Restoring the Residual Speaker Information in Total Variability Modeling for Speaker Verification

Ce Zhang[1], *Rong Zheng*[1], *Bo Xu*[1,2]

[1]Digital Content Technology Research Center, Institute of Automation
[2]National Lab of Pattern Recognition, Institute of Automation
Chinese Academy of Sciences, Beijing 100190, China

{czhang,rzheng,xubo}@hitic.ia.ac.cn

Abstract

In this paper, we introduce the residual space into the Total Variability Modeling by assuming that the speaker super-vectors are not totally contained in a linear subspace of low dimension. Thus the feature reduction carried out by Probabilistic Principal Component Analysis(PPCA) leads to information loss including information of speaker as well as channel. We add the residual factor to restore the missing speaker information which is lost during the PPCA process. To utilize the recovered information effectively, we propose two fusion methods that combine the principal components with the residual factor. We compare the fusion results that are obtained with direct scoring and Support Vector Machines for classification, respectively. The experiments on NIST SRE 2006 show that the performance can be improved consistently by involving the residual factor, e.g. the best result achieves 6% relative improvement on Equal Error Rate(EER) compared to the baseline system.

Index Terms: speaker verification, total variability, i-vector, residual space, score combination

1. Introduction

In recent years, many approaches have been tested to improve performance of the Gaussian Mixture Models based on Universal Background Model (GMM-UBM) [1]. The most popular one is Joint Factor Analysis (JFA)[2] which deals with speaker variabilities and channel/session inconsistencies simultaneously.

Recently, [3] proposed a factor analysis-based approach to speaker verification. Unlike JFA which models inter-speaker and within-speaker variability separately in a high dimensional space of super-vectors, total variability modeling consists of finding a low dimensional subspace of the GMM super-vector space, named the total variability space that represents both speaker and channel variability. This new model can be seen as a PPCA that allows us to project the super-vectors onto the total variability space.

However, there must be some information loss during the projection from the super-vector space to the total variability space. The lost information can be divided into two parts: speaker-dependent and channel-dependent. In this paper, we focus on the effective retrieval of speaker information which is lost in the feature reduction. Based on the total variability modeling, we first introduce a residual factor to capture the speaker variability which does not belong to the total variability space. Then we consider combining the total factor with the residual factor in both direct scoring methods and Support Vector Machines.

The paper is organized as follows. Section 2 gives a brief introduction to the total variability modeling (TVM). In section 3, we first introduce the residual space for the TVM and then explain how to combine the principal and residual information effectively. We present our experimental configuration and results in section 4 and the last section draws some conclusions.

Supported by the National Natural Science Foundation of China (Grant No. 90820303)

2. Total variability background

Total variability modeling outlined by Dehak [3] is based on conventional GMM-UBM approach, which simultaneously models the speaker and channel variabilities in only one space. Let F and C to be the acoustic feature dimension and the total number of Gaussian mixture components. The GMM super-vector is a CF-dimensional vector which is formed by concatenating the means of each mixture component. Given an utterance, the speaker- and channel- dependent GMM super-vector $\boldsymbol{M}$ is written as follows:

$$\boldsymbol{M} = \boldsymbol{m} + \boldsymbol{T}\boldsymbol{w} \tag{1}$$

where $\boldsymbol{m}$ is speaker- and channel- independent super-vector representing the center of the full parameter space, usually UBM super-vector is a good estimate of $\boldsymbol{m}$, $\boldsymbol{T}$ is a rectangular matrix of low rank and $\boldsymbol{w}$ is a standard normally distributed random vector. $\boldsymbol{T}$ and $\boldsymbol{w}$ are commonly referred to as total variability space and total factor, respectively.

3. Residual factor in total variability modeling

In the total variability modeling, a super-vector is represented by a low dimensional total factor $\boldsymbol{w}$ through PPCA projection. However, there must be some speaker information loss through

the projection especially when the rank of $\boldsymbol{T}$ is not enough to capture all the speaker variability. In this section, we consider adding a residual factor to the total variability model which means to capture the residual speaker information excluded from the total factor.

3.1. Extending total variability model with residual factor

First of all, we introduce the residual factor to compensate for the fact that it may be difficult to find enough training speakers to estimate $\boldsymbol{T}$ reliably. In this case, the speaker- and channel-dependent GMM super-vector $\boldsymbol{M}$ is rewritten as:

$$\boldsymbol{M} = \boldsymbol{m} + \boldsymbol{T}\boldsymbol{w} + \boldsymbol{D}\boldsymbol{z} \tag{2}$$

where $\boldsymbol{D}$ is a $CF \times CF$ diagonal matrix and $\boldsymbol{z}$ is a normally distributed CF-dimensional random vector. To be consistency with JFA, we refer to $\boldsymbol{D}$ as residual space and $\boldsymbol{z}$ as the residual factor.

We split the multi-session training set in two subsets and use the larger of the two to estimate $\boldsymbol{T}$ and the smaller to estimate $\boldsymbol{D}$. We also decouple the estimation of $\boldsymbol{T}$ and $\boldsymbol{D}$, as well as $\boldsymbol{w}$ and $\boldsymbol{z}$. The detailed estimation of the two spaces is presented in Figure 1. Firstly, we estimate $\boldsymbol{T}$ using the EM algorithm described in [2], by assuming each training utterance is spoken by a different speaker. Then we calculate the point estimate of $\boldsymbol{w}$ for each utterance of the residual training set. Thirdly we centralize the speaker's Baum-Welch statistics by subtracting $\boldsymbol{m} + \boldsymbol{T}\boldsymbol{w}$ and pool all of the recordings of each speaker in the centralized training set regardless of channel effects. Finally we use these centralized statistics to estimate a diagonal model $\boldsymbol{D}$. The assumption here is that channel effects can be averaged out if enough recordings are available for each speaker.

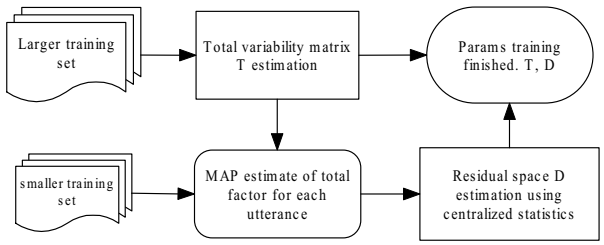

Figure 1: The flowchart showing the estimation of T and D.

3.2. Fusion based on direct scoring

In this section, we consider how to fuse the information provided by $\boldsymbol{w}$ and $\boldsymbol{z}$ using direct scoring. We first represent each speaker by the parameter vector $\boldsymbol{s}$, where $\boldsymbol{s} = \left(\boldsymbol{w}^t, \boldsymbol{z}^t\right)^t$. In [3], the authors proposed the cosine distance scoring in the total variability space which has been proved highly effective,

$$\mathrm{S}_w(\boldsymbol{w}_1, \boldsymbol{w}_2) = \frac{\boldsymbol{w}_1^t \boldsymbol{w}_2}{\|\boldsymbol{w}_1\|\|\boldsymbol{w}_2\|} \tag{3}$$

In order to utilize the residual factor, we consider two approaches. The first is to calculate the cosine distance between two residual factors just as the total factor. That is,

$$\mathrm{S}_z(\boldsymbol{z}_1, \boldsymbol{z}_2) = \frac{\boldsymbol{z}_1^t \boldsymbol{z}_2}{\|\boldsymbol{z}_1\|\|\boldsymbol{z}_2\|} \tag{4}$$

The second is to project the residual factor back to the super-vector space and use the Kullback-Leibler(KL) distance to measure the similarity of two residual super-vectors [4],

$$\mathrm{S}_{KL}(\boldsymbol{z}_1, \boldsymbol{z}_2) = (\boldsymbol{D}\boldsymbol{z}_1)^t \boldsymbol{\Sigma}^{-1} \boldsymbol{\Omega} (\boldsymbol{D}\boldsymbol{z}_2) \tag{5}$$

where $\boldsymbol{\Omega}$ is a $CF \times CF$ diagonal matrix whose diagonal blocks are $\omega_i \boldsymbol{I}$, ω_i is weight of the i-th component of UBM model and $\boldsymbol{I}$ is an identity matrix, $\boldsymbol{\Sigma}$ is a $CF \times CF$ super-covariance matrix whose diagonal blocks are covariance matrices $\boldsymbol{\Sigma}_i$ of mixture component i. Finally, we simply use the linear combination of (3) and (4), or (3) and (5), leading to the following two equations,

$$\mathrm{S}(\boldsymbol{s}_1, \boldsymbol{s}_2) = \alpha \mathrm{S}_w(\boldsymbol{w}_1, \boldsymbol{w}_2) + (1-\alpha)\mathrm{S}_z(\boldsymbol{z}_1, \boldsymbol{z}_2) \tag{6}$$

$$\mathrm{S}(\boldsymbol{s}_1, \boldsymbol{s}_2) = \alpha \mathrm{S}_w(\boldsymbol{w}_1, \boldsymbol{w}_2) + (1-\alpha)\mathrm{S}_{KL}(\boldsymbol{z}_1, \boldsymbol{z}_2) \tag{7}$$

where $\alpha \subset [0, 1]$ is the weight parameter.

3.3. Fusion based on SVM

Support Vector Machine(SVM) is a binary classifier which means to find a separator that maximizes the margin. Kernel functions used in speaker verification system, such as Generalized Linear Discriminate Sequence (GLDS) kernel [5], KL kernel [6] applied in super-vector space, have been proved as effective as generative models. Recently, [3] presented that the cosine kernel used in the total variability space achieves good performance. The cosine kernel between two vectors is given by:

$$\mathrm{K}_w(\boldsymbol{w}_1, \boldsymbol{w}_2) = \frac{\boldsymbol{w}_1^t \boldsymbol{w}_2}{\|\boldsymbol{w}_1\|\|\boldsymbol{w}_2\|} \tag{8}$$

As mentioned in previous section, we also propose two kernels to incorporate the residual factor information. The first one is direct application of cosine kernel in residual factor space,

$$\mathrm{K}_z(\boldsymbol{z}_1, \boldsymbol{z}_2) = \frac{\boldsymbol{z}_1^t \boldsymbol{z}_2}{\|\boldsymbol{z}_1\|\|\boldsymbol{z}_2\|} \tag{9}$$

Second, as defined in (5), we use the KL distance to define the kernel between two residual super-vectors,

$$\mathrm{K}_{\mathrm{KL}}(\boldsymbol{z}_1, \boldsymbol{z}_2) = (\boldsymbol{D}\boldsymbol{z}_1)^t \boldsymbol{\Sigma}^{-1} \boldsymbol{\Omega} (\boldsymbol{D}\boldsymbol{z}_2) \tag{10}$$

Because of the closure property of kernel function which shows that the linear combination of two kernels is also a kernel, we propose here a linear weighted combination of kernel functions (8)(9)(10),

$$\mathrm{K}(\boldsymbol{s}_1, \boldsymbol{s}_2) = \beta \mathrm{K}_w(\boldsymbol{w}_1, \boldsymbol{w}_2) + (1-\beta)\mathrm{K}_z(\boldsymbol{z}_1, \boldsymbol{z}_2) \tag{11}$$

$$\mathrm{K}(\boldsymbol{s}_1, \boldsymbol{s}_2) = \beta \mathrm{K}_w(\boldsymbol{w}_1, \boldsymbol{w}_2) + (1-\beta)\mathrm{K}_{\mathrm{KL}}(\boldsymbol{z}_1, \boldsymbol{z}_2) \tag{12}$$

where $\beta \subset [0, 1]$ is the weight parameter.

3.4. Channel compensation

In our new modeling based on total variability, total factors are extracted regardless of speaker variability and channel variability. However, the residual space models the residual speaker variability which is not captured in the total variability space. We assume that residual factors are channel-independent. And therefore, we only need to perform channel compensation in the total factor space. The channel compensation approach we adopt in this work is Linear Discriminant Analysis (LDA) followed by Within-Class Covariance Normalization (WCCN) [3] [7], which was successfully applied in the speaker factor space in [8].

4. Experiments and results

4.1. Experimental setup

4.1.1. Data set

The results of our experiments are reported on the telephone condition of NIST 2006 SRE corpus [9], English trials only. Both train and test conversations have an average duration of five minutes and there are no cross-gender trials.

4.1.2. Feature extraction

We extract the first 12 Mel frequency cepstrum coefficients together with a log energy feature using a 25 ms Hamming window and a 10 ms frame advance. Mean and variance normalization is used to remove the linear channel effects. Delta and double delta coefficients are then calculated using a 5 frames window and then we obtain a set of 39-dimensional feature vectors. We train bi-gauss model to prune out silence and low energy frames for Speech Activity Detection module. Finally, these 39-dimensional feature vectors are subjected to feature warping using a 3-second sliding window.

4.1.3. Total variability modeling

First, we train two gender-dependent UBMs with 1024 Gaussian components on NIST SRE 2004 telephone data. In total, there are 370 recordings for female and 246 recordings for male. The UBMs are used to collect zero and first order Baum-welch statistics.

For each gender-dependent system, the total variability space is trained on Switchboard II Phase 2, Switchboard Cellular Parts 2, SRE04 and SRE05 telephone data including 9766 recordings from 738 female speakers and 7112 recordings from 514 male speakers. We make use of those speakers, for which eight or more recordings are available in the same set, to train LDA projection matrix of top rank 300 as well as WCCN matrix.

To estimate the gender-dependent residual space, we use the reserved data set including 1022 utterances from 137 female speakers and 899 utterances from 98 male speakers. All the training procedures include seven iterations in the EM algorithm.

4.1.4. Normalization and SVM negative samples

zt-norm is applied in the verification decision phase. We use 200 female and 200 male t-norm models, 500 female and 446 male z-norm utterances, which are derived from NIST SRE 2002, 2004 and 2005 data.

For SVM negative background selection, we make use of a disjoint dataset from the same corpus as zt-norm . We implement the SVM training procedure with LIBSVM [10].

4.2. Results

Table 1 gives the traces of the matrices $\boldsymbol{TT}^t$, $\boldsymbol{DD}^t$ for the gender-dependent models according to different dimensions of the total factor. It is clear that the trace of $\boldsymbol{DD}^t$ decreases as the number of dimensions increases for both gender. With the growth of the dimension, the total variability space is able to capture more variabilities in the super-vector space including both the speaker and channel variability. As a result, there is substantially smaller variability remaining in the residual space.

Table 1: *Total and residual variability in female and male model.*

	female		male	
	tr($\boldsymbol{TT}^t$)	tr($\boldsymbol{DD}^t$)	tr($\boldsymbol{TT}^t$)	tr($\boldsymbol{DD}^t$)
dim=300	1633.5	361.2	2663.9	361.7
dim=400	1668.9	330.7	2669.8	329.8
dim=500	1680.9	308.4	2720.1	306.9
dim=600	1751.9	290.4	2828.8	289.5

Figure 2 shows the value of EER versus the number of eigenvectors in $\boldsymbol{T}$. Some insights into this figure can be obtained by analyzing the multiple curves simultaneously. With the increase of the rank of $\boldsymbol{T}$, the performance of total factor increases at the very beginning but decreases slightly at the end(after dim = 500). However, the performance of residual factor decreases monotonically and slowly in the whole process. An explanation here is that after the rank of $\boldsymbol{T}$ is saturated, the total variability space is prone to capture more channel variability than speaker variability which is harmful to speaker verification system. Because the total speaker information is a constant and we ignore the channel effects for the residual space, the speaker information in the residual factor is reduced.

Considering the performance of the fusion results and Occam's razor, we choose rank($\boldsymbol{T}$)=500 for further analysis(as indicated by the black dot vertical line in Figure 2. Additionally, α, β are set to 0.7 according to a small development set.). Table 2 and 3 present the EER and minDCF(Minimum value of Detection Cost Function) of total factor, residual factor and different fusion methods under this condition by using direct scoring and SVM, respectively. Figure 3 shows the Detection Error Tradeoff(DET) plots according to Table 2 and 3. An important remark is that KL distance in residual super-vector space is more appropriate than the cosine distance in the residual space. An intuitive explanation here is that KL distance involves the GMM model to weight the different dimensions and different components. From the second and the third row of Table 2 and

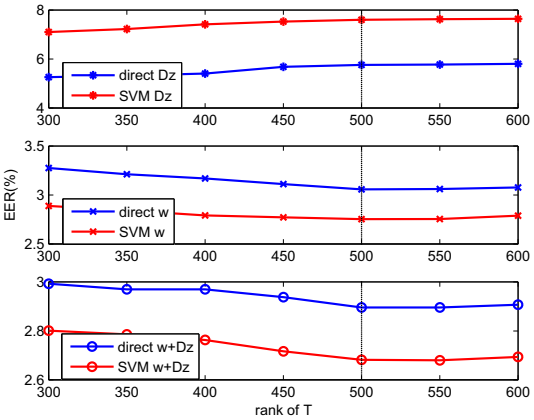

Figure 2: EER on the NIST SRE 2006 set corresponding to different rank of $\boldsymbol{T}$.

3, it seems that direct scoring is more effective than SVM in the residual space $\boldsymbol{z}$ and $\boldsymbol{Dz}$. At last, the best result appears in the last line of Table 3 which is obtained by the combination of cosine kernel on $\boldsymbol{w}$ and KL kernel on $\boldsymbol{Dz}$ for SVM training and achieves 6% relative improvement on EER compared with the original system.

Given the fact that SVM scoring seems better for total variability (the first row in Table 3) and direct scoring better for residual variability(the third row in Table 2), it is interesting to try an empirical combination of them which results in 2.68 and 0.015 for EER and minDCF, respectively. The result shows that this fusion strategy can not improve the performance compared to that of the last line of Table 3.

Table 2: *Comparison of results obtained with and without residual factor using direct scoring. Rank of* $\boldsymbol{T} = 500$.

	EER(%)	minDCF
cosine distance on $\boldsymbol{w}$	3.081	0.0162
cosine distance on $\boldsymbol{z}$	7.272	0.0309
KL distance on $\boldsymbol{Dz}$	5.758	0.0243
cosine $\boldsymbol{w}$+cosine $\boldsymbol{z}$	3.005	0.0160
cosine $\boldsymbol{w}$+KL	**2.895**	**0.0151**

Table 3: *Comparison results obtained with and without residual factor using SVM classification. Rank of* $\boldsymbol{T} = 500$.

	EER(%)	minDCF
cosine kernel on $\boldsymbol{w}$	2.753	0.0155
cosine kernel on $\boldsymbol{z}$	8.879	0.0358
KL kernel on $\boldsymbol{Dz}$	7.601	0.0317
cosine $\boldsymbol{w}$+cosine $\boldsymbol{z}$	2.708	0.0155
cosine $\boldsymbol{w}$+KL	**2.682**	**0.0150**

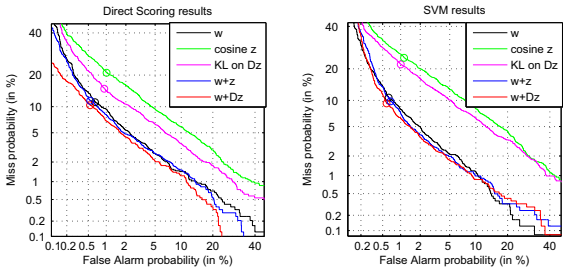

Figure 3: Left panel: Direct scoring results. Right panel: SVM results. Rank of $\boldsymbol{T} = 500$.

5. Conclusion

This work addresses the information loss problem in speaker verification system. We illustrate the fact that the speaker information originally contained in the high dimensional supervector space can not be totally projected onto a much lower dimensional total variability space. We incorporate a residual factor in the total variability modeling to capture the residual variability. According to whether SVM for classification is used or not, two fusion methods are proposed to combine total factor with residual factor. Experiments on NIST SRE 2006 data show that the residual factor improve the performance of the baseline system consistently by restoring the lost speaker information.

6. References

[1] D.A. Reynolds, T.F. Quatieri, and R.B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital signal processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[2] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of interspeaker variability in speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 5, pp. 980–988, 2008.

[3] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*.

[4] Ce Zhang, Rong Zheng, and Bo Xu, "An investigation into direct scoring methods without svm training in speaker verification," in *Proc.Interspeech*. ISCA, 2010, pp. 1437–1440.

[5] W.M. Campbell, "Generalized linear discriminant sequence kernels for speaker recognition," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing- Proceedings*, 2002, vol. 1.

[6] W.M. Campbell, DE Sturim, and DA Reynolds, "Support vector machines using GMM supervectors for speaker verification," *Signal Processing Letters, IEEE*, vol. 13, no. 5, pp. 308–311, 2006.

[7] A.O. Hatch, S. Kajarekar, and A. Stolcke, "Within-class covariance normalization for SVM-based speaker recognition," in *Proc. Interspeech*. Citeseer, 2006, vol. 4.

[8] L. Burget, N. Brummer, D. Reynolds, P. Kenny, J. Pelecanos, R. Vogt, F. Castaldo, N. Dehak, R. Dehak, O. Glembek, Z. Karam, J. Noecker, E. Na, C. Costin, V. Hubeika, S. Kajarekar, N. Scheffer, and J. Cernocky, "Robust speaker recognition over varying channels," *Technical report*, 2009.

[9] "The NIST year 2006 speaker recognition evaluation plan," `http://www.itl.nist.gov/iad/mig/tests/sre/2006/index.html`, 2006.

[10] C. C. Chang and C. J. Lin, "LIBSVM: a library for support vector machines," 2001, `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.