INTERFUSING THE CONFUSED REGION SCORE OF SPEAKER VERIFICATION SYSTEMS

Yanhua Long, Wu Guo, and Lirong Dai

iFly Speech Lab, Department of Electronic Engineering and Information Science,
University of Science and Technology of China, Hefei, Anhui 230027
{lyhsa,guowu}@mail.ustc.edu.cn, lrldai@ustc.edu.cn

ABSTRACT

In the text-independent speaker recognition field, there have been many excellent techniques based on the cepstral acoustic features. Recently, high level prosodic features have been widely used to verify the speaker's identity as they are less sensitive to the channel and noisy effect. But how to combine the existed prosodic system's scores with the scores of the system based on acoustic features to achieve a superior performance becomes a very difficult issue. This paper presents a combination method called interfusing the confused region scores (ICRS) to achieve a better recognition precision. We report results on the NIST 2006 speaker recognition evaluation (SRE) using two component systems: a standard MFCC-SVM and PGCP-SVM prosodic system, and show that the proposed interfusing technique results in 9.25% reduction in equal error rate (EER) and at last gets a EER = 4.9% after system's combination.

Index Terms— ICRS, Prosodic, Acoustic, Speaker verification

1. INTRODUCTION

In current speaker verification field, the best performance is obtained by fusing the scores of several subsystems. There has been many fusion techniques proposed: kernel combination [1], logistic regression [2, 3], Neural Network [4] and so on. In many cases, the combination leads to significant improvements if subsystems have comparably good results, as in [1, 2]. However, there are cases in which combining several theoretically complementary systems do not result in improvements over the single best system [5]. Most of these systems perform the fusion at score level: Modelling each type of features using a certain model independently, and then combine their scores in the last stage to produce the final score and the decision. Ideally, the ultimate goal of the fused system is to optimize the resulting performance after combining the existing subsystems.

In the last decades, the acoustic features had led to significant success. But the use of only acoustic features is

limiting because they suffer direct performance degradation in the presence of noise and environmental mismatch. Due to this, recent research directions have broadened to incorporate so called high level features in an effort to make speaker recognition systems more robust. However, dislike the cepstral acoustic features, the high level prosodic features, such as pitch and energy contours and speaking rates, are more difficult to extract from speech signals. The performances of existed prosodic systems are always worse than acoustic feature systems. If we combine the two type scores using usual techniques, it would not bring improvement in recognition precision though their scores are theoretically had complementarities.

In this work, we consider a fusion method called interfusing the confused region scores (ICRS) at the case of two available systems: acoustic and prosodic feature systems. Given two independent systems' scores on one certain verification task, we ignore the score region that have no complementary information between the two different features systems, and focus on the confused score region that too hard to decision a "yes" or "no" judgment. In this confused region, both the acoustic and prosodic systems have the worst recognition precision. We only combine the confused region scores using the usual combination techniques. Then incorporate the fused scores into two other parts of the acoustic scores to achieve the final scores for decision.

The outline of this paper is as follows: In section 2, we discuss how to interfuse the confused region scores between two existing systems to improve the final performance. In section 3, we describe the construct of our prosodic feature system. In section 4, the standard MFCC-SVM acoustic system will be introduced. Section 5 details the experimental databases, configurations, results and discussions. We conclude the paper in Section 6.

2. INTERFUSING THE CONFUSED REGION SCORES

The objective of score fusion method is to fuse multiple subsystems into a single effective one. In this work, we demonstrate the case of two existing subsystems to combine. As we know, the most important step in combination is to

define which score region has the strongest complementary property. Then the appropriate fusion methods are needed to be considered. We describe our ICRS method as follows:

Given two independent systems' trials scores on one certain verification task. There inevitably exists three score regions as described in Fig1: positive safe region (we can easy to decide these test trails should be indicated by "yes"), negative safe region (scores of those trials whose test speeches are said by imposters and we can easy to decide these test trails should be indicated by "no") and the confused scores region (the score region we can hard to give a right decision).

Figure 1: Score distribution of one verification system

Now, we obtain two subsystems' scores on evaluation task. How to combine them to get a final result, and decide which score region has the best complementary property without any prior knowledge become the most important step in combination. In order to complete this step, task-matched development corpus (NIST2004 core test) is used to train the appropriate transform parameters. The following six steps are used to address our ICRS method:

(1) **Calculate the decision cost:** We first set the acoustic system's scores as the base reference scores for combination. Then, we sort these base scores from low to high to get a last ranked score list Q , partition Q into M score blocks every empirically score intervals. Calculating the decision cost for each score block to form a decision cost table.

(2) **Define the confused score region:** Choosing those trials in continuous N score blocks whose decision costs are larger than other blocks. So we think the confused score region is composed by the corresponding N score ranges because of their large detect costs. As a matter of convenience, we denote this confused region of development task by A .

Both of the above two steps are used to train the confused score region. They are operated on the development corpus. But our objective is to better the performance of final fusion system after score combine. So the next steps are directly applied to evaluation subsystems.

(3) **Pick out the complementary trials from acoustic and prosodic scores of evaluation task:** On the evaluation task, we first find those test trials of acoustic features whose scores lie in the confused score region A defined in step (2) and denote these trials by B . Then we make B as an

indicator of the complementary test pairs list, picking out the corresponding trials in B from the prosodic features system and denote them by C .

(4) **Linear score normalization:** In order to make full use of independent system's complementarities, we should normalize those scores which are to be combined up to an identical score value range. The scores of part C are normalized by the following formulation:

$$\begin{aligned} ax_{MaxProsodic} + b &= x_{MaxAcoustic} \\ ax_{MinProsodic} + b &= x_{MinAcoustic} \end{aligned} \quad (1)$$

where a, b are the transformation parameters. $x_{MaxProsodic}$ and $x_{MaxAcoustic}$ respectively represent the maximum score in C and B . Similarly, $x_{MinProsodic}$ and $x_{MinAcoustic}$ respectively represent the minimum score in C and B . It can be easily calculate out the linear parameters a and b through (1). We give symbol D to present scores of C after normalized for simplification.

(5) **Combine the confused region scores on evaluation task:** After doing the score normalization, we use an equal weight to linear combine the confused region scores.

(6) **Integrate all scores for decision:** At last, we put three region scores together to get a final scores list for giving a threshold to make a decision: one is the combines scores by step (5), both of the other two are composed by the acoustic scores except the confused region scores.

The above six steps have described our ICRS methods. Experimental results will approve its effectiveness in Section 6.

3. PROSODIC SPEAKER VERIFICATION

Prosodic speaker verification exploits interpersonal variations in pitch and volume patterns to discriminate between speakers. The prosodic speaker verification subsystem implemented by the robust prosodic features, channel effect elimination method (within-class covariance normalization [7]) and GMM-SVM modeling techniques [8]. The reader should consult the reference papers for more details.

The 6 dimensional features of our prosodic speaker system can be roughly divided into two types: prosodic related real value features (log pitch and log energy) and the dynamics parameters (the first and second derivatives of pitch and energy). We extract log pitch values calculated at 25ms intervals with the Praat package [9]. Pitch is calculated with the autocorrelation method and is undefined in unvoiced region. Pitch is normalized on an utterance base by subtracting the maximum value of the utterance. Energy is extracted by HTK tool. Further processing including Rasta filtering, VAD detection, CMS compensation, and Gaussian-ization are applied to the energy.

For our prosodic modelling, we use the GMM-SVM technique for classification. The GMM mixture is set to 64

and GMM mean super vectors are classified by support-vector machines (SVMTorch [10]). WCCN are used to eliminate unwanted channel variability. The smooth factor $\alpha = 0.3$.

4. ACOUSTIC SPEAKER VERIFICATION

This system is also based on GMM-SVM. But we generate a 12-dimensional MFCC obtained from pre-emphasized speech every 10ms using a 25ms Hamming window. The C0 and the Delta-cepstral coefficients are appended to form a 39 dimensional feature vector. VAD, CMS and RASTA are sequentially applied to alleviate channel and noise effects.

The GMM UBM consists of 256 gender-dependent mixture components. We use MAP to adapt each training, testing and background segment. An adaptation relevance factor of $\tau = 8$ is adopted. Mean vectors of all GMM mixture components are then concatenated to form a GMM-mean super vector. Linear kernel [8] is used to train SVM models.

Dislike the prosodic system, in our acoustic features system, Nuisance attribute projection (NAP [6]) is used to remove channel variability.

5. EXPERIMENTS AND RESULTS

This section is dedicated to the experimental results and discussions. Results are provided for both acoustic and prosodic systems and for ICRS combination techniques.

5.1. Evaluation and development database

We performed our test experiments on the core condition of NIST 2006 SRE [11] corpus (all trials). The NIST 2006 database consists of : *i*) a subcorpus of the MIXER database and *ii*) a significant amount of additional multi-channel and multi-language data acquired in order to complete the corpus for the evaluation. Each train and test conversation has an average duration of 5 minutes, with 2.5 minutes of speech on average after silence removal. Although there are speakers of both genders in the corpus, no cross-gender trials are defined. We use equal error rate (EER) and the minimum decision cost value (minDCF) as metrics for performance evaluation.

NIST 2004 SRE core condition task is used to train ICRS score fusion method and for tuning system's parameters. NIST SRE 2005, Switchboard II and MIX5 corpuses are used to compute the projection matrix P in NAP for development 2004 core condition task. But we use NIST2004 data for the NIST2006 evaluation NAP matrix training. The rank of P matrix is empirically set to 64. The positive example is extracted from the target speaker and the negative examples are taken from the same data as NAP training for different task during SVM training process.

5.2. Experimental results and analysis

In this section, we use MFCC-SVM to represents our acoustic system and PGCP-SVM to represent our prosodic system for convenience.

Table 1 shows the decision cost table on NIST2004 development task. In this table, FR-Ratio = target trials in a fix score region / all target trials in the test task, FA-Ratio = notarget trials in a fix score region / all notarget trials in the test task. We only list the score region $[-1.31, -0.04]$ which is nearby the threshold of system in table 1. As scores which are around the decision threshold are always hard to separate, so the confused score region will probably lie in score region listed in table 1.

Table 1: Decision cost table on NIST2004 development task

Score Region	FR-Ratio	FA-Ratio
$[-1.31, -1.03]$	0.02%	5.20%
$[-1.03, -0.99]$	0.07%	4.41%
$[-0.99, -0.97]$	0.26%	3.41%
$[-0.97, -0.89]$	0.20%	3.00%
$[-0.89, -0.82]$	0.90%	2.08%
$[-0.82, -0.70]$	0.52%	4.43%
$[-0.70, -0.58]$	2.60%	2.12%
$[-0.58, -0.04]$	3.28%	1.06%

From the table 1, we see that, if the threshold is set to smaller than -0.99 , the notarget trials will be all easy to give "no" decision, and meanwhile, target trials' miss alarm is also very low. Similarly, if threshold is larger than -0.70 , the target trials will be all give the "yes" judgment and notarget trials' false alarm is also very low. So we can define region $[-0.99, -0.70]$ as our development confused score region. This region will be directly applied to the evaluation task for fusion.

Table 2: Performance of the confused score region on NIST2004 coretest task

	EER	minDCF
MFCC-SVM	24.1%	9.52%
PGCP-SVM	31.7%	9.91%
Combine	21.6%	8.87%
Improvement	10.37%	6.83%

Table 3: Performance of the confused score region on NIST2006 coretest task

	EER	minDCF
MFCC-SVM	19.03%	7.99%
PGCP-SVM	31.3%	9.78%
Combine	17.6%	7.32%
Improvement	7.35%	8.38%

Table 2 and 3 present results of the confused score region on NIST04 and NIST06 task. After using our ICRS method to complement two systems' fusion, there has a 10.37% relative improvement in EER and 6.83% in minDCF. When the same score region [-0.99, -0.70] is applied to NIST2006 evaluation for combination by ICRS, it results in performance gains of 7.35% in EER and 8.38% in minDCF.

Table 4: Performance of all trials of NIST2006 core test task

	EER	minDCF
MFCC-SVM	5.40%	2.50%
PGCP-SVM	18.3%	7.27%
Fusion-Linear	12.2%	4.85%
ICRS-SVM	4.90%	2.30%
Improvement	9.25%	8.00%

Table 4 gives the performance of all trials in evaluation task. We see that the acoustic features system MFCC-SVM is better than the prosodic features system PGCP-SVM both in EER and minDCF. When we combine them all the trials using the general linear equal weight, EER=12.2% and minDCF=4.85% are obtained. This combination is worse than any single system in performance. The fifth row of table 4 is our fusion results using ICRS technique. Comparing the results Fusion-Linear with ICRS-SVM, the result we have obtained using ICRS in confused score region is clearly better in all score distribution region.

6. CONCLUSION

In this paper, we presented a new method ICRS to combine the acoustic and high level system. This method made full use of the complementary information among different hierarchical features systems and greatly improved the performance of the fusion system. Comparing with those traditional fusion methods, ICRS discarded the prosodic scores region in which the acoustic features can perform very excellently, and combined the scores lied in the confused score region in which both of acoustic and prosodic features performed very bad. This led to a better result than those traditional fusion methods performed. In our experiment part, we only demonstrated the proposed ICRS technique on two existing systems. In fact, it can be easily applied to any multi-system statistical detection task.

7. REFERENCES

[1] Reda Dehak, Najim Dehak, Patrick Kenny, et. al. "Kernel Combination for SVM Speaker Verification," in *Odyssey*, 2008.

[2] P. Matejka, L. Burget, P. Schwarz, et. al. "STBU System for the NIST 2006 Speaker Recognition Evaluation," in *ICASSP, Hawaii, USA*, 2007.

[3] N. Brummer, L. Burget, J.H. Cernocky, et. al. "Fusion of Heterogeneous Speaker Recognition Systems in the STBU Submission for the NIST Speaker Recognition Evaluation 2006." *IEEE Trans. On Audio, Speech and Language Processing*, vol.15, no 7, pp2072-2084, 2007.

[4] W.M. Campbell, D.A. Reynolds, and J.P.Campbell. "Fusing Discriminative and Generative Methods for speaker Recognition: Experiments on Switchboard and NFI/TNO Field Data," in *Speaker Odyssey*, Toledo, Spain, June 2004, pp, 41-44.

[5] D. Reynolds, W. Campbell, T. Gleason, C. Quillen, D. Sturim, and P. Torres-Carrasquillo, "The 2004 MIT Lincoln Laboratory speaker recognition system," in *Proc. ICASSP*, Philadelphia, Mar.2005.

[6] A. Solomonoff, C. Quillen, and W. Campbell, "Channel Compensation For SVM Speaker Recognition," in *Proc. of Odyssey: The Speaker and Language Recognition Workshop*, Toledo, Spain, 2004.

[7] Andrew O. Hatch, Sachin Kajarekar, and Andreas Stolcke, "Within-Class Covariance Normalization for SVM-based Speaker Recognition," in *ICSLP* 2006.

[8] W. M. Campbell, D. E. Sturim, and D. A. Reynolds, "Support vector machines using GMM supervectors for speaker verification," *IEEE Signal Processing Letters*, 2005.

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

[10] R. Collobert and S. Bengio, "SVM-Torch: Support vector machines for large-scale regression problems," *J. Mach. Learn. Res.*, vol. 1, pp. 143-160, 2001.

[11] Nation Institute of Standards and Technology, *NIST speech group website*, 2006, <http://www.nist.gov/speech>.