EXPLOITING PROSODIC INFORMATION FOR SPEAKER RECOGNITION

Yanhua Long[1], *Bin Ma*[2], *Haizhou Li*[2,3], *Wu Guo*[1], *Eng Siong Chng*[3], *Lirong Dai*[1]

[1]iFly Speech Lab, EEIS, University of Science and Technology of China (USTC), China
[2]Institute for Infocomm Research (I^{2}R), Singapore
[3]Nanyang Technology University, Singapore
lyhsa@mail.ustc.edu.cn, {mabin,hli}@i2r.a-star.edu.sg,
{guowu, lrdai}@ustc.edu.cn, ASESChng@ntu.edu.sg

ABSTRACT

In this paper, we study speaker characterization using prosodic supervectors with negative within-class covariance normalization (NWCCN) projection and speaker modeling with support vector regression (SVR). We also propose a segmental weight fusion (SWF) technique that combines acoustic and prosodic subsystems effectively, despite the big performance gap between the subsystems. We validate the effectiveness of our proposed techniques on the NIST 2006 Speaker Recognition Evaluation (SRE) in comparison with other prominent solutions. The experiments have reported competitive results of 17.72% Equal Error Rate for the prosodic subsystem alone and 4.50% for the fusion system on NIST 2006 SRE core test condition.

Index Terms— Negative within-class covariance normalization, Support vector regression, Segmental weight fusion.

1. INTRODUCTION

One of the fundamental issues in speaker recognition is to characterize speakers by discriminative cues. The cues, varying from low level acoustic features to high level prosodic features, reflect different aspects of speaker characteristics. Another issue is how to effective organize and combine the speaker cues in the speaker recognition system design for the best performance. It has been proved that fusion of multiple sources of information will boosts the performance of speaker recognition.

Significant improvements have been achieved through exploiting the acoustic features representing the temporal properties of speech spectrum, such as Gaussian mixture modeling based on universal backgrouond model (GMM-UBM) [1], generalized linear discriminant sequence (GLDS) kernel by expanding acoustic features using a monomial basis [2], support vector machine modeling on GMM supervectors (GMM-SVM) [3], MLLR transforms as features for support vector machine modeling (MLLR-SVM) [4], and joint factor analysis (JFA) for compensating session and channel variabilities [5]. In recent years, prosodic features [6][7][8] have attracted much attention for their robustness to channel variability and their complementary advantage to the acoustic features in speaker recognition.

There are two major challenges in applying prosodic features in speaker recognition. One is the availability of training data. Compared with acoustic features, prosodic features are more easily affected by the size of training data. For example, pitch, one of the important prosodic features, does not exist in unvoiced region. Another challenge is how to make good use of their complementary advantage in a fusion system with both acoustic features and prosodic features, where the recognition accuracy using prosodic features is generally much lower than that using acoustic features. Improper score fusion may not result in an improved performance over the acoustic features.

In this paper, we propose a new strategy for modeling prosodic features in speaker recognition by applying the negative within-class covariance normalization (NWCCN) on the supervectors of prosodic features. The NWCCN is a variation of WCCN [9] which makes use of the negative data from impostor speakers to estimate the expected within-class covariance matrix. We further apply support vector regression (SVR) [10] to model the prosodic supervectors for better generalization and approximation.

We design a segmental weight fusion (SWF) algorithm to effectively fuse both acoustic scores and prosodic scores. The score fusion are deployed in three score regions to minimize the errors separately. This fusion algorithm works well even when there is big performance gap between the acoustic features and prosodic features.

The paper is organized as follows. In Section 2, we introduce the SVR-NWCCN modeling approach on the prosodic supervectors. In Section 3, we present the segmental weight fusion algorithm by combining both acoustic and prosodic scores. The experimental results on the 2006 NIST Speaker Recognition Evaluation (SRE) corpus are shown in Section 4. Finally we conclude in Section 5.

This work is partly supported by Microsoft Research Funding (07122803).

2. SVR-NWCCN MODELING FOR PITCH FEATURES

The SVR-NWCCN (support vector regression - negative within-class covariance normalization) approach includes three major modules: prosodic feature extraction for prosodic supervectors, NWCCN projection for channel and session variability normalization, and SVR modeling for speaker recognition.

2.1. Prosodic Feature Extraction

Six dimensional prosodic features are used in the speaker recognition. They are log value of pitch, log value of energy, and their first and second order derivatives. The pitch is extracted at every 25ms frames using the Praat toolkit [11]. Since the pitch values are estimated with the autocorrelation method, they are not valid in unvoiced frames. We scale the log pitch linearly into the range of [-1, 1]. To calculate the energies, the speech signals are further processed with RASTA filtering, voice activity detection (VAD), utterance-based cepstral mean subtraction (CMS), and short-time Gaussianization [12]. The Gaussian mixture models are trained on the 6-dimension feature vectors and the means vectors of all the mixture components compose the prosodic supervector.

2.2. NWCCN Projection

WCCN is a technique for training generalized linear kernels that was recently introduced in [9], where **W** is the expected within-class covariance matrix over all classes in the training data. Negative within-class covariance normalization (NWCCN) is designed for the speaker modelling with limited training data from the target speaker in the one-versus-all classification while the dimensionality of the feature space of supervectors is very high [9]. The negative samples from all impostor speakers with the class (speaker) labels are used to identify orthonormal directions in feature space.

With a set of negative prosodic supervectors, we define the expected within-class covariance matrix in NWCCN projection as follows:

$$\mathbf{W} = (1-\lambda)\{\Sigma_{i=1}^{M} p(i) E\{(X_i - \bar{X}_i)(X_i - \bar{X}_i)^T\}\} + \lambda \mathbf{I}, \quad (1)$$

where *M* is the number of negative classes, X_i is a random vector from class *i*, $\bar{X}_i$ is the expected value of X_i, *p(i)* is the prior probability of class *i*, and $\lambda \in [0, 1]$ is a smoothing factor over an identity matrix.

Given the within-class covariance matrix, we make the NWCCW projection on prosodic supervectors using the following transformation:

$$P(X) = \mathbf{U}^T X \; with \; \mathbf{W}^{-1} = \mathbf{U}\mathbf{U}^T. \quad (2)$$

Since **W** is a full-rank matrix, there exists a Cholesky factorization **U** of $\mathbf{W}^{-1}$.

NWCCN optimally weights each of orthonormal directions to minimize a particular upper bound on error rate [9]. It thus can maximize the speaker-relevant information that is in the underlying feature space but is affected by channel and session variability. Therefore, applying NWCCN to the prosodic feature space will benefit the robustness and performance of speaker recognition system.

2.3. SVR Modeling

There are two reasons why we adopt the support vector regression (SVR) instead of SVM to model the prosodic supervectors. One is that the amount of training data available for the target speaker is very limited and there always exist many unvoiced frames in the utterance. In such case, we need a more general approach to avoid the over-fitting, like SVR, to train the speaker model. Another reason is that SVR aims to find a good approximation of the features.

SVR aims to minimize a regularized error function [10] given by:

$$C \sum_{n=1}^{N} E_\epsilon(y(X_n) - t_n) + \frac{1}{2} \parallel \mathbf{W} \parallel^2, \quad (3)$$

where $y(X_n)$ is the prediction of X_n and t_n is the corresponding target value in training data set. The quadratic error function in SVM has been replaced by an $\epsilon-$*insensitive error function* in SVR, which have a linear cost associated with the errors outside the insensitive region. The introduced slack variables allows points to lie outside the $\epsilon-$tube provided the slack variables are nonzero. It tolerates some degree of mismatch by the use of an margin controlled by the ϵ parameter. Therefore, SVR has a better generalization than SVM, and is thus suitable for a prosodic speaker recognition system.

3. SEGMENTAL WEIGHT FUSION

The goal of a fusion method in speaker recognition is to properly combine the scores from multiple subsystems for complementary effect. Given the fact that an individual subsystem on acoustic features generally has a much better performance than that on prosodic features, using a unique fusion weight may not achieve the fusion target which is to improve the speaker recognition performance with the complementary advantage of prosodic features.

To make better use of the discriminative cue of prosodic features, we conduct the score fusion separately in different score regions, using a segmental weight fusion (SWF) strategy. The fusion weights are estimated on the development data set, and applied on evaluation data accordingly. The two pre-processing steps before the score fusion are as follows:

- Score region partition. Assume A_{min} and A_{max} are the minimum and maximum acoustic scores in the development data set. Two Gaussian distributions, shown in

Fig.1 are constructed using the scores of target speakers and impostor speakers, respectively. All the score values are divided into three regions, Q_1, Q_2, and Q_3, by two thresholds, $Th1$ as the minimum score value of target speakers and $Th2$ as the maximum score value of impostor speakers. Each of the three score region represents a unique score characteristics, and should be handled differently in the score fusion. For example, Q_2 is the most confusable region in speaker recognition.

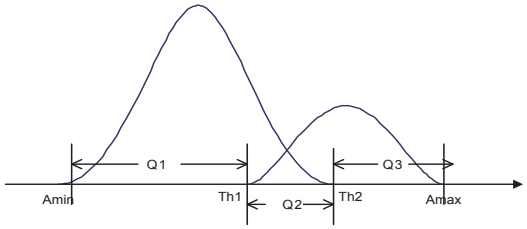

Fig. 1. Score segments in target/impostor score distributions

- Score scaling. As the acoustic scores and prosodic scores may have a different dynamic ranges, we conduct a linear transformation on prosodic scores for scaling as follows:

$$\begin{cases} \kappa P^r_{min} + \rho = A^r_{min} \\ \kappa P^r_{max} + \rho = A^r_{max} \end{cases} \Rightarrow \kappa\ , \rho \qquad (4)$$
$$\Rightarrow\ \bar{P} = \kappa P + \rho$$

The score transformation are made in each of the three regions independently by training the three sets of parameters κ, ρ. P^r_{min}, P^r_{max}, A^r_{min}, and A^r_{max} represent the region-dependent minimum and maximum values of prosodic and acoustic scores. $\bar{P}$ and P denotes the scaled and original prosodic scores.

As shown in Fig. 1, the decision threshold should be in the interval $[Th1, Th2]$ after the score fusion, in order to obtain the minimum detection cost. If the score of a test trial lies in the region Q_1, this trial will be classified as non-target. Therefore, we use the following formula to train the fusion weight by minimum $P_{Miss|Target}$ given threshold *Th1* .

$$\begin{cases} (1-\alpha)A_{Q_1} + \alpha\bar{P}_{Q_1} = F_{Q_1} \\ min\{P_{Miss|Target}\}\ ,\ threshold = Th1 \end{cases} \qquad (5)$$

In contrary to region Q_1, we apply the following formula to train the fusion weight by minimum $P_{FalseAlarm|NonTarget}$ given threshold *Th2* for the region Q_3.

$$\begin{cases} (1-\beta)A_{Q_3} + \beta\bar{P}_{Q_3} = F_{Q_3} \\ min\{P_{FalseAlarm|NonTarget}\}\ , \\ threshold = Th2 \end{cases} \qquad (6)$$

The region Q_2 is the most confusable region for the speaker recognition decision. The fusion weight is trained by minimizing the equal error rate (EER):

$$\begin{cases} (1-\gamma)A_{Q_2} + \gamma\bar{P}_{Q_2} = F_{Q_2} \\ min\{EER\} \end{cases} \qquad (7)$$

In the above three formulas, $\alpha, \beta, \gamma \in [0, 1]$, A_{Q_*}, $\bar{P}_{Q_*}$, and F_{Q_*} represent the acoustic score, transformed prosodic score and the final fused score.

4. EXPERIMENTS

In the following experiments, we will show the advantages with the proposed SVR-NWCCN approach for modeling prosodic features in speaker recognition, and compare the fusion strategies using the common equal weight, LLR and the proposed SWF methods for combining the acoustic and prosodic scores. All the results in this section are without score normalization, such as Tnorm and Znorm.

4.1. Experiment Setup

The speaker recognition experiments are conducted on the core test condition of the 2006 NIST Speaker Recognition Evaluation (SRE) corpus with 51448 test trails. Each trial contains about 2.5 minutes of speech. The speech data from the core test condition of the 2004 NIST SRE corpus are used as the development set, for tuning the smooth factor of NWCCN and the parameters of score fusion. The speech data from 2004/2005 NIST SRE corpus and Switchboard are used as negative samples and as the training set for Nuisance Attribute Projection (NAP) [3] on the supervectors of acoustic features.

EER and minDCF (minimum decision cost function) are adopted to measure the performance. Both the prosodic GMM and acoustic GMM are gender-dependent, with 64 Gaussian mixture components on the 6-dimension prosodic features and 512 Gaussian mixture components on the 39-dimension MFCC features. The smooth factor of the proposed SVR-NWCCN system is estimated using cross validation is set to 0.3.

4.2. Experiment Results

Table 1 shows the experimental results of four speaker recognition systems based on prosodic features. They are basic GMM-UBM system, GMM-SVM with SVM modeling on the prosodic supervectors, GMM-SVR with SVR modeling on the prosodic supervectors, and SVR-NWCCN with SVR modeling on the NWCCN projection. It is shown that both SVR modeling and NWCCN projection can improve the speaker recognition performance on prosodic features. The SVR-NWCCN obtains an EER of 17.72% which is a competitive result compared with recently reported prosodic feature systems on the same task [13].

Table 1. Performances comparison of four prosodic systems

Prosodic system	EER	minDCF
GMM-UBM	30.61%	9.98%
GMM-SVM	20.40%	7.52%
GMM-SVR	19.80%	7.47%
SVR-NWCCN	17.72%	7.13%

Table 2 shows the speaker recognition results of two separate systems with acoustic and prosodic features, as well as two fusion systems with different fusion strategies. The acoustic system is GMM-SVM based on MFCC features with NAP for channel and session compensation. The prosodic system is the SVR-NWCCN mentioned in Table 1. The two fusion strategies are fusion two systems with equal weight, and the proposed segmental weight fusion (SWF). The results show that the prosodic features help to improve the speaker recognition performance of fusion system when the correct fusion strategy is used. A relative 12.28% improvement is achieved over the acoustic system based on SWF. On the contrary, using the LLR fusion method [1] with a unique weight, the prosodic features do not give us effective complementary contribution because the estimated weight for the prosodic features is near to zero. Meanwhile, fusion strategy with a equal weight, shown in the Table 2, does not work well either due to the big performance gap between the acoustic and prosodic system.

Table 2. Performance comparison of two fusion strategies

System	EER	minDCF
Acoustic GMM-SVM-NAP	5.13%	2.31%
Prosodic SVR-NWCCN	17.72%	7.13%
Fusion-Equal Weight	7.12%	3.25%
Fusion-SWF	4.50%	2.08%

In Fusion-SWF, the weights of three regions are trained on the 2004 NIST SRE corpus, and applied to 2006 NIST SRE evaluation task. It is interesting to find that the best fusion result is obtained at $\alpha = 0$, $\beta = 0$, $\gamma = 0.6$. In the regions Q_1 and Q_3, Acoustic GMM-SVM-NAP system gives much higher recognition accuracy than Prosodic SVR-NWCCN system. The prosodic features can not help in the fusion system. In region Q_2, prosodic features has shown a great complementary advantage to acoustic features.

5. CONCLUSION

In this paper, we propose a new strategy to effectively exploit prosodic features for speaker recognition. The negative within-class covariance normalization provides a robust representation for prosodic features and the support vector regression helps to achieve good generalization and approximation in speaker modelling. It is expected to obtain a better performance by involving more prosodic features besides pitch and energy. The proposed segmental weight fusion strategy has shown to be able to effectively combine the acoustic and prosodic informations in different score regions for a reliable fusion system.

[1] http://www.dsp.sun.ac.za/nbrummer/focal/index.htm

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, pp. 19–41, 2000.

[2] W. M. Campbell, J. P. Campbell, D. A. Reynolds, E. Singer and P. A. Torres-Carrasquillo, "Support vector machines for speaker and language recognition," *Computer Speech and Language*, vol. 20, pp. 210-229, 2006.

[3] W. M. Campbell, D. Sturim, and D. A. Reynolds, "SVM based speaker verification using a GMM supervector kernel and NAP variability compensation," *Proc. ICASSP*, 2006.

[4] A. Stolcke, L. Ferrer, S. Kajarekar, E. Shriberg and A. Venkataraman, "MLLR transforms as features in speaker recognition," *Proc. ICASSP*, 2005.

[5] P. Kenny, G. Boulianne, P. Ouellet and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Trans. Audio Speech and Language Processing*, vol. 15, no. 4, pp. 1435-1447, 2007.

[6] S. Kajarekar, L. Ferrer, A. Venkataraman, and .et.al, "Speaker recognition using prosodic and lexical features," *Proc. IEEE ASRU*, pp. 19–24, 2003.

[7] A.G. Adami, R. Mihaescu, D.A. Reynolds, and J.J. Godfrey, "Modeling prosodic dynamics for speaker recognition," *Proc. ICASSP*, pp. 788–791, 2003.

[8] K. Sonmez, J. Zheng, E. Shriberg, S. Kajarekar, L. Ferrer and A. Stolcke, "Modeling nerfs for speaker recognition," *Proc. Odyssey*, pp. 51–56, 2004.

[9] A. Hatch and A. Stolcke, "Generalized linear kernels for one-versus-all classification: application to speaker recognition," *proc. of ICASSP*, 2006.

[10] C. M. Bishop, "Pattern recognition and machine learning," *Springer*, vol. 339-344, March 2006.

[11] P. Boersma and D. Weenink, "Praat: doing phonetics by computers," *http://www.praat.org/*.

[12] B. Xiang, U.V. Chaudhari, J. Navratil, G. N. Ramaswamy and R. A. Gopinath, "Short-time Gaussianization for robust speaker verification," *Proc. ICASSP*, 2002.

[13] E. Shriberg, "High-level features in speaker recognition," *Springer-Verlag Berlin Heidelberg*, pp. 241–257, 2007.