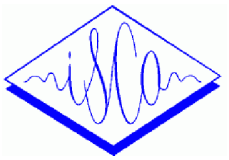

INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

Unsupervised NAP Training Data Design for Speaker Recognition

Hanwu Sun and Bin Ma

Institute for Infocomm Research (I^2R), A*STAR, Singapore 138632

{hwsun, mabin}@i2r.a-star.edu.sg

Abstract

The Nuisance Attribute Projection (NAP) with labeled data provides an effective approach for improving the speaker recognition performance in the state-of-art speaker recognition system by removing unwanted speaker channel and handsets variation. However, the requirement for the labeled NAP training data may limit its practical application. In this paper, we propose an unsupervised clustering strategy to design NAP training data without a priori knowledge about channel and speaker information. A fast clustering and purifying algorithm is introduced to group the unlabeled NAP training data into speaker dependent clusters to drive the NAP training data. The GMM-SVM based speaker recognition system is adopted to evaluate the performance. The system with the unsupervised NAP training data design achieves a similar performance with that using labeled NAP training data on both SRE06 1conv-1conv all English trials and SRE08 short2-short3 Tel-Tel All English trials subtasks.

Index Terms: speaker recognition, speaker diarization, speaker cluster, Nuisance Attribute Projection

1. Introduction

A critical problem for speaker recognition is to effectively deal with the mismatch between training and testing conditions for the same speaker. The mismatch is caused by a number of factors, such as microphone differences, ambient noise, and communication channel differences and different recording sessions [1, 2, 3].

Currently, the Nuisance Attribute Projection (NAP) [1, 2, 3] technique has been widely adopted to compensate the mismatch by removing the nuisance attributes. The NAP approach is generally conducted through a data-driven approach over a large size of labeled background training data. The NAP technique works well with a labeled data set which has been labeled with the nuisance attributes and/or the desired speaker/channel attributes such as in the NIST speaker recognition evaluations (SREs) [4, 5].

However, in many practical applications, it may be hard or time consuming to collect such labeled information of the training data for NAP design. For the old collected or recorded data, such information may be unavailable or lost forever. So such labeled data requirements may hamper the practical application of the NAP for speaker recognition. In this paper, we seek to introduce a discriminative clustering method for the NAP design without a priori knowledge about channel and speaker information, and investigate its performance in comparison with the ideal labeled NAP data sets.

We take advantage of the previous study on speaker diarization [6, 7, 8, 9] for the unsupervised NAP training data design. We propose a fast purification process to cluster the NAP training datasets into individual separate speaker dependent clusters. Such speaker dependent clusters are then used to drive NAP matrix to compensate the channel variations.

In the paper, we are interested in speaker recognition of two of the subtasks, NIST06 1conv-1conv all English trials [4] and NIST08 short2-short3 Condition 7 Tel-Tel all English trials [5]. Both training and testing data consist of one conversational excerpt of approximately 5 minutes of speech each excerpt, only involving the about 2.5 minutes target speaker speech on the designated side. Four speaker recognition experimental comparisons are presented in the paper, without any NAP compensation, NAP with a random data clustering, NAP with the proposed NAP data design and NAP with ground-truth labels, respectively.

The paper is organized as follows. In Section 2 we give an overview of the speaker recognition system. The proposed unsupervised NAP training data design is introduced in Section 3. The experimental results are reported in Section 4. Finally, we conclude in Section 5.

2. Speaker Recognition System

The speaker recognition system in this study was based on the GMM-SVM classifier [3] with MFCC features. A 16-dimenison MFCC features were generated for each speech frame with a window of 30ms and a frame shift of 12.5ms. By including the 16-dimension of the first derivatives and the 14-dimension of the second derivatives, a MFCC feature vector consists of 46 dimensional features.

We used the spectral subtraction process [10, 11] for noise reduction to assist the voice activity detection (VAD) process which was used to select the useful speech frames. The detailed description for the VAD can be found in [12]. The spectral subtracted speech signal was used for frame selection only while the MFCC features for speaker recognition and unsupervised NAP clustering were still derived from the original speech signals. The selected feature vectors were further processed by RASTA filtering [13] and cepstral mean and variance normalizations (MVN).

Gender-dependent universal background models (UBMs) with 1024 Gaussian mixture components were trained for the GMM-SVM speaker recognition system, and the speaker GMM models were trained by adapting the corresponding UBM using the MAP adaptation algorithm [14]. The relevance factor of MAP adaptation was set to 2.4. For each utterance, its mean vectors of the mixture components in the GMM were concatenated for a GMM supervector:

$$m(s) = [m_1(s),\ m_2(s)\ ,\cdots m_n(s)] \qquad (1)$$

where $m_i(s)$ are the mean vectors of individual Gaussian component. The mean vectors are further normalized by its standard deviation and weighted by the squared root of the weights of the Gaussian mixtures as:

$$m'((i-1)G+s) = \sqrt{w_i}\,\frac{m_i(s)}{\sigma_i(s)} \tag{2}$$

where G is the dimension of the feature vector, $m_i(s)$ is the s-th coefficient of m_i, $\sigma_i(s)$ is the standard deviation equivalent to the square root of s-th diagonal element of $\sum_i$ matrix. The w_i is the weight of the i-th Gaussian component.

The SVMTorch [15] was used to train SVM model. The 2004 NIST SRE (SRE04) corpus was used as the background data set for the UBM training and the background speaker data set for the SVM training. Meanwhile, we also used the SRE04 dataset to evaluate our proposed NAP clustering method. Based on the derived NAP matrix, the NAP projection compensated supervector is:

$$\hat{m} = \left(I - EE^t\right)\cdot m' \tag{3}$$

Here the E is the eigenvectors of the NAP matrix and the rank of NAP was set to be 60 in the experiments.

For the speaker recognition, a variety of score normalization approaches [16] have been proposed for a robust decision. We compared the Tnorm, Znorm, TZnorm and ZTnorm, and found that the TZnorm score normalization gave an overall better performance than others in this GMM-SVM speaker system. As a result, we reported the experimental results based on the TZnorm scores normalization. In the experiments, the 2005 NIST SRE (SRE05) 1-side training data were used for training the cohort models in Tnorm and the SRE04 data were used as imposter speech utterances in the Znorm.

3. Unsupervised NAP Training Data Design

The unsupervised NAP training data design is to cluster the unlabeled utterances into speaker related groups and use these grouped clusters for the NAP training so that various channel and handsets variation can be compensated without a priori knowledge about channel and speaker information. Like what we have done in the NIST RT speaker diarization [6, 7, 8, 9], the speaker clustering method contains three stages: Firstly, the N utterances related supervectors were random divided into small groups. Then, we purified these groups using the speaker supervectors based on their highest mean scores against the clustered supervisors. The final stage was to discard those clusters with small number supervectors and relocate them to their highest mean score related clusters. In order to compute these testing scores among supervectors efficiently, we adopted the supervector's dot product value as the test score and pre-computed the test scores among these supervectors and kept them in a matrix.

We summarize the clustering method as follows:

1) Train a gender-dependent Root GMM, λ_{Root} (same as the one used in the GMM-SVM system).
2) Compute all the NAP related mean supervectors $V=[m_1, m_2, ..., m_N]$ via MAP [14].
3) Random divide the N supervectors into Q initial clusters (N>Q> expected speakers).
4) Compute the N supervectors dot product matrix A (for fast tabular score search),

$$A = V \bullet V^T \tag{4}$$

5) For each supervector, compute the averaged dot product scores against the Q clusters using pre-computed matrix score A (not include itself dot product score).

$$S(i,j) = \sum_{k=1}^{Q(j)} A[i,k]/Q(j) \quad i = 1, ..., N,\ j = 1, ..., Q \tag{5}$$

where $Q(j)$ is the number of supervectors in the j^{th} cluster.

6) Relocate the supervectors into the Q clusters by using their highest averaged score.
7) Repeat the steps 5) and 6) until no supervector change is found.
8) Discard the cluster which contains small number of supervisors (<=3 in experiment) and relocate them to the clusters with the highest averaged score with each of the supervisors.
9) Repeat step 5) until no cluster contains more than the given small number of supervisors and no supervector change is found.

Since we used the same supervectors in the NAP clustering as the SMM-SVM speaker recognition experiments, no extra computation is required. Meanwhile, we only need to do the dot product scores once among supervectors, so the computational requirement is very low.

4. Speaker Recognition Experiments

The experiment in this work were focused on two subtasks in the 2006 NIST SRE (SRE06) 1conv-1conv all English trials [4] and the 2008 NIST SRE (SRE08) short2-short3 Condition 7 Tel-Tel all English trials [5]. In the following, we first introduce the measure of the unsupervised NAP clustering performance in comparison to the labeled NAP training data. Then, we apply the unsupervised NAP training design to the speaker system and compare the results with different NAP approaches to evaluate its performance.

4.1. Speaker Recognition Evaluation Measure

We evaluate the speaker recognition performance by both the Equal Error Rate (EER) and the Detection Cost Function (DCF) [4, 5]. The DCF is a weighted sum of miss detection and false alarm rates defined in the NIST SRE evaluation plans [4, 5], and are given as follows:

$$\begin{aligned} DCF &= C_{Miss} \times P_{Miss|\text{Target}} \times P_{\text{Target}} \\ &+ C_{FalseAlarm} \times P_{FalseAlarm|Non\text{Target}} \times (1 - P_{\text{Target}}) \end{aligned} \tag{6}$$

where $C_{\text{Miss}} = 10$, $P_{\text{Target}} = 1$ and $C_{\text{FalseAlarm}} = 0.01$.

Moreover, we also demonstrate the DET curves [4] to show the system performance between the tradeoff false rejections and false acceptances rates.

4.2. Unsupervised NAP Results

To evaluate the unsupervised clustering of the NAP training data, we used the SRE04 training data set as the evaluation data set. We set each initial cluster to contain four random supervectors and applied the purification and clustering. Since we know the NAP labels of all the speakers in SRE04, we are able to evaluate how good the proposed clustering method is in comparison to the ground-truth labeling. We use a simple speaker cluster rate and the speaker purification rate to do the measurement. The speaker clustering rate and speaker purification rate are given as below:

$$S_{\text{Cluster\_rate}} = \frac{\text{No. of Speaker Clusters}}{\text{No. of Speakers}} \tag{7}$$

and

$$S_{Purification\_rate} = \sum_{i=1,\ldots M} P(i)/N \tag{8}$$

where M is final clustering number, N is total number of supervectors in the NAP training data set and $P(i)$ is the biggest same speaker number within the ith cluster. Table 1 shows the results for the male, female and all speakers based on NIST04 NAP ground-truth, respectively.

Table 1. Speaker Cluster Rates and Purification Rates.

	Male	*Female*	*All*
Clustering rates	93.8%	93.2%	93.4%
Purification rates	90.7%	89.5%	90.1%

We achieved about overall 93% speaker cluster rate and 90% speaker purification rate for the unsupervised speaker clustering.

4.3. Speaker Recognition Results

To appreciate how we benefit from the clustering NAP design, we have conducted four speaker recognition experiments under four different NAP compensation schemes on SRE06 1conv-1conv all English trials [4] and SRE08 short2-short3 Condition 7 Tel-Tel all English trials:

1) without any NAP compensation (No NAP);

2) with a random clustered NAP (Random NAP) training data set;

3) with the proposed unsupervised NAP training data design (Clustering NAP);

4) with ground-truth NAP labels (Labeled NAP) .

The experimental results for NIST SRE06 1conv-1conv all English trails and SRE08 short2-short3 condition 7 Tel-Tel all English trials tasks are shown in Table 2 and Table 3, respectively. The related DET plots are also illustrated in Figure 1 and Figure 2, where minimum DCF is marked with red cycle.

From Table 2 and Figure 1, it is not surprised that the worst results is the one without any NAP compensation condition. Meanwhile, we can also see that the random grouped NAP training data set can also improve the system performance for more than 20% for both EER and DCF compared with the performance without NAP. It is more important to note that the proposed unsupervised NAP training data design significantly improves the speaker recognition performance in terms of both EER and minimum DCF compared to both No NAP and Random NAP conditions. With the help of the unsupervised NAP training data design approach, the Clustering NAP achieved an EER of 2.89% and a minimum DCF of 1.62%, representing a 41.7% relative improvement in EER, and 32.2% relative improvement in minimum DCF over no NAP condition. Of course, the best result was still the labeled NAP data condition. However we can see that the difference between the proposed unsupervised NAP training data design and labeled NAP training data is rather small in both EER and DCF. Figure 1 also illustrates that the DET curves and minimum DCF points between the unsupervised NAP training data design and the labeled NAP system are very close.

Table 2. EER and min DCF for the SRE06 1conv-1conv All English Trials under Different NAP Approaches.

NAP Conditions	**Male**		**Female**		**All**	
	EER %	**DCF x100**	**EER %**	**DCF x100**	**EER %**	**DCF X100**
No NAP	4.56	2.05	5.26	2.47	4.97	2.39
Random NAP	3.56	1.63	4.06	2.12	3.89	1.97
Clustering NAP	2.70	1.36	3.19	1.81	2.89	1.62
Labeled NAP	2.50	1.30	2.92	1.70	2.72	1.57

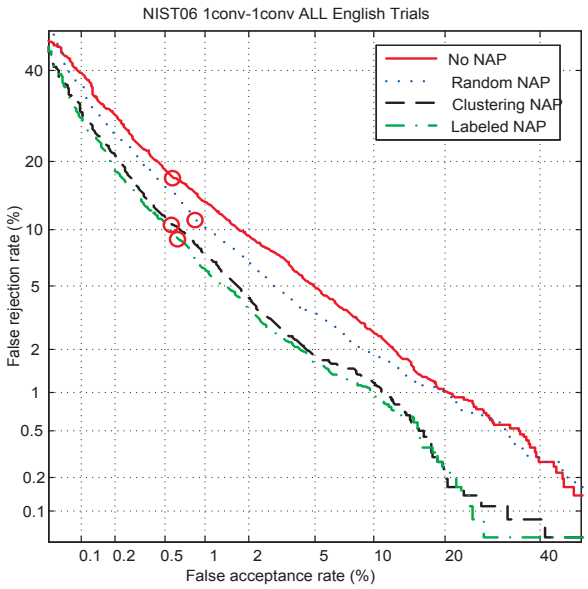

Figure 1. SRE06 1conv-1conv English Trials Subtask DET Curves under Different NAP Approaches.

We also report the results for SRE08 short2-short3 condition 7 Tel-Tel all English trials as shown in Table 3 and Figure 2. It is observed that the unsupervised NAP training data design achieved 36.8% EER reduction over No NAP condition and

25.5% EER reduction over the Random NAP approach. It is only about 1% relative EER worse in comparison to the Labeled NAP condition. The analysis indicates that the proposed unsupervised NAP training data design consistently works well on both SRE06 and SRE08 Tel-Tel subtasks.

Table 3. EER and min DCF for the SRE08 Condition 7 Tel-Tel All English Trials under Different NAP Approaches.

NAP Conditions	Male		Female		All	
	EER %	DCF x100	EER %	DCF x100	EER %	DCF x100
No NAP	4.04	1.59	4.55	1.87	4.21	1.79
Random NAP	3.37	1.27	3.54	1.53	3.57	1.51
Clustering NAP	2.39	0.93	2.87	1.38	2.66	1.23
Labeled NAP	2.51	0.91	2.69	1.27	2.63	1.18

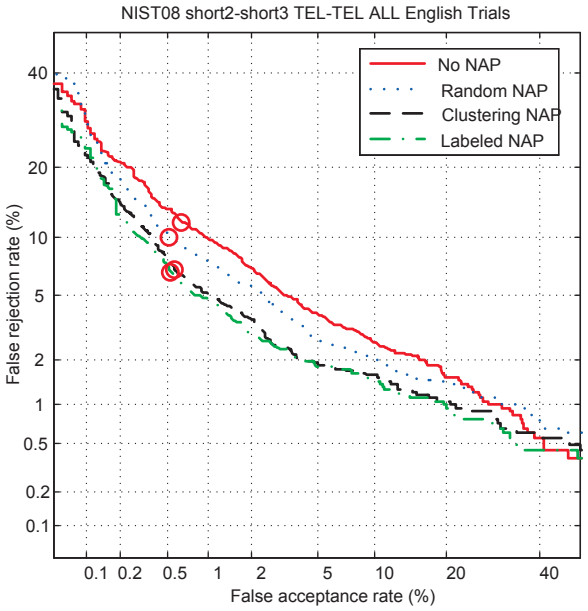

Figure 2. SRE08 Condition 7 Tel-Tel All English Trials DET Curves under Different NAP Approaches.

5. Conclusions

In this paper, we present an unsupervised clustering method to design NAP training data without knowledge about the labels of the data. A fast clustering and purifying algorithm was introduced to group the NAP training data into speaker dependent clusters. The unsupervised NAP training data design strategy achieved an EER of 2.89% and a minimum DCF of 1.62%, and reduced the speaker recognition EER and DCF by 41.7% and 32.2% for SRE06 1conv-1conv over No NAP condition. It shows that the results using the proposed strategy are very close to that by using the labeled NAP training data. The similar conclusion is also obtained for SRE08 Tel-Tel subtask. The current NAP clustering experiment was conducted on the SRE04 telephone based dataset, a future study will focus on using the algorithm to handle more complicated NAP training conditions, such as the NIST SRE microphone channel's dataset and NIST SRE interviewer's channel dataset.

6. References

[1] A. Solomonoff, C. Quillen and W.M. Campbell, "Channel Compensation for SVM Speaker Recognition", *In Proc. Odyssey*: The Speaker and Language Recognition Workshop in Toledo, Spain, ISCA, pp. 41–44, 2004.

[2] W.M. Campbell, A. Solomonoff and I Boardman, "Advances in Channel Compensation for SVM Speaker Recognition". in *Proc. ICASSP*, pp. 18-23 Philadelphia, 2005.

[3] W. M. Campbell, D. E. Sturim, D. A. Reynolds, and A. Solomonoff, "SVM based speaker verification using a GMM supervector kernel and NAP variability compensation," in *Proc. ICASSP*, pp. 97–100, 2006.

[4] NIST 2006 Speaker Recognition Evaluation Plan,

http://www.itl.nist.gov/iad/mig/tests/sre/2006/sre-06_evalplan-v9.pdf.

[5] NIST 2008 Speaker Recognition Evaluation Plan,

http://www.itl.nist.gov/iad/mig//tests/sre/2008/sre08_evalplan_release4.pdf.

[6] H. Sun, B. Ma, Z. Swe. and H. Li., "Speaker Diarization System for FT07 and RT09 Meeting Room Audio," in *Proc. ICASSP*, pp.4982–4985, 2010.

[7] T.L. Nwe, H. Sun, B. Ma, and H. Li," Speaker Clustering and Cluster Purification Methods for RT07 and RT09 Evaluation Meeting Data ", *IEEE Transactions on Speech, Language Processing*, vol. 20, no. 2, pp. 461–473 2012.

[8] "Spring 2007 (RT-07) Rich Transcription meeting recognition evaluation plan," http://www.nist.gov/speech/tests/rt/rt2007/docs/rt07-meeting-eval-plan-v2.pdf.

[9] "2009 (RT-09) Rich Transcription meeting recognition evaluation plan,"

http://www.itl.nist.gov/iad/mig//tests/rt/2009/docs/rt09-meeting-eval-plan-v2.pdf.

[10] S.F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE. Trans. Acoustics, Speech, Signal Processing*, vol. 27, pp. 113–120, 1979.

[11] R. Martin "Spectral Subtraction Based on Minimum Statistics," in *Proc. EUSPICO*, vol. 2, pp.1182–1185, 1994.

[12] H. Sun, B. Ma and H. Li, "An Efficient Feature Selection Method for Speaker Recognition," in *Proc. ISCSLP*, pp. 181–184, 2008.

[13] H. Hermansky and N. Morgan, "RASTA Processing of Speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 578–589, 1994.

[14] D.A. Reynolds, T.F. Quatieri and R.B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," *Digital Signal Processing*, 10(1):19-41, 2000.

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

[16] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score Normalization for Text-independent Speaker Verification Systems," *Digital Signal Processing*, vol. 10, no 1-3, pp. 42–54, Jan 2000.