SVM-BASED SPEAKER VERIFICATION BY LOCATION IN THE SPACE OF REFERENCE SPEAKERS

Xianyu Zhao[1], Yuan Dong[1,2], Hao Yang[2], Jian Zhao[2], Haila Wang[1]

[1]France Telecom R&D Center (Beijing), Beijing, 100080, P.R.China
[2]Beijing University of Posts and Telecommunications, Beijing, 100876, P.R.China
{xianyu.zhao, yuan.dong, haila.wang}@orange-ft.com

ABSTRACT

In this paper, we investigate SVM-based speaker verification by location in the space of reference speakers. Speaker location is represented by a vector of log-likelihoods of utterance data given reference speaker models. Channel or session variability in speaker locations due to microphone, acoustic environments etc. would impair verification performance. To reduce such variability, Within-Class Covariance Normalization (WCCN), Nuisance Attribute Projection (NAP) and their combination are applied, and significant performance improvements are obtained. Experimental results on a NIST SRE 2006 task show that this location SVM system achieves comparable performance to a state-of-art cepstral GMM-UBM verification system, and their fusion can give additional performance gains.

Index Terms— supporting vector machines, session variability, speaker location, speaker recognition, speech processing.

1. INTRODUCTION

The use of a set of reference speakers to model a speaker has been extensively studied for many tasks in speech processing, e.g. rapid speaker adaptation [1, 2], speaker recognition [3, 4] and tracking [5]. In [3], anchor models and speaker location vector were proposed for speaker verification and indexing purposes; although its computational efficiency for speaker indexing in large audio database was shown to be superior to that of cepstral GMM-UBM, the verification performance by location fell short of a state-of-art cepstral GMM-UBM system. This is related with the issue of channel or session variability in speaker location vectors (due to microphones, acoustic environments, etc.). In [4], statistically modeling of speaker location vectors with a Gaussian distribution was proposed to address such kind of variability, and some promising results were reported. As we will show later, the statistical approach in [4] has much in common with using proper kernel function in SVM-based classifiers.

In recent years, support vector machines have become one of the most important and widely used classification techniques in the field of speaker recognition. In [6, 7], the Within-Class Covariance Normalization (WCCN) technique for training generalized linear kernel of SVMs was introduced, which identifies and optimally weights directions of underlying feature space to maximize task-relevant information. In [8], Nuisance Attribute Projection (NAP) was developed to do projection to remove dimensions from the SVM feature space that are irrelevant to the classification problem.

In this study, we apply WCCN and NAP to reduce or compensate session variability in speaker location vectors, and performance improvements are obtained for location SVM verification systems; furthermore, their combination is found to provide significant performance improvements of up to 19% in Equal Error Rate (EER) and 18% in the NIST minimum Detection Cost Function (DCF) than single technique alone.

This paper is organized as follows. In Section 2, we describe the concepts of anchor models and speaker location. In Section 3, we present the construction of location SVM systems and some discussions about WCCN and NAP. In Section 4, we report experimental results on a NIST speaker recognition evaluation (SRE) 2006 task. We end with conclusions and future work in Section 5.

2. ANCHOR MODELS AND SPEAKER LOCATION

Speaker location in the space of reference speakers is represented by the following vector, $\mathbf{v}$ [3] – [5]:

$$\mathbf{v}_{\mathbf{x}} = \left[\tilde{p}\left(\mathbf{x}\middle|\bar{\lambda}_1\right) \quad \tilde{p}\left(\mathbf{x}\middle|\bar{\lambda}_2\right) \quad \cdots \quad \tilde{p}\left(\mathbf{x}\middle|\bar{\lambda}_E\right) \right]^T \qquad (1)$$

where $\left\{\bar{\lambda}_i; i = 1, 2, \cdots, E\right\}$ is a set of well trained reference speaker models (called anchor models), which are usually modeled as Gaussian Mixture Models (GMM) and adapted from a Universal Background Model (UBM) [9]; $\tilde{p}\left(\mathbf{x}\middle|\bar{\lambda}_i\right)$ is the normalized log-likelihood of the speaker utterance data $\mathbf{x}$ (of L acoustic feature vectors) for the i-th anchor

model, $\bar{\lambda}_i$, relative to the Universal Background Model, λ_{UBM},

$$\tilde{p}\left(\mathbf{x}\middle|\bar{\lambda}_i\right)=\frac{1}{L}\log\frac{p\left(\mathbf{x}\middle|\bar{\lambda}_i\right)}{p\left(\mathbf{x}\middle|\lambda_{UBM}\right)} \tag{2}$$

3. SVM-BASED SPEAKER VERIFICATION BY LOCATION IN THE SPACE OF REFERENCE SPEAKERS

In the speaker location SVM systems, the location vector $\mathbf{v}$ is treated as input feature and modeled using support vector machines. In the standard formulation, a SVM, $f(\mathbf{v})$, is given by

$$\begin{aligned} f(\mathbf{v}) &= \sum_{i=1}^{M}\alpha_i k\left(\mathbf{v},\bar{\mathbf{v}}_i\right)+b \\ &= \sum_{i=1}^{M}\alpha_i\left\langle\Phi(\mathbf{v}),\Phi\left(\bar{\mathbf{v}}_i\right)\right\rangle+b \end{aligned} \tag{3}$$

where $k\left(\mathbf{v}_1,\mathbf{v}_2\right)$ is a kernel function and $\Phi(\mathbf{v})$ is a feature transformation function. The relationship between the feature transformation Φ and the kernel function k is that

$$k\left(\mathbf{v}_1,\mathbf{v}_2\right)=\left\langle\Phi\left(\mathbf{v}_1\right),\Phi\left(\mathbf{v}_2\right)\right\rangle \tag{4}$$

where $\langle\bullet,\bullet\rangle$ stands for inner product. The b and $\left\{\alpha_i,\bar{\mathbf{v}}_i;i=1,\cdots,M\right\}$ are obtained through a training process that maximizes the margin between two classes (positive vs. negative). One of the critical problems in SVM-based systems is the choice of kernel or feature transformation function. In this study, the generalized linear kernel is used, i.e.

$$k\left(\mathbf{v}_1,\mathbf{v}_2\right)=\mathbf{v}_1^T\mathbf{R}\mathbf{v}_2 \tag{5}$$

Three location SVM systems are constructed based on different choices of kernel parameterization or feature transformation function.

1. In the first system (denoted as **Location SVM I**), the $\mathbf{R}$ in (5) is simply set to be the identity matrix, $\mathbf{I}$, i.e. a linear kernel SVM.
2. In the second system (denoted as **Location SVM WCCN**), $\mathbf{R}$ is set to be $\mathbf{W}^{-1}$, i.e.
$$k\left(\mathbf{v}_1,\mathbf{v}_2\right)=\mathbf{v}_1^T\mathbf{W}^{-1}\mathbf{v}_2 \tag{6}$$
where $\mathbf{W}$ is the expected within-class covariance matrix of location vectors over all classes. It can be represented mathematically as [6, 7],
$$\mathbf{W}=\sum_{j=1}^{J}p_j\Sigma_j \tag{7}$$
where p_j and Σ_j represent respectively the priori probability and covariance matrix of the *j*-th class, *J* is the total number of classes. Through Within-Class Covariance Normalization (WCCN), the generalized linear kernel in (6) can be implemented through the following feature transformation function,
$$\Phi(\mathbf{v})=\mathbf{\Lambda}^{-1/2}\mathbf{U}^T\mathbf{v} \tag{8}$$
where $\mathbf{\Lambda}$ and $\mathbf{U}$ can be obtained through the eigenvalue decomposition of $\mathbf{W}$,
$$\mathbf{W}=\mathbf{U}\mathbf{\Lambda}\mathbf{U}^T \tag{9}$$
$\mathbf{\Lambda}$ is a diagonal matrix with $\mathbf{W}$'s eigenvalues on the main diagonal, columns of $\mathbf{U}$ are eigenvectors of $\mathbf{W}$.
3. In the third SVM system (denoted as **Location SVM NAP**), Nuisance Attribute Projection (NAP) is used to do feature projection. In this case, the feature transformation function Φ is given by
$$\Phi(\mathbf{v})=\left(\mathbf{I}-\mathbf{U}_m\mathbf{U}_m^T\right)\mathbf{v} \tag{10}$$
where $\mathbf{U}_m$ is a matrix whose columns are composed of *m* eigenvectors with largest eigenvalues of the within-class covariance matrix, $\mathbf{W}$. (In our following experiments, the value of *m* was set to be 10.)

3.1. Discussion

In the above location SVM systems, both WCCN and NAP use the within-class covariance matrix as means to capture session variability in speaker location. After projecting the original feature onto the orthogonal space spanned by the eigenvectors of $\mathbf{W}$, NAP totally discards such directions which cause much variability (corresponding to have large eigenvalues of $\mathbf{W}$) in the kernel and restricts classification to the remaining subspace; alternatively, WCCN inversely weights the contribution of each direction based on their extent of variability. In the following section, we will investigate their performance for location SVM systems and explore if they could be combined properly.

In [4], session variability in speaker location was addressed, and the within-class covariance matrix was used in statistically modeling of speaker location vectors. Speaker verification was carried out through

$$\begin{aligned} \text{score} &= \log\frac{N\left(\mathbf{v}_\mathbf{x}\middle|\mu_{spk},\mathbf{W}\right)}{N\left(\mathbf{v}_\mathbf{x}\middle|\mu_{UBM},\mathbf{W}\right)} \\ &= \left(\mathbf{v}_\mathbf{x}-\mu_{spk}\right)^T\mathbf{W}^{-1}\left(\mathbf{v}_\mathbf{x}-\mu_{spk}\right)-\left(\mathbf{v}_\mathbf{x}-\mu_{UBM}\right)^T\mathbf{W}^{-1}\left(\mathbf{v}_\mathbf{x}-\mu_{UBM}\right) \end{aligned} \tag{11}$$

where μ_{spk} and μ_{UBM} are respectively the average of location vectors of target and background speakers. This scheme has much in common with using $\mathbf{W}^{-1}$ in generalized linear kernel. In Section 4, we will briefly compare its performance with SVM-based systems.

3.2. Implementation

During training stage, a database which contains multiple sessions of recordings for each speaker is used. For each session of utterance, a location vector is calculated; and

location vectors of utterances from the same speaker are assigned to one class. The within-class covariance matrix is estimated as a weighted average of each class's sample covariance matrix,

$$\hat{\mathbf{W}} = \frac{1}{N}\sum_{j=1}^{J} N_j \hat{\mathbf{\Sigma}}_j \qquad (12)$$

where N is the total number of background speaker location vectors and N_j is the number of vectors in the j-th class whose sample covariance matrix is $\hat{\mathbf{\Sigma}}_j$.

We use SVMTorch to train SVM-based speaker models [10]. Each speaker model is trained using the location vectors of the speaker's enrollment utterances as positive examples, and the location vectors of all utterances from background speakers as negative examples.

4. EXPERIMENTS AND RESULTS

In this section, we report speaker verification experiments by location in the space of reference speakers. Section 4.1 presents some general experiment setup information about the task, corpora, features and configuration of the reference speaker space. The results of these experiments are discussed in Section 4.2.

4.1. Experiment Setup

Speaker verification experiments were conducted on the 2006 NIST SRE corpus [11]. We focused on the single-side 1 conversation train, single-side 1 conversation test task. This task involves 3,612 true trials and 47,836 false trials.

A cepstral GMM-UBM system [9] was setup as a baseline for performance comparison; it also provided the basis for the construction of anchor models and calculation of speaker location in the reference speaker space. For cepstral feature extraction, a 13-dimensional PLP is calculated every 10 ms using a 25ms Hamming window. First, second and third order derivatives over a ±2 frame span are computed and appended to each feature vector, which results in dimensionality 52. Heteroscedastic linear discriminant analysis (HLDA) is then used to decorrelate the features and reduce the dimensionality from 52 to 51. RASTA, feature mapping and histogram equalization (HEQ) are applied to improve channel and noise robustness. Gender independent UBM with 2048 Gaussians is trained using about 40 hours of data from the Switchboard corpora. Speaker GMM models are adapted from UBM by MAP-adaptation with relevance fact set to be 16 (only the means are adapted).

Some reference speakers are selected from the Switchboard corpora. Their models are adapted from the UBM and used as anchor models in the reference speaker space. Considering computational efficiency and no cross gender trials in the NIST SRE task, gender dependent reference speaker spaces are constructed with 248 anchor models in the male space and 335 in the female space. Male and female speakers are represented and tested separately in corresponding reference space.

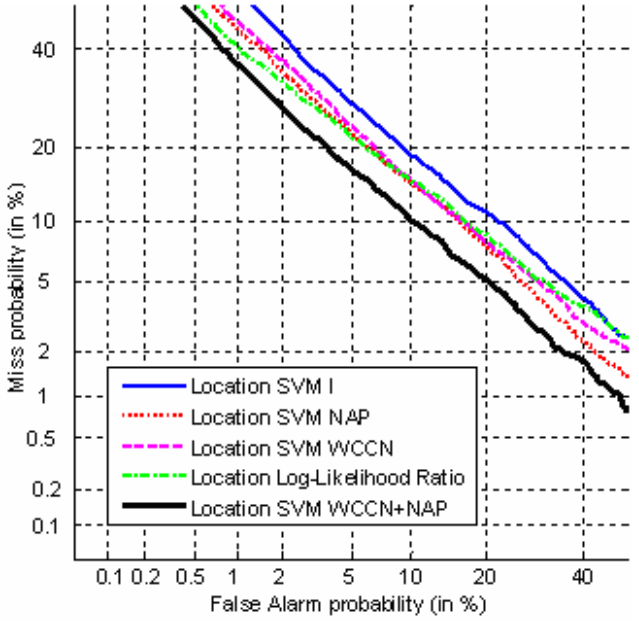

Fig.1 Verification performance by location in the space of reference speakers

In the reference speaker spaces, location vectors of utterances in a subset of the 2004 NIST SRE corpus (the single-side, 1 conversation train, single-side, 1 conversation test part) are calculated as negative examples of SVM training; and these background location vectors, whose speaker id information is extracted from the answer key provided by NIST, are also used to estimate gender dependent within-class covariance matrices. In the male reference speaker space, there are 763 background location vectors from 130 speakers; in the female space, there are 1027 background location vectors from 191 speakers (on average, each background speaker has about 5 sessions).

4.2. Results

In Fig.1, we show verification performance for various systems that use speaker location in the reference speaker spaces. From this figure, it can be seen that "Location SVM WCCN" and "Location SVM NAP" have comparable performance; and they are better than that of the basic "Location SVM I" system. After reducing or compensating session variability in location vectors through WCCN or NAP, speaker location in the reference spaces becomes more stabilized; and this facilitates SVM-based verification.

In this figure, "Location SVM WCCN+NAP" stands for the combination of "Location SVM WCCN" and "Location SVM NAP" systems. This combination was done in the score level, and scores from these two systems were linearly combined with equal weights, i.e.

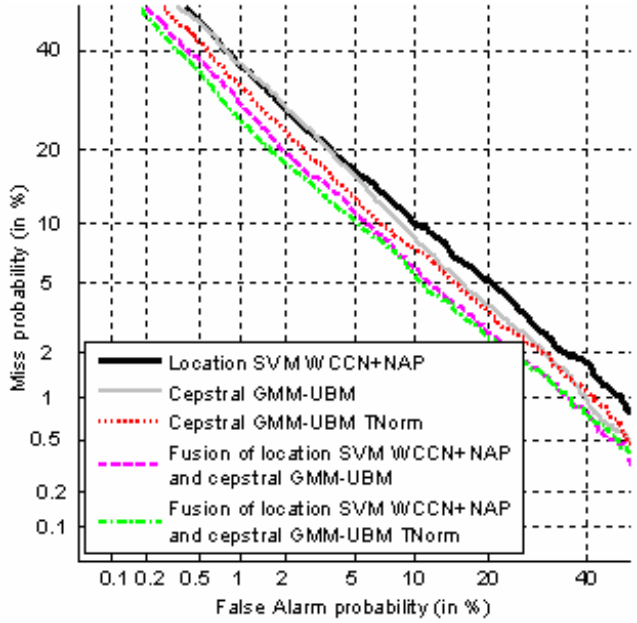

Fig.2 Verification performance comparison of location SVM and cepstral GMM-UBM systems

$$S_{Loc.\ SVM\ \ WCCN+NAP} = S_{Loc.\ SVM\ \ WCCN} + S_{Loc.\ SVM\ \ NAP} \qquad (13)$$

The "Location SVM WCCN+NAP" system achieved significant performance improvements of up to 19% in EER and 18% in the NIST minimum DCF [11] compared with "Location SVM WCCN". This is an interesting result which indicates that "Location SVM WCCN" and "Location SVM NAP" could provide complementary decision results because of their different ways of disposing of session variability.

In Fig.1, "Location Log-Likelihood Ratio" shows verification performance using the score given by equation (11), it can be seen that it has comparable EER with "Location SVM WCCN".

In Fig.2, the performance of location SVM system was compared with the cepstral GMM-UBM systems. It can be seen that "Location SVM WCCN+NAP" achieved comparable performance to the GMM-UBM baseline which is denoted as "Cepstral GMM-UBM" in Fig.2. These two systems were then fused using logistic linear regression developed by Brümmer [12], and better performance was obtained than that of single system alone. The "Cepstral GMM-UBM TNorm" system stands for cepstral GMM-UBM system with TNorm score normalization [13]. The set of imposter speakers used for TNorm is the same as that for anchor models. As shown in Fig.2, the fusion of "Location SVM WCCN+NAP" and "Cepstral GMM-UBM" provided better performance than that of a state-of-art cepstral GMM-UBM system with TNorm; and the fusion of "Location SVM WCCN+NAP" and "Cepstral GMM-UBM TNorm" can further gain performance improvements.

5. CONCLUSION

We have proposed a speaker verification approach based on SVM modeling of speaker location in the space of reference speakers. Steady performance improvements are achieved after applying Within-Class Covariance Normalization and Nuisance Attribute Projection to reduce or compensate session variability in location vectors. It is also found that the SVM system using WCCN normalized location and the system using NAP projected location could provide complementary decision results and be combined in score level to significantly improve performance. The combined location SVM system achieves comparable performance to cepstral GMM-UBM systems on a NIST SRE 2006 task. Further performance improvements are obtained by fusing the location SVM and cepstral GMM-UBM systems. Refining the reference speaker space for location SVM systems will be studied in future work.

6. REFERENCES

[1] T. Hazen, "The use of speaker correlation information for automatic speech recognition," Ph.D. Thesis, Mass.Inst.Technol., Cambridge, Jan. 1998.

[2] R. Kuhn, J.-C. Junqua, P. Nguyen and N. Niedzielski, "Rapid speaker adaptation in Eigenvoice space," *IEEE Trans. Speech and Audio Processing*, vol.8, no.6, pp. 695-707, Nov. 2000.

[3] D. Sturim, D. Reynolds, E. Singer, and J. P. Campbell, "Speaker indexing in large audio database using anchor models," in *Proc. ICASSP'2001*, pp. 429-432, 2001.

[4] Y. Mami, D. Charlet, "Speaker recognition by location in the space of reference speakers," *Speech Communication*, vol.48, pp. 127-141, 2006.

[5] M. Collet, D. Charlet and F. Bimbot, "A Weighted measure of similarity for speaker tracking," in *Proc. IEEE Odyssey 2006: The Speaker and Language Recognition Workshop*, 2006.

[6] A. O. Hatch, A. Stolcke, "Generalized linear kernels for one-versus-all classification: application to speaker recognition," in *Proc. ICASSP'2006*, 2006.

[7] A. O. Hatch, S. Kajarekar and A. Stolcke, "Within-Class Covariance Normalization for SVM-based Speaker Recognition," in *Proc. ICSLP'2006*, 2006.

[8] A. Solomonoff, W. M. Campbell and I. Boardman, "Advances in channel compensation for SVM speaker recognition," in *Proc. ICASSP'2005*, 2005.

[9] D. Reynolds, T. Quatieri and R. Dunn, "Speaker verification using adapted Gaussian Mixture Models," *Digital Signal Processing*, vol.10, pp. 19-41, 2000.

[10] R. Collobert, S. Bengio, "SVMTorch: Support vector machines for large–scale regression problems," *Journal of Machine Learning Research*, vol. 1, pp. 143-160, 2001.

[11] "The NIST 2006 speaker recognition evaluation plan," http://www.nist.gov/speech/tests/spk/spk/2006/ .

[12] N. Brümmer, J. Preez, "Application-independent evaluation of speaker detection," *Computer, Speech and Language*, vol. 20, pp. 230-275, 2006.

[13] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, pp. 42-54, 2000.