Speaker Recognition using Supervised Probabilistic Principal Component Analysis

Yun Lei, John H.L. Hansen

CRSS: Center for Robust Speech Systems
Erik Jonsson School of Engineering & Computer Science
University of Texas at Dallas, Richardson, Texas 75083,USA
yxl059200@utdallas.edu, John.Hansen@utdallas.edu

Abstract

In this study, a supervised probabilistic principal component analysis (SPPCA) model is proposed in order to integrate the speaker label information into a factor analysis approach using the well-known probabilistic principal component analysis (PPCA) model under a support vector machine (SVM) framework. The latent factor from the proposed model is believed to be more discriminative than one from the PPCA model. The proposed model, combined with different types of intersession compensation techniques in the back-end, is evaluated using the National Institute of Standards and Technology (NIST) Speaker Recognition Evaluation (SRE) 2008 data corpus, along with a comparison to the PPCA model.

Index Terms: speaker recognition, factor analysis, supervised modeling

1. Introduction

Factor analysis approach has been successfully applied for the state-of-the-art speaker recognition systems [1, 2, 3]. Eigenvoice, eigenchannel, and joint factor analysis (JFA) have emerged as efficient approaches in solving the speaker and channel variability under the Gaussian Mixture Model combined with a Universal Background Model(GMM-UBM) framework. At the same time, the SVM has also presented a powerful ability to recognize speakers, using diverse feature inputs, such as the GMM supervector (GMM-SVM) [4]. Under a SVM framework, some successful approaches have been proposed. For example, Nuisance Attribute Projection (NAP) [4] was proposed to remove the channel distortion, and the Within Class Covariance Normalization (WCCN) [5] tried to minimize the expectation error rate of false alarms and false rejections in the one-versus-all case.

In addition, some successful efforts have combined factor analysis and SVM. For example, the speaker GMM supervector obtained from JFA was used as the feature input for the SVM; alternatively, the SVM was trained using the speaker factor from the JFA in [6]. In JFA modeling, the speaker and channel spaces are assumed to be independent, however, some analysis suggests the presence of channel factors in the JFA model also contain speaker information [7]. As such, to release the independence assumption, the total variability model was proposed in [7, 8]. In the total variability model, the speaker and channel spaces are combined and represented by one latent factor, which was used as the feature to train the SVM. In addition, the total variability model also avoids the assumption that the channel distortion of the enrollment data can be ignored in the JFA model, since the total latent variables are extracted from both enrollment and test data.

Since the total variability model is a classical application of the probabilistic principal component analysis (PPCA) [9], the main effect of the total variability model is to reduce the dimension of the GMM supervector so that the latent variables can be estimated well using limited data. As a type of PCA, the total variability model does not need speaker information, however, the speaker label is generally available for the training data. To incorporate the speaker label information into the dimension-reduction projection processing, a supervised PPCA, similar as the method in [10], is proposed here for the speaker recognition task. The SPPCA, as an extension of the PPCA, inherits benefit of the total variability and integrate the speaker information into the projection processing so it is believed that more speaker related information is retained in the SPPCA model rather than more energy in the PPCA model.

This study is organized as follows. The next section begins with a brief review of the total variability model, followed by the description of the proposed SPPCA model. The estimation algorithm of SPPCA is derived based on the EM algorithm of PPCA in Sec.3. In Sec.4, the combination of SPPCA and PPCA, named weighted SPPCA, is derived based on the SPPCA algorithm. Sec.5 presents a series of experimental results with different intersession compensation techniques under the SVM framework. Finally, research findings are summarized along with a discussion of the impact in Sec.6.

2. The SPPCA model

2.1. Total variability review

In [8], the total variability model was proposed in place of two separate models representing the speaker space and channel space in classical JFA modeling. The method is composed of the eigenvectors with the greatest eigenvalues of the total variability covariance matrix, in order to release the independence assumption between speaker and channel spaces. Given an utterance, the speaker and channel dependent GMM supervector M can be written as follows:

$$M = m + T\omega, \tag{1}$$

where m is the UBM supervector, T is a rectangular matrix of low rank and ω is a random vector of dimension D having a

This project was funded by AFRL under a subcontract to RADC Inc. under FA8750-09-C-0067 and partially by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J. Hansen. Approved for public release; distribution unlimited

standard normal distribution $N(0, I)$. The components of the vector ω are the total factors. Clearly, the total variability can be considered as a classical PPCA model here, so the latent variable ω can best explain the data covariance.

In general, some intersession compensation approaches following the PPCA model can be used to set aside the distortion, such as WCCN, LDA, and NAP. The WCCN is an approaches applied in one-versus-all SVM modeling to minimize the expectation error rate of false alarms and missing by using the inverse of the within class covariance. The LDA attempts to find a new projection that minimize the within-class scatter caused by channel effects, and to maximize the between-class scatter between speakers. The NAP tries to find an appropriate projection matrix removing the nuisance direction by using the eigenvectors having the largest eigenvalues of the within class covariance. The WCCN followed by LDA and WCCN followed by NAP have been proven to obtained better performance in [8]. After these front-end process, SVM classifier is used to recognize the speakers.

2.2. SPPCA modeling

As an unsupervised technique, there is no label information required using PPCA so given an utterance without the label, PPCA is the best method to represent the utterance using a small number of uncorrelated variables (ω in Eq.1). However, if label information is available in the training data (such as Switchboard corpora), it is reasonable to extend PPCA including the label information, especially for the classification task, since extracting the latent label related information is more important than one representing the acoustic data in speaker recognition. Considering the utterances in the training data, there are two kinds of observations for each utterance, including the acoustic data and the speaker label. PPCA provides a way to find the latent variable using the acoustic data, however, there is no direct way to use the speaker label into the training processing. To integrate the speaker label into the training process, a similar method to the eigenvoice model is used where the speaker label is represented by a collection of the acoustic data from the same speaker. As such, the supervised PPCA model can be written as follows:

$$\begin{aligned} M &= m + T\omega, \\ S &= m + V\omega, \end{aligned} \tag{2}$$

where S is the supervector of the speaker, V is a rectangular matrix of low rank. The pair (M, S) from Eq.2 are an observed supervector pair of the utterance with the speaker label. In the SPPCA model, the property of conditional independence is kept so the supervector M and S are conditionally independent to each other given the latent variable ω. In other words, the latent variable ω is used to represent a supervector $[M^T S^T]^T$. In the model, the latent variable ω does not only represent the acoustic data of the utterance, but also represents the speaker information of the utterance which is expected to be extracted prior to speaker recognition. In PPCA modeling, since the main energy of the acoustic data is kept using the latent variable ω, it is hoped that the variable ω carries sufficient speaker information for recognition. However, in SPPCA modeling, the variable vector ω directly represents both the utterance and the corresponding speaker label information, so it is certain that the variable ω contains the speaker information.

Since speaker information is used to supervise the training of the projection matrix T, the maximum likelihood solution of this approach is identical to that of LDA from the analysis of [11, 12] so the approach can be considered a probabilistic LDA when compared with PPCA. It is clear that the latent variable ω in the SPPCA model is expected to carry more speaker information than that in the PPCA model so better performance is expected in the following recognition process. It is noted that the SPPCA model is different from the JFA model, where the speaker and channel spaces are assumed to be independent with each other. In the JFA model, the variable ω can be split into two parts ω_s and ω_c, so that the supervector M in Eq.2 can be represented by concatenating ω_s and ω_c, and the supervector S is represented only by ω_s. However, in the SPPCA model, the variable ω can not be split to represent the speaker and channel individually, and must be considered as a whole, thereby avoiding the independence assumption between speaker and channel spaces. The difference between SPPCA and JFA is illustrated in Fig. 1

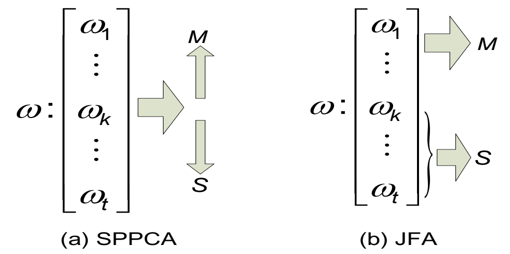

Figure 1: The difference between SPPCA and JFA model.

3. Learning in the SPPCA model

First, some SPPCA parameters are defined: $\mathcal{X}(i)$ denotes the ith utterance data, $\mathcal{Y}(s)$ denotes the sth speaker data including all session from the speaker, where the ith utterance comes from the sth speaker, n_s denotes number of utterances from the sth speaker. The relationship between $\mathcal{X}(i)$ and $\mathcal{Y}(s)$ can be formalized as:

$$\mathcal{X}(i) \subseteq \mathcal{Y}(s), \text{ for the case of speaker of } \mathcal{X}(i) \text{ being } s. \tag{3}$$

Therefore, given a training data pool, the likelihood function, which serves as the estimation criterion, is

$$\begin{aligned} &\prod_i \max_\omega P_{GMM}(\mathcal{X}(i)|m + T\omega), \\ &\prod_i \max_\omega P_{GMM}(\mathcal{Y}(s)/n_s|m + V\omega). \end{aligned} \tag{4}$$

In the training phase, both observations $\mathcal{X}(i)$ and $\mathcal{Y}(s)$ are available, therefore the posterior distribution of the variable ω given ($\mathcal{X}(i)$ and $\mathcal{Y}(s)$) can be calculated as follows:

$$\begin{aligned} P(\omega|\mathcal{X}(i), \mathcal{Y}(s)) &\propto P(\mathcal{X}(i), \mathcal{Y}(s)|\omega)P(\omega) \\ &= P(\mathcal{X}(i)|\omega)P(\mathcal{Y}(s)|\omega)P(\omega), \end{aligned} \tag{5}$$

where the property of the conditional independence is applied. To easily calculate the mean and covariance of the posterior distribution from the existing PPCA approach, the Eq.2 can be rewritten as:

$$\begin{bmatrix} M \\ S \end{bmatrix} = \begin{bmatrix} m \\ m \end{bmatrix} + \begin{bmatrix} T \\ V \end{bmatrix} \omega. \tag{6}$$

Therefore, employing the conditional independence, the mean μ_ω and covariance Σ_ω of the posterior distribution can be obtained as follows:

$$\begin{aligned} \mu_\omega &= l^{-1}(T^*\Sigma^{-1}\tilde{F}(i) + V^*\Sigma^{-1}\tilde{F}(s)/n_s), \\ \Sigma_\omega &= l^{-1}, \end{aligned} \tag{7}$$

where $l = I + T^*\Sigma^{-1}N(i)T + V^*\Sigma^{-1}N(s)/n_sV$, Σ is the covariance matrix of the UBM, and $N(i)$, $N(s)$, $\tilde{F}(i)$, and $\tilde{F}(s)$ are the Baum-Welch statistics defined in [3]. Since the two matrices T and V are independent, the update formulas of both matrices given the variable ω are the same as that for PPCA, which is defined in [3] as well.

In the speaker recognition phase, given an input utterance, which can be the enrollment file or test file, the posterior distribution of the variable ω can be estimated without the speaker information as:

$$\begin{aligned} P(\omega|\mathcal{X}(i)) &\propto P(\mathcal{X}(i)|\omega)P(\omega) \\ &= P(\mathcal{X}(i)|\omega)P(\omega), \end{aligned} \tag{8}$$

and then the mean of the distribution can be calculated as:

$$\mu_\omega = l^{-1}(T^*\Sigma^{-1}\tilde{F}(i)), \tag{9}$$

where $l = I + T^*\Sigma^{-1}N(i)T$. Although the solution looks similar to that in the PPCA model, the projection is supervised since the training of matrix T is influenced by the speaker information.

4. Weighted SPPCA model

In general, PCA is a unsupervised method which aims at extracting a subspace in which the variance of the projected data is maximized, and LDA is a supervised method which maximizes the between-class scatter and simultaneously minimizes the within-class scatter. Therefore, LDA is expected to include more discriminative information than PCA, which is the same in probabilistic case. However, under the SVM speaker recognition framework, after PPCA or SPPCA projection, the following classifier is SVM which is a classifier that maximizes the margin distance between the classes [13]. Since there are different criterion between LDA and SVM, it is possible that the combination of LDA and SVM do not generate a good performance. On the other hand, the combination of PCA and SVM does not have the possible combination problem since PCA minimizes the reconstruction error and only produce a compact representation of the utterance. Therefore, a combination of PPCA and SPPCA, named weighted SPPCA, is proposed to smooth the gap between the criterion of LDA and SVM. Considering the Eq.5, if the term $P(\mathcal{Y}(s)|\omega)$ is ignored, SPPCA model becomes PPCA model. Therefore, the posterior distribution of the variable ω given $\mathcal{X}(i)$ and $\mathcal{Y}(s)$ in the weighted SPPCA model can be formalized as:

$$P(\omega|\mathcal{X}(i), \mathcal{Y}(s)) \propto P(\mathcal{X}(i)|\omega)P^\alpha(\mathcal{Y}(s)|\omega)P(\omega), \tag{10}$$

where α is a parameter from 0 to 1. When α is 0, the model is PPCA model; when α is 1, the model is SPPCA model. Since $P(\mathcal{X}(i)|\omega)$ and $P(\mathcal{Y}(i)|\omega)$ are Gaussian distribution with the same covariance matrix Σ, $P^\alpha(\mathcal{Y}(s)|\omega)$ is also a Gaussian distribution with the covariance matrix $\frac{1}{\alpha}\Sigma$. Therefore, the mean μ_ω and covariance Σ_ω of the posterior distribution in the weighted SPPCA can be obtained as follows:

$$\begin{aligned} \mu_\omega &= l^{-1}(T^*\Sigma^{-1}\tilde{F}(i) + V^*\alpha\Sigma^{-1}\tilde{F}(s)/n_s), \\ \Sigma_\omega &= l^{-1}, \end{aligned} \tag{11}$$

where $l = I + T^*\Sigma^{-1}N(i)T + V^*\alpha\Sigma^{-1}N(s)/n_sV$. As such, the learning process of the weighted SPPCA is the same as the process of the SPPCA except using the Gaussian distribution $P^\alpha(\mathcal{Y}(s)|\omega)$ with the covariance matrix $\frac{1}{\alpha}\Sigma$ in place of $P(\mathcal{Y}(s)|\omega)$.

5. Experiments and Discussion

The proposed algorithm is evaluated based on the 5min-5min telephone-telephone male condition from the NIST 2008 speaker recognition evaluation (SRE) corpus.

5.1. Experimental setup

For parametrization, a 60-dimension feature vector (19 MFCC with log energy + Δ + $\Delta\Delta$) using a 25 ms analysis window with 10 ms shift, filtered by feature warping using a 3-s sliding window is employed [14]. In order to set aside silence frames, a phone recognizer developed from BUT [15] and the energy based voice activity detection (VAD) were employed.

Next, a gender dependent UBM with 1024 mixtures was trained on the NIST 2004, 2005, 2006 SRE enrollment data. The Switchboard II Phase 2 and 3, Switchboard Cellular Part 1 and 2, and the NIST 2004, 2005, 2006 SRE enrollment data were used to train the PPCA and SPPCA model. A total of 400 factors were used in both PPCA and SPPCA models. The same data set as PPCA was used for both LDA and NAP training, and WCCN was trained on the same data as the UBM. The SVM with cosine kernel was used here and the NIST 2004 and 2005 SRE enrollment data were used as the SVM background impostors.

5.2. SPPCA evaluation

In order to evaluate SPPCA and PPCA models directly, all other setups in the system are fixed. Table 1 presents the performance of PPCA and SPPCA under the SVM framework, followed by different session compensation approaches including WCCN, NAP-WCCN, LDA-WCCN. Here, WCCN is always used as a part of SVM and no variables in WCCN needs to be optimized. In NAP and LDA cases, since the performance of the systems depend on the dimension of LDA and NAP, only the best EER and minDCF are presented with the corresponding dimension in the table 1. The best EER and minDCF is a directly evaluation of the speaker recognition ability of the latent factors extracted from PPCA and SPPCA. For example, in PPCA-NAP-WCCN case, the best EER 6.14 is achieved with the NAP of 260 dimension and the best minDCF 2.83 is achieved with the NAP of 300 dimension. From Table 1, SPPCA significantly outperforms PPCA in WCCN and NAP-WCCN cases, and is slightly better than PPCA in LDA-WCCN case. The reason that the performance of PPCA with LDA-WCCN is closer to the performance of SPPCA with LDA-WCCN is that SPPCA is actually a probabilistic LDA so the similar criteria is used in both cases, and 400 factors in PPCA model can cover the most energy of the utterance so the output from PPCA hold the most speaker information which can be extracted by the following LDA. However, if the dimension of PPCA is not large enough, then the speaker information will be partly lost after PPCA extraction, resulting in weak performance.

In addition, the details of the minDCFs of the PPCA-NAP-WCCN and SPPCA-NAP-WCCN SVM systems with different dimensions are shown in Fig.2. After 300 dimension, the minDCF increases sharply in both PPCA and SPPCA cases.

Table 1: Performance of SPPCA and PPCA followed different session compensation approaches for male

Algorithms	EER (%)	minDCF ($\times$ 100)
PPCA-WCCN	6.77	3.12
SPPCA-WCCN	6.39	2.83
PPCA-NAP-WCCN	6.14 (260)	2.83 (300)
SPPCA-NAP-WCCN	6.07 (220)	2.67 (240)
PPCA-LDA-WCCN	6.00 (160)	2.58 (120)
SPPCA-LDA-WCCN	5.82 (300)	2.53 (300)

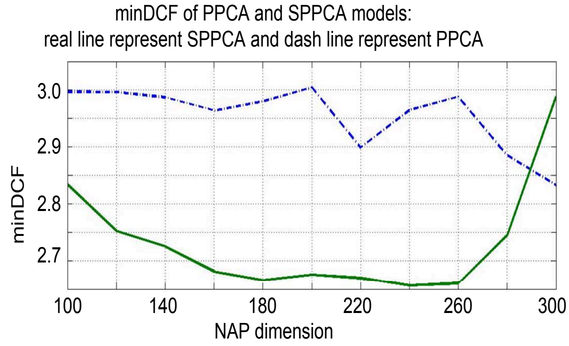

Figure 2: minDCFs of the PPCA-NAP-WCCN and SPPCA-NAP-WCCN SVM systems.

5.3. Evaluation of Weighted SPPCA

Since the different criterion between LDA and SVM, it is not guaranteed that the combination of LDA and SVM results in better performance although the improvement was shown in [8] and our experiments. As such, the weighted SPPCA, as a combination of probabilistic LDA and PCA is evaluated here to present the possible improvement by the combination. Since the training process of SPPCA cost a lot of computation so it is very difficult to evaluate the influences of different α. As such, only the case of $\alpha = 0.2$ is evaluated and shown in Table2. From the table, the performance of the weighted SPPCA with $alpha = 0.2$ is better than the performance of PPCA and worse than the performance of SPPCA. Clearly, the performance of the weighted SPPCA depends on the parameter α, so it is possible that the best performance can be achieved by optimizing the parameter α.

6. Conclusion

This study has considered an supervised PPCA and weighted SPPCA models as an extension of PPCA model under the SVM framework for speaker recognition. The corresponding EM-algorithm based model learning and recognition algorithms are derived. SPPCA is proposed to obtain more discriminative speaker information for recognition and weighted SPPCA is designed to smooth the different criterion between LDA and SVM, resulting in optimizing the performance of speaker recognition. The resulting study shows that the SPPCA model achieves an improvement in the SVM based speaker recognition system and the weighted SPPCA suggests an optimization approach when LDA and SVM are combined for speaker recognition.

Table 2: Performance of weighted SPPCA with $\alpha = 0.2$ followed different session compensation approaches for male

Algorithms	EER (%)	minDCF ($\times$ 100)
SPPCA-WCCN	6.68	2.98
SPPCA-NAP-WCCN	5.98 (250)	2.75 (250)
SPPCA-LDA-WCCN	5.91 (240)	2.54 (160)

7. References

[1] P. Kenny, G. Boulianne, and P. Dumouchel, “Eigenvoice modeling with sparse training data,” *IEEE Trans. on Speech and Audio Processing*, vol. 13, pp. 345–354, May 2005.

[2] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, “Joint factor analysis versus eigenchannels in speaker recognition,” *IEEE Trans. on Audio Speech and Language Processing*, vol. 15, pp. 1435–1447, May 2007.

[3] P. Kenny, P. Oueleet, N. Dehak, V. Gupta, and P. Dumouchel, “A study of inter-speaker variability in speaker verification,” *IEEE Trans. on Audio Speech and Language Processing*, vol. 16, pp. 980–988, July 2008.

[4] W. Campbell, D. Sturim, D. Reynolds, and A. Solomonoff, “Svm based speaker verification using a gmm supervector kernel and nap variability compensation,” vol. 1. ICASSP-2006, Toulouse, 2006.

[5] A. Hatch, S. Kajarekar, and A. Stolcke, “Within-class covariance normalization for svm-based speaker recognition.” Interspeech-2006, Pittsburgh, PA, USA, 2006, pp. 1471–1474.

[6] N. Dehak, P. Kenny, R. Dehak, O. Glembek, P. Dumouchel, L. Burget, and V. Hubeika, “Support vector machines and joint factor analysis for speaker verification.” ICASSP-2009, Taipei, Taiwan, 2009, pp. 4237–4240.

[7] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, “Front-end factor analysis for speaker verification.” [Online]. Available: http://www.crim.ca/perso/patrick.kenny/Najim_TASLP2009.pdf

[8] N. Dehak, R. Dehak, P. Kenny, N. Brummer, P. Ouellet, and P. Dumouchel, “Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification.” Interspeech-2009, Brighton, UK, 2009, pp. 1559–1562.

[9] M. Tipping and C. Bishop, “Mixtures of probabilistic principal component analysers,” *Neural Computation*, vol. 11, pp. 435–474, February 1999.

[10] S. Yu, K. Yu, V. Tresp, H.-P. Kriegel, and M. Wu, “Supervised probabilistic principal component analysis.” Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2006, pp. 464–473.

[11] Y. Zhang and D. Yeung, “Heteroscedastic probabilistic linear discriminant analysis with semi-supervised extension,” vol. 5782. Lecture Notes In Artificial Intelligence, 2009, pp. 602–616.

[12] F. Bach and M. Jordan, “A probabilistic interpretation of canonical correlation analysis,” *Technical Report 688, Department of Statistics, University of California, Berkeley*, 2005.

[13] T. Xiong and V. Cherkassky, “A combined svm and lda approach for classification,” vol. 3. Neural Networks, 2005. IJCNN ’05., 2005, pp. 1455 – 1459.

[14] J. Pelecanos and S. Sridharan, “Feature warping for robust speaker verification.” IEEE Odyssey 2001: The Speaker and Language Recognition Workshop, June 2001, pp. 18–22.

[15] L. Burget, P. Matejka, P. Schwarz, O. Glembek, and J. Cernochy, “Analysis of feature extraction and channel compensation in a gmm speaker recognition system,” *IEEE Trans. on Audio Speech and Language Processing*, vol. 15, pp. 1979–1986, September 2007.