaker specific attributes are modeled by the mean shift vectors from SD model to SI model, which are demonstrated in figure 2.

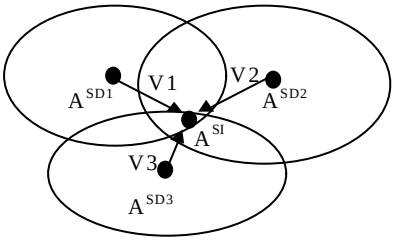

Figure 2. Shift vector from SD model space to SI model space

Here V1, V2 and V3 are shift vectors of speaker 1, 2 and 3 respectively for unit A.

In experiments, because of the speaker specific data are limited, which cannot give a well-trained SD model, the speaker adapted (SA) model trained with MAP estimation takes the place of SD model. Then, the shift vector can be calculated with equation (1).

$$v_{ik} = \frac{\tau_{ik} m_{ik}^{SI} + \sum_{t=1}^{T} c_{ikt} x_t}{\tau_{ik} + \sum_{t=1}^{T} c_{ikt}} - m_{ik}^{SI} = \frac{\sum_{t=1}^{T} c_{ikt} (x_t - m_{ik}^{SI})}{\tau_{ik} + \sum_{t=1}^{T} c_{ikt}} \quad (1)$$

Here,

$$c_{ikt} = \frac{w_k N(x_t \mid \lambda_{ik}^{SI})}{\sum_k w_k N(x_t \mid \lambda_{ik}^{SI})} \quad (2)$$

$\lambda_{ik}^{SI} = (m_{ik}^{SI}, \Sigma_{ik}^{SI})$ is the *k*-th Gaussian component of state *i* of SI model, with corresponding mixture weight w_k. $X = (x_1, \Lambda, x_T)$ are observation samples assigned to state *i*. v_{ik} is corresponding shift vector.

It is apparently that the equation (1) directly converges to maximum likelihood estimation formula along with the increase of $\sum_{t=1}^{T} c_{ikt}$ like other MAP estimation equation:

$$v_{ik} = \frac{\sum_{t=1}^{T} c_{ikt} (x_t - m_{ik}^{SI})}{\sum_{t=1}^{T} c_{ikt}} \quad (3)$$

Figure 3 is the block diagram of above normalization.

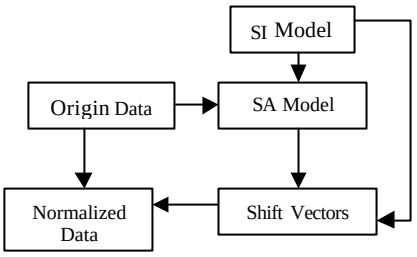

Figure 3. System block diagram of speaker normalization with direct mean shift

With subtracting corresponding shift vector, speaker specific features are normalized, and then used to train the seed model for future adaptation. The advantage of this method is direct aiming at the reduction of cross-unit overlaps caused by inter-speaker differences. Because the normalized feature space is much more compact than original one, it may be more appropriate for adaptation.

3. SPEAKER ADAPTATION WITH MAP ESTIMATION AND ORDER-ALTERABLE WNR

In our previous work, a speaker adaptation framework based on the combination of MAP estimation and WNR is presented[5]. In such framework, after MAP estimation, the model interpolation of unseen model and smoothing of final model are performed with a transformation-based method, WNR. This method gives a general framework, into which various regression models can be brought. In this paper, an order-alterable regression is adopted, which can be described as following.

In WNR, a threshold of regression reliability is used to prevent the weak regression. For example, the correlative coefficient is adopted with linear regression model applied. On the other hand, this threshold can be used to dynamically choose the proper

order of regression model. In this paper, if the correlative coefficient of linear regression is less than the threshold, vector field smoothing (VFS) will be used instead. Here VFS can be regard as zero-order regression with only the transfer vectors as regression parameters.

The linear regression and zero-order regression can be expressed with equation (4) and (5) respectively.

$$\tilde{m}_{ik} = Bm_{ik} + b_0 \quad (4)$$

$$\tilde{m}_{ik} = m_{ik} + \sum_{m \in A} w_m (\tilde{m}_m - m_m) \quad (5)$$

Here, $\boldsymbol{l}_{ik} = (m_{ik}, \Sigma_{ik})$ and $\tilde{\boldsymbol{l}}_{ik} = (\tilde{m}_{ik}, \tilde{\Sigma}_{ik})$ denote the Gaussian distribution in the SI and SA model space respectively. *A* is the adapted *M*-nearest neighbors set of $\boldsymbol{l}_{ik}$. w_m is the weight of *m*-th component $\boldsymbol{l}_m = (m_m, \Sigma_m)$ in *A*, which is also used in the estimation of linear regression parameters: transform matrix B and offset vector b_0. The details of WNR can be obtained in [5].

Figure 4 is the block diagram of speaker adaptation with MAP estimation and order-alterable WNR. Here, WNLR is the abbreviation of weighted neighbor linear regression, in which linear regression is applied.

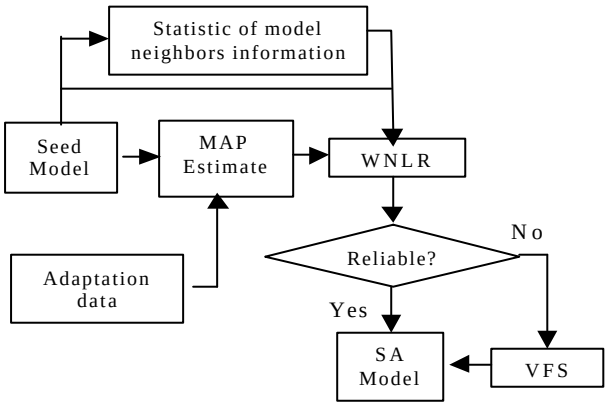

Figure 4. Speaker adaptation framework based on MAP estimation and order-alterable WNR

4. EXPERIMENTS

This section is the experimental evaluation of speaker normalization training and adaptation framework described before. After database and speech recognizer introduction, details of experimental results and corresponding analysis are presented

In all experiments, a continuous read Chinese speech corpus selected from 863 materials is used, where approximately 520 sentences are available for each of 15 male speakers. This corpus is syncopated and labeled to Chinese syllable level in advance, then parameterized using 17 dimensional MFCC and the corresponding 17 dimensional auto regression coefficients. The arrangement of database is shown in Table 1. Moreover, for each testing speaker, 100 sentences are used as testing data. The adaptation data are selected from the rest sentences.

A simplified CDHMM with six left-to-right states is used to build a Mandarin syllable recognition system. The simplification means that only the observation probability density function (PDF) described with mixture Gaussian distributions is preserved. Three different numbers of mixture are selected: 2, 4, and 8.

TABLE 1. Database arrangement

Information	Training Set	Adaptation and Testing Set
Syllables covered	398	398
Speakers	12	3
Utterances	6,244	1,561
Total Samples	78,027	19,553

With normalization, the SN model is more compact than SI model, which can be demonstrated with Table 2.

TABLE 2. Comparison of Recognition Accuracy of Training-set before and after normalization (%)

Number of Mixture	SI Model	SN Model
2	79.8	87.8
4	89.6	94.7
8	95.2	98.9

Apparently, the feature distribution can be reduced with normalization to a large extent, so the training-set can be modeled well with less parameters. The accuracy of training-set may not be very significative for real testing, but it is useful to demonstrate the status of model distribution.

Table 3 gives the results of comparison experiments of speaker adaptation with SI model as seed model and SN model as seed model. Here, all results are average over 3 testing speakers. The last column is the error rate reduction from adapted SI mode to adapted SN model.

TABEL 3. Comparison of adaptation with SI model and SN model as seed model (%)

Adaptation Sentences	SI model Accuracy	SN model Accuracy	Error Rate Reduction
Number of Mixture = 4			
0	67.50	67.76	-
20	70.47	71.94	5.0
50	73.23	74.67	5.4
100	77.93	79.80	8.5
250	82.57	85.16	14.9
Number of Mixture = 8			
0	70.87	70.37	-

20	73.87	75.18	5.0
50	75.80	77.13	5.5
100	80.47	81.97	7.7
250	84.77	86.13	8.9

According to Table 3, we can draw following conclusion.

- The SN model can be more efficient adapted than SI model either with sparse adaptation data or abundant data. In Table 3, the error rate of adapted SI model can be reduced from 5% to 15% with SN model as seed model.
- The performance of SN model is comparable with SI model even without adaptation, which demonstrates that the phonetic information can be preserved well after normalization.

5. CONCLUSION

In this paper, a speaker normalization process is combined into adaptation framework. In this process, the inter-speaker variances of training set are removed by direct subtracting the corresponding shift vector between SI model mean and SD model mean. Moreover, in adaptation stage, an extended WNR algorithm is presented for model interpolation and smoothing, in which the order of regression model can be dynamically chose according to the reliability estimation result.

In some speaker normalization methods, the testing data can be normalized with the same technique like training data. However, in those methods, a physical model assumption is usually indispensable, and the speaker specific attribute can only model with a few parameters. For example, in VTLN, the multi-tube lossless model of the vocal tract is necessary, and only a global vocal tract length is used to describe the differences among speakers. Apparently, there are too many simplifications in those methods. On the other hand, some mathematics assumptions are used to model and reduce the inter-speaker differences in some other methods, in which the SN model need be adapted before testing. Here, the adaptation method can be similar to the normalization method. For example, normalization and adaptation are both implemented with MLLR. Consequently, such normalization process can be called speaker adaptive training (SAT) too. In this paper, the bias vector is used to model the speaker specific attribute. Because the bias vector for each state is independent of other states, it is state-relative, which can be detailed enough to describe the inter-speaker variances. Moreover, subtraction those bias vectors can directly remove the overlaps among different units. Accordingly, this normalization method is efficient despite its simplicity, which has been proved by experimental results.

6. References

[1] T. Anastasakos, J. McDonough, and J. Makhoul, "Speaker Adaptive Training: A Maximum Likelihood Approach to Speaker Normalization", *ICASSP*,1997,vol.2,pp1043-1046.

[2] T. Anastasakos, F.Kubala, J.Makhoul, and R.Schwartz, "Adaptatoin to New Microphone using tied-mixture Normalization", *ICASSP*, 1994,pp433-436.

[3] T. Claes, I. Dologlou, L. Bosch, and F. Compernolle, "A Novel Feature Transformation for Vocal Tract Length Normalization in Automatic Speech Recognition", *IEEE Trans. Speech and Audio Processing*, vol.6, No.6, pp549-557, Nov. 1998.

[4] E. Eide, H. Gish, "A Parametric Approach to Vocal Tract Length Normalization", *ICASSP*,1996,vol.1,pp346-348.

[5] L. He, J. Wu, D. Fang, and W. Wu, "Speaker Adaptation Based on Combination of MAP Estimation and Weighted Neighbor Regression", *ICASSP*,2000,vol.2.pp981-984.

[6] L. Lee, and R. Rose, "A Frequency Warping Approach to Speaker Normalization", *IEEE Trans. Speech and Audio Processing*, vol.6, No.1, pp49-60, Jan. 1998.

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

[8] T. Matsui, T. Matsuoka, and S. Furui, "Frequency-Warping and Speaker-Normalization", *ICASSP*,1997,vol.2,pp1015-1018.

[9] M. Rahim, and B-H. Juang, "Signal Bias Removing by Maximum Likelihood Estimation for Robust Telephone Speech Recognition", *IEEE Trans. Speech and Audio Processing*, vol.4, No.1, pp19-30, Jan. 1996.

[10] S. Umesh, L. Cohen, and D. Nelson, "Frequency-Warping and Speaker-Normalization", *ICASSP*,1997,vol.2,pp983-986.

[11] P. Zhan, and M. Westphal, "Speaker Normalization Based on Frequency Warping", *ICASSP*,1997,vol.2,pp1039-1042.