



An Investigation into Direct Scoring Methods without SVM Training in Speaker Verification

Ce Zhang¹, Rong Zheng¹, Bo Xu^{1,2}

¹Digital Content Technology Research Center, Institute of Automation

²National Lab of Pattern Recognition, Institute of Automation
Chinese Academy of Sciences, Beijing 100190, China

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

Abstract

In the paper, we first propose a new method to handle the problem of scoring a test utterance against a speaker model in the JFA Speaker Verification System, called Symmetric Scoring. The SS method is derived from the GMM log-likelihood-ratio approximation and is both symmetrical and efficient. Then we show that SS method and the JFA-SVM system using GMM super-vector space as input have nearly the same form in scoring phase. We also show that the performance of SS method is better than the JFA-SVM system, which indicates that applying the KL kernel function directly to obtain a score in GMM super-vector space is as effective as the JFA-SVM trained using the same kernel. As an inspiration of this direct scoring method in which kernel function is only used to calculate score without SVM training procedure, we try to extend the relationship to speaker factor space and evaluate some results based on different kernels.

Index Terms: Speaker Verification, Joint Factor Analysis, Symmetric Scoring, Direct Scoring, JFA-SVM.

1. Introduction

In recent years, two speaker modeling approaches have been the dominant framework: generative model based on MAP adaption of a UBM [1] and discriminative model based on the separating hyperplane like SVM [2]. Joint Factor Analysis (JFA) approach introduced by Patrick Kenny [3][4] has become the state-of-the-art technique in reducing the effects of channel/session variability based on the generative model-GMM. It can deal with the speaker and session variability simultaneously, by the means of factorizing a speaker- and channel-dependent super-vector into two independent components: speaker-dependent super-vector and channel/session-dependent super-vector.

JFA has been used in the latest NIST evaluations by many sites. However, JFA variants [6] result in different ways to train the JFA hyper-parameters and to calculate Log-Likelihood-Ratio (LLR) scores. Different implementations give different performances and different explanations. In [7], the authors present five methods to produce LLR scores under the JFA framework. The method called Linear Scoring outperformed the rest four methods if we take speed (measured by RTF) into consideration.

However, the form of Linear Scoring is asymmetric that it treats the enrollment and test utterance in an unequal situa-

tion. In this work, we investigate a new method for LLR scoring called Symmetric Scoring (SS) and then compare this to classic JFA scoring methods.

Because of the symmetry of the SS technique, we compare the formula to the KL kernel function used in the SVM based system. Surprisingly, the SS method can be seen as a direct application of KL kernel function in the verification phase and can obtain even better results on the SRE 2006 test sets. As an extension, we also compare the direct scoring method and SVM training method in speaker factor space.

The outline of the paper is as follows. Section 2 describes the theory and technique. Starting with a simple description about JFA in Section 2.1, the sections 2.2 and 2.3 present the SS method and compare it with JFA-SVM. Section 2.4 describes the kernels used in speaker factor space. The details of experimental configurations is presented in section 3 followed by results and analysis of the proposed algorithm in section 4.

2. Theoretical Background

2.1. Joint Factor Analysis

In JFA modeling, each speaker is represented by the weights, means and diagonal-covariances of a GMM which is composed with C multivariate Gaussian densities. The basic and important assumption of JFA is that the super-vector M of a real utterance can be decomposed into a sum of two independent super-vectors: a speaker super-vector s and a channel super-vector c .

Furthermore, let us assume that s and c can be described by a low dimensional vector, y and x respectively. Mathematically, leading to three equations:

$$M = s + c \tag{1}$$

$$s = m + Vy + Dz \tag{2}$$

$$c = Ux \tag{3}$$

where m is speaker- and channel independent super-vector, representing the center of the full parameter space, V is a subspace with high speaker variability, U is a subspace with high inter-session variability and D is a diagonal matrix describing the residual variability.

All the hyper-parameters are estimated by the EM algorithm. We decoupled the estimation of V and D , and then the speaker factor y and channel factor x are jointly estimated.

For more details about JFA, we refer the readers to [3][4][5].

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

2.2. Symmetric Scoring

In [7], the authors proposed five scoring methods including Linear Scoring. There are three assumptions in Linear Scoring : firstly, using fixed alignment of frames to Gaussians by assuming that each frame is generated by a single best scoring Gaussian; secondly, the channel factor x of a test utterance is known (point estimate by MAP) and thirdly only the first order Taylor series (linear term of the model super-vector in the GMM mean space) is kept. The formula of Linear Scoring is given as follows:

$$LS(\chi, s) = \frac{1}{T} (Vy_e + Dz_e)^t \Sigma^{-1} (F - Nm - NUx_t) \quad (4)$$

where subscripts e and t indicate the information produced by enrollment data and test data, T is total frame length of the test utterance. y_e and z_e is the speaker factor and residual factor of the enrolling utterance of a target speaker, x_t is the channel factor of the test utterance. F is a column vector obtained by concatenating the first order statistics. N is a diagonal matrix, whose diagonal blocks are the occupation counts for each Gaussian.

Realizing that this formula can be seen as an inner product between the centralized super-vector of target model and the centralized first-order Baum-Welch statistics of test utterance. It does not treat the two utterance equally.

To overcome this asymmetry, (4) can be rewritten to

$$\begin{aligned} LS(\chi, s) &= \frac{1}{T} \sum_{c=1}^C (V_c y_e + D_c z_e)^t \Sigma_c^{-1} (F_c - N_c m_c - N_c U_c x_t) \\ &= \sum_{c=1}^C \frac{N_c}{T} (V_c y_e + D_c z_e)^t \Sigma_c^{-1} \left(\frac{F_c}{N_c} - m_c - U_c x_t \right) \end{aligned} \quad (5)$$

where subscript c corresponds the c^{th} component in a GMM.

Furthermore, we use the UBM weights to approximate $\frac{N_c}{T}$ and note that $\frac{F_c}{N_c}$ is a Maximum-Likelihood estimation of new GMM mean vector in the EM algorithm. That is,

$$\frac{N_c}{T} \approx \omega_c \quad (6)$$

$$\frac{F_c}{N_c} \approx M_c \quad (7)$$

Substituting these approximations (6) (7) to (5), and therefore we obtain the definition of Symmetric Scoring:

$$\begin{aligned} SS(\chi, s) &= \sum_{c=1}^C \omega_c (V_c y_e + D_c z_e)^t \Sigma_c^{-1} (M_c - m_c - U_c x_t) \\ &= \sum_{c=1}^C \omega_c (V_c y_e + D_c z_e)^t \Sigma_c^{-1} (V_c y_t + D_c z_t) \\ &= \sum_{c=1}^C \omega_c (s_{ec} - m_c)^t \Sigma_c^{-1} (s_{tc} - m_c) \end{aligned} \quad (8)$$

where ω_c and Σ_c are the c^{th} UBM mixture weights and diagonal covariance matrix, s_{ec} and s_{tc} are the subvectors of s_e and s_t respectively, corresponding to the mixture component c .

Note that when we use (8) to compute scores, the enrolling utterance and the testing utterance have the same position. We call this method as Symmetric Scoring.

2.3. Comparison of SS and JFA-SVM in Super-vector Space

The Support Vector Machine(SVM) is a classifier used to find a separator between two classes. In [2], the authors show that the application of SVM in GMM super-vector space gets interesting results. In conventional GMM-super-vector SVM system, KL divergence is used to measure the distance of the two GMMs. Recently, [8] propose a framework that combines JFA with SVM by using the information given by JFA as SVM input features.

As discussed in [2], when applied some approximation to the KL divergence for two GMM probability distributions, the KL kernel function that satisfies the Mercer condition can be given as follows,

$$K_{KL}(s, s') = \sum_{c=1}^C (\sqrt{\omega_c \Sigma_c^{-\frac{1}{2}}} s_c)^t (\sqrt{\omega_c \Sigma_c^{-\frac{1}{2}}} s'_c) \quad (9)$$

Recall that the SS method has a well symmetric property as stated before and if we give a more insight into (8), we can see that it is the KL kernel used in the JFA-SVM algorithm. In other words,

$$\begin{aligned} SS(\chi, s) &= \sum_{c=1}^C (\sqrt{\omega_c \Sigma_c^{-\frac{1}{2}}} (s_{ec} - m_c))^t (\sqrt{\omega_c \Sigma_c^{-\frac{1}{2}}} (s_{tc} - m_c)) \\ &= K_{KL}(s_t, s_e) - K_{KL}(s_t, m) - K_{KL}(s_e, m) + K_{KL}(m, m) \end{aligned} \quad (10)$$

where the third term of (10) can be canceled out through z -norm, and the fourth term will be equal for all target models and test utterances, thus it will also be canceled out. To prove the rationality of canceling these terms, an experiment is carried out on the SRE2006 core condition. Table 1 shows the results.

Table 1: An illustration of the equivalence between SS method and its approximation form in terms of EER and DCF.

| | No-norm | | zt-norm | |
|----------|---------|--------|---------|--------|
| | EER | DCF | EER | DCF |
| SS | 6.34 | 0.0334 | 3.63 | 0.0188 |
| aprox-SS | 43.2 | 0.0992 | 3.63 | 0.0188 |

The aprox-SS in Table 1 denotes the usage of only the first and second term of the last equation in (10). That is, we directly apply $K_{KL}(s_t, s_e) - K_{KL}(s_t, m)$ in the verification phase to score an unknown utterance against a target model. It's clear that when zt -norm is applied, the approximation form of SS method give the same results as the original form in terms of EER and DCF. (EER refers to the equal error rate and DCF to the minimum value of the NIST detection cost function.)

It's interesting that we find the consistent form in scoring phase between discriminative model and generative model. However, the explanation can be different between SS method and JFA-SVM using $K_{KL}(s, s')$ as its kernel.

When we train a SVM using the kernel (9), the decision function in verification phase comes to

$$\begin{aligned} f(s) &= \sum_{i=1}^L \alpha_i t_i K_{KL}(s_i, s) + b \\ &= \left(\sum_{i=1}^L \alpha_i t_i \Phi(s_i) \right)^t \Phi(s) + b \\ &= w^t \Phi(s) + b \end{aligned} \quad (11)$$

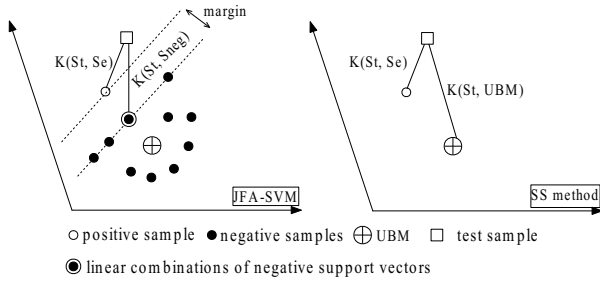


Figure 1: Explanations of JFA-SVM and SS method in super-vector space .

where $\Phi()$ satisfies $K_{KL}(x_i, x_j) = \Phi(x_i) \cdot \Phi(x_j)$. w is a linear combination of L support vectors. This means that we only have to compute a single inner product between the GMM super-vector and w to obtain a score.

Figure 1 shows the difference of scoring function between JFA-SVM and SS method. When we trained a SVM, the test phase can be interpreted as comparing two kernel values: $K(s_t, s_e)$ and $K(s_t, s_{neg})$, where s_{neg} is linear combination of a set of negative support vectors found by SVM training algorithm and is a most confusable position in the feature space. However, SS method replace s_{neg} by UBM . The difference between $K(s_t, s_e)$ and $K(s_t, m)$ is calculated. Section 4 will present the different results obtained by SS method and JFA-SVM algorithm.

2.4. Direct scoring and JFA-SVM in Speaker Factor Space

In section 2.2 and 2.3, we discussed both the similarity and difference between two scoring methods in GMM super-vector space. The SS method can be seen as a direct scoring using a kernel function. As an extension of the relationship between direct calculation and JFA-SVM algorithm, we compare the two methods in speaker factor space. Because the KL kernel can not be directly used in speaker factor space, four classical kernels are tested on the core condition of the NIST 2006 SRE. Precisely, they are given by the following:

$$k(x, y) = x^t y \quad (12)$$

$$k(x, y) = \frac{x^t y}{\|x\| \|y\|} \quad (13)$$

$$k(x, y) = \exp(-\gamma \|x - y\|^2) \quad (14)$$

$$k(x, y) = (\gamma x^t y + r)^d \quad (15)$$

We do not make use of residual factor z .

3. Experiment

3.1. Data Set

The results of our experiments are reported on NIST 2006 SRE corpus . Both train and test conversations have an average duration of five minutes and there are no cross-gender trials.

3.2. 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. MVA filtering (Mean subtraction, Variance normalization, and ARMA filtering) 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. A simple energy based VAD is used to prune out silence frames. Finally, these 39-dimensional feature vectors are subjected to feature warping using a 3s sliding window.

3.3. Configuration of JFA

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 JFA system, eigenvoice subspace spanned by 300 vectors 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 only make use of those speakers for which five or more recordings are available.

The matrix with a set of 100 eigenchannels is trained on the same data set as used in eigenvoice, so as residual space. All the training procedures of hyper-parameters carry out ten iterations in the EM algorithm.

3.4. Normalization and SVM negative samples

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

For SVM training, imposters are selected from the same data set as z -norm. We implemented the SVM training procedure with LIBSVM.

4. Results and Discussion

4.1. SS method and KL Kernel

We start with the comparison between the results obtained by SS method and KL kernel based JFA-SVM. The performance of Linear Scoring is also listed. We compare these methods on two data sets: 1conv4w-1conv4w and 1conv4w-1conmic of the NIST 2006 SRE. Another set of 50 eigenchannels are trained on SRE 2005 microphone data to concatenate with 100 telephone eigenchannels when cross channel trials involved. Both the raw scores and z -norm scores are presented in the table 2 and 3 . On phone-phone condition, Linear Scoring performs a little better than SS method. However, SS method gives better result than Linear Scoring on the phone-microphone condition. We explain this by the fact that SS method is symmetric between training and testing condition and it only compares the same part of the two utterances regardless of the cross channel information. We also find that when score normalization is not available, JFA-SVM always performs best.

Figure 2 and figure 3 compares three methods on the core condition and on the tel-mic condition of NIST 2006 SRE respectively. These curves are based on normalized scores.

4.2. Direct scoring and JFA-SVM algorithm in speaker factor space

In this section, we present the results obtained by direct scoring and JFA-SVM in speaker factor space. Both methods are tested using the linear, cosine, Gaussian and poly kernel functions. Table 4 gives these results.

The results show that the performance of direct scoring in speaker factor space is worse than that of SVM training fol-

Table 2: Comparison results between SS method and KL kernel SVM in terms of EER and DCF. The results are given on the core condition of the NIST 2006 SRE.

| | No-norm | | zt-norm | |
|---------------|-------------|---------------|-------------|---------------|
| | EER | DCF | EER | DCF |
| LS | 5.95 | 0.0296 | 3.53 | 0.0178 |
| KL kernel SVM | 5.23 | 0.0261 | 4.90 | 0.0233 |
| SS | 6.34 | 0.0334 | 3.63 | 0.0188 |

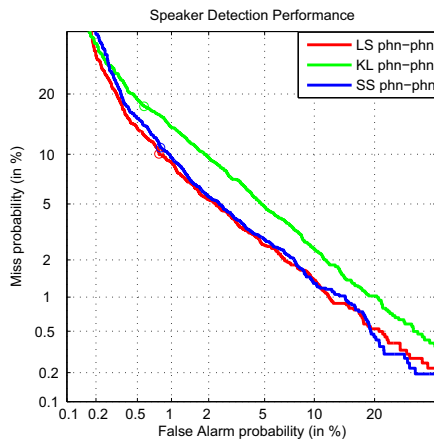


Figure 2: DET curves on the core condition of NIST 2006 SRE corresponding to table 2.

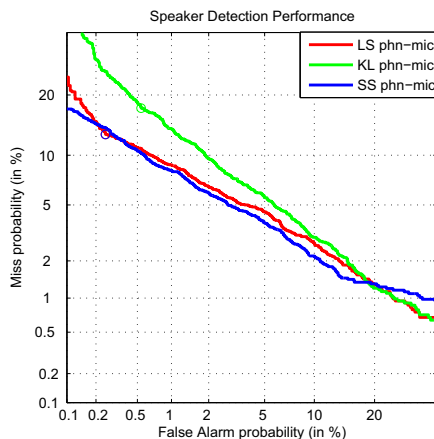


Figure 3: DET curves on the tel-mic condition of NIST 2006 SRE corresponding to table 3.

lowed by a SVM prediction procedure. These results can be explained as that the speaker factors obtained by JFA model is non-separable and confusable because of the low dimensional speaker factor space.

5. Conclusion

In this paper, we investigated into different direct scoring methods for speaker verification on the NIST 2006 SRE platform. We showed that the SS method, a kind of direct scoring in GMM super-vector space, performs as well as Linear Scoring on the tel-tel condition and outperforms Linear Scoring while cross-channel trials involved. This can be explained by the symmetry of SS method which only take the same variability into consid-

Table 3: Comparison results between SS method and KL kernel SVM in terms of EER and DCF. The results are given on the 1conv4w-1convmic condition of the NIST 2006 SRE.

| | No-norm | | zt-norm | |
|---------------|-------------|---------------|-------------|---------------|
| | EER | DCF | EER | DCF |
| LS | 9.19 | 0.0323 | 4.69 | 0.0154 |
| KL kernel SVM | 8.09 | 0.0284 | 5.33 | 0.0228 |
| SS | 8.57 | 0.0333 | 4.25 | 0.0154 |

Table 4: Comparison results between JFA-SVM and direct scoring method corresponding to the same kernel function in terms of EER and DCF on the core condition of the NIST 2006 SRE. zt-norm is used as score normalization.

| | direct score | | SVM based | |
|-----------------|--------------|--------|-----------|--------|
| | EER | DCF | EER | DCF |
| Linear kernel | 4.68 | 0.0238 | 4.75 | 0.0224 |
| Cosine kernel | 4.67 | 0.0242 | 4.63 | 0.0224 |
| Gaussian kernel | 5.56 | 0.0302 | 5.21 | 0.0278 |
| Poly kernel | 4.87 | 0.0254 | 4.71 | 0.0227 |

eration. We also showed that direct scoring applied in super-vector space gives better result than JFA-SVM using KL kernel function. However, when applied in speaker factor space, the direct scoring using classical kernel functions can not make any improvements. The explanation is that due to the high dimensional super-vector, the data space is sparse and well separable. But it is not the case in the speaker factor space which has a much lower feature dimension. And therefore, the data in speaker factor space is not well separable that we should make use of SVM to find the separating line.

6. References

- [1] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Process.*, vol. 10, pp. 19-41, 2000.
- [2] W. M. Campbell, D. E. Sturim, D. A. Reynolds, "Support vector machines using GMM supervectors for speaker verification," *IEEE Sign. Process. Lett.* 13(5), 308-311 (2006)
- [3] P. Kenny, "Joint factor analysis of speaker and session variability: Theory and algorithms," *Tech. Rep. CRIM-06/08-13*, 2005 [Online]. Available: <http://www.crim.ca/perso/patrick.kenny>
- [4] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 4, pp. 1435-1447, May 2007.
- [5] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A Study of Inter-Speaker Variability in Speaker Verification," *IEEE Trans. Audio, Speech and Language Processing*, 16(5):980-988.
- [6] L. Burget, P. Matejka, V. Hubeika and J. Cernocky, "Investigation into variants of Joint Factor Analysis for speaker recognition," in *Proc. Interspeech 2009*, Brighton, U.K., Sept. 2009, pp. 1263-1266.
- [7] O. Glembek, L. Burget, N. Dehak, N. Brummer and P. Kenny, "Comparison of scoring methods used in speaker recognition with Joint Factor Analysis," *Proc. ICASSP 2009*, pp. 4057-4010.
- [8] N. Dehak, P. Kenny, R. Dehak, O. Glembek, P. Dumouchel, L. Burget, V. Hubeika and F. Castaldo, "Support vector machines and joint factor analysis for speaker verification," *Proc. ICASSP 2009*, pp. 4237-4240.