Improved Unsupervised NAP Training Dataset 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 Project (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 channel and handset variation. However, the requirement for the labeled NAP training data may limit its practical application. In our previous study, a simple unsupervised clustering algorithm based on dot products between supervectors was introduced for designing NAP training dataset without a prior knowledge about channel and speaker information. Using such clustering results as the initial training dataset, in this paper, we make a further improvement of the training dataset by enhancing similarity measurement of supervectors via NAP projection and score normalization. The effectiveness of this unsupervised NAP training dataset design strategy has been verified in the experiments using the in-house development dataset of IIR submission for the 2012 NIST SRE.

Index Terms: *speaker recognition, speaker clustering, Nuisance Attribute Projection*

1. Introduction

One of critical problems 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 microphones, ambient noises, and communication channels and different recording sessions [1, 2, 3].

The Nuisance Attribute Projection (NAP) [3] technique has been widely adopted to compensate the mismatches by removing the nuisance attributes in the high dimensional space. The NAP approach is generally conducted through a data-driven approach over a large size of labeled background training data. It works well with a labeled data set such as in the NIST speaker recognition evaluations (SREs) corpora [4, 5, 6, 7], which have been labeled with the nuisance attributes of the speaker/channel information. However, in many practical applications, it may be hard or time consuming to collect such labeled information of the training data for NAP design. Such a requirement of labeled data may hamper the practical deployment of NAP in speaker recognition applications.

In the previous study [8], we introduced a fast clustering method to group the speech utterances into speaker related clusters based on the simple dot product scores between GMM supervectors, and then applied the unsupervised NAP training by using the clustering results. The experiments conducted by using the 2004 NIST SRE (SRE04) as the unsupervised NAP training data has demonstrated an effective compensation to overcome channel effects in the speaker recognition.

However, we have noticed that the SRE04 corpus [4] mainly consists of telephone channel speech data and the amount of data is rather small, and thus we need a more diversified speech corpus for verifying the effectiveness of unsupervised NAP training. Since 2006 to 2012, NIST has conducted more complicated speech data collection efforts [5, 6, 7] by adding the speech data from microphone channel as well as interview channel. In this paper, we would like to study the unsupervised NAP training with the mixed channel speech corpus and develop more discriminative clustering methods for the NAP training dataset design. We will introduce an improved discriminative clustering method for the NAP training data design using mixed channel speech data, without a priori knowledge about channel and speaker information, and investigate the performance of the proposed method by the comparison with the NAP training using ground-truth channel and speaker labels.

We take advantage of the previous study on speaker diarization [10, 11, 12] for the cluster purification in the unsupervised NAP training data design. We first adopt the previous proposed fast clustering method based on the dot product scores between GMM supervectors [8] as the initial NAP training dataset for a NAP training and then apply speaker verification with the trained NAP, as well as score normalization to generate the score for each of utterances instead of the simple dot product scores. Finally the same clustering approach as in [8] is applied to cluster the utterances into speaker related clusters. Such speaker dependent clusters are then used to derive NAP matrix to compensate the channel variations. We adopted our in-house development set for SRE12 [9] which were extracted from SRE06, SRE08 and SRE10 corpora for this NAP study, which contain these three type channels, telephone, microphone and interview.

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

2. KL-SVM-NAP speaker recognition system

The speaker recognition system using in this study is based on the support vector machine (SVM) using the *Kullback-Leibler* (KL) divergence kernel and the *nuisance attribute projection* (NAP) technique for channel compensation, in short, the KL-SVM-NAP as reported in [1, 2].

We use the MFCC feature in this study. In particular, a 16-dimenison MFCC features are generated for each speech frame with a window of 30ms and a frame shift of 12.5ms. By including the 16-dimension first and second derivatives, the resulting MFCC feature vector consists of 48 elements.

The spectral subtraction technique [13, 14] is used to assist the voice activity detection (VAD) for selecting useful speech frames [15]. The MFCC feature vectors are then processed by RASTA filtering [16] and followed by *mean and variance* normalization (MVN).

In the KL-SVM-NAP speaker recognition, each of the utterances in variety of durations will be represented by a

high-dimensional vectors referred to as the GMM supervector. Channel compensation and speaker detection are then performed in the high-dimensional vector space.

Let $\Lambda = \{ \omega_i, \mathbf{\mu}_i, \mathbf{\Sigma}_i; i = 1, 2, \ldots M \}$ be the parameters of the universal background model (UBM), where M is the number of mixture components, ω_i are the mixture weights, $\mathbf{m}_i$ are the mean vectors, and $\mathbf{\Sigma}_i$ are the covariance matrices assumed to be diagonal. For a given utterance X_s, the Baum-Welch statistics are used to adapt the mean vectors of the UBM using the maximum *a posterior* (MAP) [17]. The adapted mean vectors are concatenated to form a GMM supervector, as follows:

$$\mathbf{m}(s) \equiv \left[\mathbf{m}_1^{\mathrm{T}}(s), \mathbf{m}_2^{\mathrm{T}}(s), \ldots, \mathbf{m}_M^{\mathrm{T}}(s) \right]^{\mathrm{T}}, \quad (1)$$

where T denotes transposition. The mean vectors are then normalized by its standard deviation and weighted by the squared root of the mixture weights:

$$\mathbf{m}'_i(s) = \sqrt{\omega_i}\, \mathbf{\Sigma}_i^{-1/2} \mathbf{m}_i(s), \; i = 1, 2, \ldots, M . \quad (2)$$

The normalization in (2) allows the similarity between two GMM supervectors to be computed by taking their inner product in accordance to the KL-divergence [3].

Based on the NAP matrix derived from the labeled or clustered NAP training dataset, the NAP projection compensated GMM supervector is given as [3]:

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

Here the E is the eigenvectors of the NAP matrix.

With KL-SVM-NAP, one SVM is trained for each target speaker using the NAP related supervector. Let Ω_k be the target speaker utterance (i.e., the positive example), and $\Re$ the set of supervectors pertaining to background speakers. An SVM solver for the dual formulation [3],

$$f_k \leftarrow \mathrm{SVM}\left(\Omega_k \,\|\, \Re \right), \quad (4)$$

with the Lagrange multipliers α associated to all the supervectors in the training set and a bias parameter β, which essentially forms a linear model f_k for a target speaker, as follows:

$$f_k(\mathbf{m}') = \left[\sum_{\Omega_k} \alpha_i \mathbf{m}'(i) - \sum_{\Re} \alpha_j \mathbf{m}'(j) \right]^{\mathrm{T}} \mathbf{m}' + \beta . \quad (5)$$

It should be noted that, the background set $\Re$ is used for all target speakers enrolled to the system. In this study, the SRE04 corpus was used as the background speaker data set for the SVM training and also used to train the gender-dependent UBMs with 1024 mixture components. The rank of NAP is set to be 60 in the experiments. TZnorm was used for score normalization [18], where SRE05 data was selected for training the cohort models for Tnorm and SRE04 data was used as imposture utterances for Znorm.

3. Improved unsupervised NAP training data design

In the previous study [8], we proposed an unsupervised NAP training dataset design to cluster the unlabeled utterances into speaker related groups and use these grouped clusters for NAP training so that channel and handset variations can be compensated without a priori knowledge about channel and speaker information. Such clustering method contains three stages: firstly, all the NAP training utterances are random divided into small groups. Then, we purified these groups based on the simple supervector dot product scores. The final stage is to discard those clusters with small number supervectors and relocate them to the nearest clusters. With a diversity of channels involved in the training data, the above-mentioned simple NAP training dataset design method might not be sufficient. Instead of the simple dot product scores, we further improve the unsupervised NAP training dataset design using the speaker verification score derived from the KL-SVM-NAP system with the TZnorm score normalization. The improved NAP training data clustering has the following two stages. The first stage is to apply the previous proposed clustering method to get the initial NAP clusters and the second stage is to enhance the clustering performance by using the KL-SVM-NAP speaker verification and score normalization. The details of these two stages are as follows:

Stage 1: Simple dot product score based NAP training data clustering [8].

1) Train a gender-dependent Root GMM, λ_{Root} (same as the one used in the KL-SVM system).
2) Compute GMM supervectors $V=[m_1, m_2, \ldots, m_N]$ via MAP [17] for all utterances.
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),

$$A = V \bullet V^{T} \quad (6)$$

5) For each supervector, compute the averaged dot product scores against the Q clusters using pre-computed matrix score A.

$$S(i,j) = \sum_{k=1}^{Q(j)} A[i,k]/Q(j) \quad i = 1, \ldots, N, \; j = 1, \ldots, Q \quad (7)$$

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

6) Relocate the supervectors into the nearest Q clusters according to the highest averaged scores against the clusters.
7) Repeat the steps 5) and 6) until no supervector change is found.
8) Discard the cluster which contains small number of supervisors (<=1 in experiment) and relocate them to the clusters according to the highest averaged scores against the clusters.
9) Repeat step 5) until no cluster contains more than the given small number of supervisors and no supervector change is found.

Stage 2: Speaker verification based NAP clustering.

Based on the above initial NAP matrix, the improved unsupervised NAP training dataset design is as follows:

1) Compensate the channel effects for all utterances of both background data and NAP training data by:

$$\hat{m} = \left(I - EE^t\right) \cdot m' \tag{8}$$

where E is the eigenvectors of the clustered NAP matrix.

2) Train all the NAP's utterance speaker models via KL-SVM-NAP, each utterance will be trained as individual speaker model.

$$f_k \leftarrow \mathrm{SVM}\left(\Omega_k \,\|\, \Re\right) \tag{9}$$

where SRE04 data is used as the background dataset $\Re$

3) Update the NAP score matrix A using the KL-SVM-NAP models, and TZnorm is applied to normalize the SVM scores. The SRE05 data is used for training the cohort models for Tnorm while SRE04 data is used as imposture utterances for Znorm.
4) Repeat what we have done in Stage 1 (from Step 3 and Step 9 in Stage 1), and group the utterances into speaker related groups according to the score matrix A.
5) Using the updated NAP matrix to repeat Stage 2 from Steps 1 to 5 for further NAP cluster purification.
6) Finally apply the clustering NAP training dataset for NAP training.

4. Experimental results

The experiment in this study was focused on our in-house development set of IIR submission for the 2012 NIST SRE [7, 9] using speech segments drawn from SRE06, SRE08 and SRE10 corpora. The NAP training data contains the speech data in telephone channel, microphone channel and interview channel. The SRE12 development set consists of training and testing sets without overlapped utterances. The training set was also used as for NAP training. In the following experiment, we first measure the performance of unsupervised NAP training dataset design compared with the labeled NAP training dataset. Then, we apply the unsupervised NAP training design to the speaker system and compare the results with different NAP approaches.

We have conducted the speaker clustering and purification experiments on the NAP data set, as well as speaker recognition experiments on the developed training and developed testing data under the following five different NAP compensation schemes:

1) without any NAP compensation (No NAP);
2) with simple supervector dot product scores based NAP (Supervector NAP) [8] ;
3) with first round KL-SVM-NAP assistant scores based NAP (1st SVM NAP);
4) with fourth round adaptive KL-SVM-NAP assistant scores based NAP (4th SVM NAP);
5) with ground-truth NAP labels (Labeled NAP) .

4.1. Speaker recognition performance evaluation

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], given as follows:

$$\begin{aligned} DCF &= C_{Miss} \times P_{Miss|Target} \times P_{Target} \\ &+ C_{FalseAlarm} \times P_{FalseAlarm|NonTarget} \times (1 - P_{\mathrm{Target}}) \end{aligned} \tag{10}$$

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

4.2. Unsupervised NAP results

To evaluate the unsupervised clustering of the NAP training data, we used the in-house SRE12 development set [9] drawn from SRE06, SRE08 and SRE10. We set each initial cluster to contain three random supervectors and applied the purification and clustering. Since the NAP labels of all the speakers are known, 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 [8]:

$$\mathrm{S}_{\mathrm{Cluster\_rate}} = \frac{\text{No. of Speaker Clusters}}{\text{No. of Speakers}} \tag{11}$$

and

$$S_{Purification\_rate} = \sum_{i=1,\ldots,M} P(i) / N \tag{12}$$

where M is final number of clusters number, N is total number of supervectors in the NAP training dataset and $P(i)$ is the largest number of utterances from the same speaker within the i^{th} cluster. Table 1 shows the results for the male and female clustering rate and purifying rate, respectively.

Table 1. Speaker Cluster Rates and Purification Rates.

	Male		*Female*	
	Clustering Rates	*Purifying Rates*	*Clustering Rates*	*Purifying Rates*
Supervec. NAP	33.65%	57.51%	29.00%	41.50%
1st SVM NAP	92.21%	88.57%	91.25%	84.83%
4th SVM NAP	93.35%	89.96%	92.76%	86.11%

From the results shown in the table 1, we can see that the previous proposed dot product score based method basically fails under such a mixed-channel training data set. After applying KL-SVM-NAP speaker verification for the clustering scores, both speaker clustering rate and purification rates are significantly improved. The results are consistent with that reported in the previous approach [8] on the SRE04 corpus. In addition, further improvements can be achieved by using multiple round of unsupervised adaptive training of NAP score matrix.

4.3. Speaker recognition results

The experimental results for our development training dataset and development testing dataset under the suggested five NAP schemes are shown in Table 2. We can consider the development training dataset as closed test while development testing dataset as open test. The EER and minimum DCF scores are also illustrated in Figures 1 and 2.

Similar as the previous study [8], from Table 2 and Figure 1, it is not surprised that the speaker verification without NAP gives the worst results. The previous proposed supervector dot product scores based clustering NAP method can improve the system performance for 30% ~ 50% in both EER and DCF compared with the performance without any NAP on both open and closed testing datasets.

However, we notice that the performance of such simple NAP training data clustering is still rather poor compared with the labeled NAP training dataset, due to the diversity of channel conditions.

Table 2: EER and min DCF for the open and closed tests under different NAP Approaches.

	Open Test		Closed Test	
NAP Conditions	Male EER%/DCF%	Female EER%/DCF%	Male EER%/DCF%	Female EER%/DCF%
No NAP	9.27/5.44	11.7/6.93	7.22/3.99	8.99/4.91
Supervector NAP	4.96/2.16	7.93/3.29	5.12/1.92	7.24/2.67
1st SVM NAP	2.26/1.06	3.87/1.66	2.22/0.78	3.16/1.20
4th SVM NAP	2.13/0.98	3.51/1.53	1.90/0.65	2.67/1.04
Labeled NAP	1.95/0.92	3.14/1.40	1.71/0.58	2.20/0.89

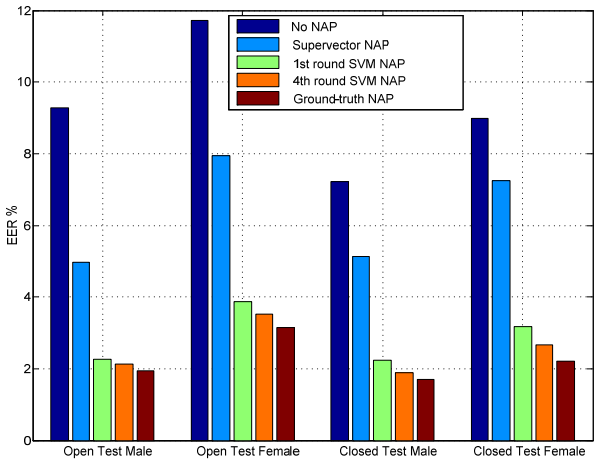

Figure 1. EER Rates Under different NAP Conditions on Open and Closed Tests.

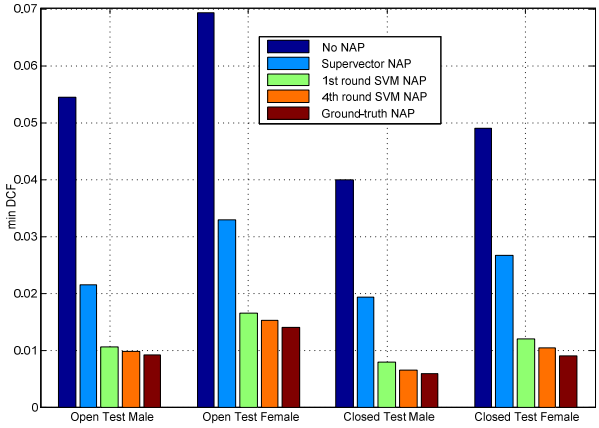

Figure 2. Minimum DCFs Under different NAP Conditions on Open and Closed Tests.

With the improved KL-SVM-NAP method for unsupervised NAP training dataset design, a significant improvement of speaker verification performance in terms of both EER and minimum DCF has been achieved. The 1st round SVM clustering NAP achieves an EER of 2.26% and a minimum DCF of 1.06% in the open male test, representing a 77.0% relative improvement in EER, and 80.5% relative improvement in minimum DCF over no NAP condition. It also has about 54.0% relative improvement in EER, and 50.9% relative improvement in minimum DCF over our previous proposed simple supervector score based NAP approach in the open male testing. The similar results have been observed for the female open testing set and the closed testing set.

Furthermore, with a recursive NAP training dataset design, an additional improvement can be achieved, as indicated with the fourth round adaptive training in Table 2. Although the performance of the unsupervised NAP training dataset design is still worse than that with labeled NAP training dataset, the performance gap has been significantly reduced.

5. Conclusions

In this paper, we present an improved KL-SVM-NAP based unsupervised clustering method to design NAP training data without knowledge about the labels of the mixed-channel training data. The experiment results show the effectiveness of such NAP clustering method while the NAP training data are in telephone channel, microphone channel and interview channel. The speaker purification and clustering results on the in-house SRE12 NAP training data sets show the significant advantage of the improved method. The speaker verification results on both open and closed test dataset demonstrate the great improvement for this KL-SVM-NAP based unsupervised NAP training data clustering approach in both EER and minimum DCF, which are quite close to that by using the NAP training data with labeled information.

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 2004 Speaker Recognition Evaluation Plan, http://www.itl.nist.gov/iad/mig/tests/sre/2004/SRE-04_evalplan-v1a.pdf

[5] NIST 2006 Speaker Recognition Evaluation Plan,

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

[6] NIST 2008 Speaker Recognition Evaluation Plan, http://www.itl.nist.gov/iad/mig//tests/sre/2008/sre08_evalplan_release4.pdf.

[7] NIST 2012 Speaker Recognition Evaluation Plan, http://nist.gov/itl/iad/mig/upload/NIST_SRE12_evalplan-v17-r1.pdf

[8] H. Sun and B. Ma, "Unsupervised NAP Training Data Design for Speaker Recognition", in *Proc. Interspeech 2012*, Portland, 2012.

[9] H. Sun, K. Lee and B. Ma, "Anti-Model KL-SVM-NAP System For NIST SRE 2012 Evaluation", accept for *ICASSP 2013,* Vancouver, Canada.

[10] 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.

[11] 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.

[12] "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.

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

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

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

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

[17] 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.

[18] 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.