AN INVESTIGATION OF SUMMED-CHANNEL SPEAKER RECOGNITION WITH MULTI-SESSION ENROLLMENT

Shanshan Zhang, Ce Zhang, Rong Zheng, Bo Xu

Interactive Digital Media Technology Research Center
Institute of Automation, Chinese Academy of Sciences, Beijing, China
{shanshan.zhang, ce.zhang, rong.zheng, xubo}@ia.ac.cn

ABSTRACT

This paper describes a general framework of speaker recognition on summed-channel condition for both enrolling and test data. We present several methods for clustering the target speaker who is involved in multiple summed-channel enrolling excerpts. In our approach, each excerpt is segmented separately by a speaker diarization system as the first stage. Then segments belonging to the same speaker are clustered to train the target speaker model, and speaker verification is applied finally. We propose several effective objective functions to measure the purity of clustered segments in multi-session enrollment. Different confidence measures for summed-channel scoring are also presented. We report experimental results on female part in the NIST 2008 speaker recognition evaluation data, which show that our approach applied on summed-channel condition loses only 1% of the performance measured by equal error rates (EER) compared to the two-channel condition.

Index Terms— speaker recognition, summed-channel, speaker clustering, multi-session

1. INTRODUCTION

Speaker recognition on summed-channel speech data is an important task in practice. It can be applied to the records of two speaker scenarios such as interviews and telephone conversations, as the two-channel recording is not always available in those application scenarios. The summed-channel speaker recognition has been one of the evaluation tasks in the NIST speaker recognition evaluation (SRE) since 2005. In such a task, there are multiple excerpts from telephone conversations for each target speaker enrollment (also addressed as multi-session enrollment). Each enrolling excerpt includes both the target speaker and a non-target speaker. The non-target speakers are assumed to be distinct between multiple excerpts. The test excerpt from summed-channel telephone conversation includes either a target speaker and a non-target speaker or two non-target speakers.

There are two differences between summed-channel condition and two-channel condition, which contains only one speaker in each enrolling and test excerpt. One is to distinguish the voice of the intended target speaker from multiple enrolling excerpts, which is the main challenge in the task, and the other is the confidence measure adjustment to adapt the summed-channel condition in test phase. We are interested in addressing the above issues in this paper especially the former.

The general framework of the speaker recognition system on summed-channel condition we present is illustrated in Figure 1.

This work is supported by Beijing Natural Science Foundation (No. 4132071) and National Program on Key Basic Research Project (973 Program) under Grant 2013CB329302.

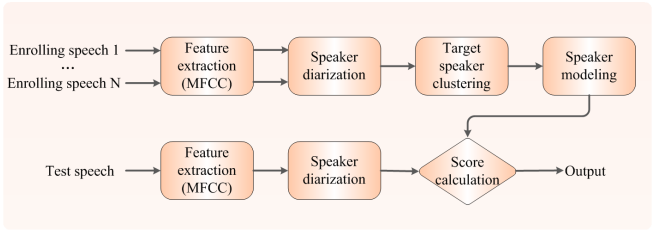

Fig. 1. General framework of the speaker recognition system on summed-channel condition

Compared with the previous reported systems [1, 2, 3] for two-channel speaker recognition, the summed-channel speaker recognition system contains two other phases: speaker diarization and target speaker clustering. The score calculation phase is also modified.

The paper is organized as follows. In Section 2, we give an overview of speaker diarization as the first stage of the system. The target speaker clustering approachs are presented in Section 3. The confidence measure in test stage is described in Section 4. Section 5 describes experimental results and discussions. Finally, we give the conclusion and discussions on relation to prior work in Section 6.

2. SINGLE-SESSION SPEAKER DIARIZATION

As the first stage of the whole recognition system, speaker diarization affects the performance of following stages greatly. P. Kenny presented variational Bayes based factor analysis for speaker diarization in [4, 5], which achieved significant performance improvement compared with previous works. Although eigenchannels are ineffective in this work, channel effects are usually found helpful in speaker diarization system because it seems to distinguish two conversation sides. So we extend this work to total variability space which potentially contains the speaker and channel variabilities simultaneously. The speaker diarization system based on variational Bayes consists of four phases: speech activity detection, speech segmentation, variational posterior calculation and Viterbi re-segmentation.

As for variational Bayes approach, we use one second uniform segmentation to begin with which assumes just one speaker in each segment. Baum-Welch statistics for each segment are accumulated previously. After that, the segment and speaker posteriors are updated alternately until convergence. On convergence, we assign each segment to the speaker which provides the maximal segment posterior (see [6] for more details).

Finally, Viterbi re-segmentation is applied to correct the crude initial segmentation of the data.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

3. TARGET SPEAKER CLUSTERING

As the NIST SRE described, a collection of enrolling excerpts for each target speaker are given on summed-channel training condition, denoted by $\{\boldsymbol{O}_1, \ldots, \boldsymbol{O}_R\}$. Here $R \geq 2$, otherwise we could not identify the target speaker with only one summed-channel excerpt.

After processed by the speaker diarization system described in Section 2, each excerpt is separated into two segments $\{\boldsymbol{O}_r(i)|1 \leq r \leq R, i \in \{0,1\}\}$. We tend to find the target speaker segments via different combinations of the $2R$ speech segments. There is only one correct combination among the 2^R possible combinations without consideration of speaker diarization errors. Hence the problem turns into how to find a binary sequence $\boldsymbol{i} = (i_1, i_2, \ldots, i_R)$ to maximize the objective function

$$\arg\max_{\boldsymbol{i}} \mathcal{F}(\boldsymbol{O}_1(i_1), \ldots, \boldsymbol{O}_R(i_R)) \tag{1}$$

$\mathcal{F}(\boldsymbol{i})$ is used as shorthand for the objective function in (1) in the remainder of this paper.

As speaker recognition technologies are dominated by the so-called iVector[1] representation of speech utterance in recent years [2], we apply a function f, called the iVector extractor, to every input $\boldsymbol{O}_r(i)$, so that $\boldsymbol{w}_r(i) = f(\boldsymbol{O}_r(i))$ is the associated iVector. On this basis, we present two iVector-based criterions to measure the confidences of different combinations $\boldsymbol{i}$.

3.1. Cosine distance criterion

Cosine distance, i.e., the angle between two iVectors, is an effective criterion used in speaker verification systems to represent the similarity of two speakers [2]. We use the cosine distance accumulation of multiple trials in $\boldsymbol{i}$ to measure the similarity of multiple speakers, written as

$$\mathcal{F}_{cos}(\boldsymbol{i}) = \sum_j \sum_{k \neq j} \frac{\boldsymbol{w}_j^{\mathsf{T}}(i_j)\boldsymbol{w}_k(i_k)}{\|\boldsymbol{w}_j^{\mathsf{T}}(i_j)\|\|\boldsymbol{w}_k(i_k)\|} \tag{2}$$

where the iVector $\boldsymbol{w}$ is projected onto a Linear Discriminative Analysis (LDA) basis.

3.2. A probability perspective

Heavy-tailed probabilistic LDA (PLDA) performed on iVectors yields state-of-the-art speaker verification results [3], which inspires us to describe the objective function from a probability perspective. In general, given R iVectors $\{\boldsymbol{w}_1(i_1), \ldots, \boldsymbol{w}_R(i_R)\}$ and a PLDA model $\mathcal{M}$, the log likelihood of those iVectors belonging to the same speaker is

$$\mathcal{F}_{like}(\boldsymbol{i}) = \mathcal{L}(i_1, \ldots, i_R|\mathcal{M}) \tag{3}$$

the definition of log likelihood $\mathcal{L}$ is presented in Section 4 of [3].

While (3) contains only the likelihood of R clustered segments, there are residual information about the other R segments which should belong to R different non-target speakers as the NIST SRE described. So we modify (3) as follows

$$\mathcal{F}_{\overline{like}}(\boldsymbol{i}) = \mathcal{L}(i_1, \ldots, i_R|M) + \mathcal{L}(j_1|\mathcal{M}) + \cdots + \mathcal{L}(j_R|\mathcal{M}) \tag{4}$$

where $i_r \oplus j_r = 1, 1 \leq r \leq R, \oplus$ is the exclusive OR operator.

Given the definition of likelihood, we can also describe the objective function from the viewpoint of posterior probability. The likelihood of the R speech segments in a combination $\boldsymbol{i}$ belonging to s speakers is written as $p(\boldsymbol{i}|s)$.

- $s = 1$, $\boldsymbol{i}$ is the unique correct combination,

$$\log p(\boldsymbol{i}|s=1) = \mathcal{L}(\boldsymbol{i}|\mathcal{M})$$

- $s = 2$, there are R kinds of combinations in total, using the sum role of probability, we obtain

$$\log p(\boldsymbol{i}|s=2) = \log \sum_r p(i_1, \ldots, i_{r-1}, i_{r+1}, \ldots, i_R|s=1)p(i_r|s=1)$$

- ...
- $s = R$, the speakers of R segments in $\boldsymbol{i}$ differ from one another ,

$$\log p(\boldsymbol{i}|s=R) = \log \prod_r p(i_r|s=1) = \sum_r \mathcal{L}(i_r|s=1)$$

Then we consider the prior $p(s=k)$ of those cases.

- $k = 1$, the unique correct combination is selected in 2^R candidate combinations, $p(s=1) = 1/2^R$
- $1 < k < R$, the selected combination contains k speakers,

$$p(s=k) = \frac{1}{2^R}\binom{R}{k-1} = \frac{R(R-1)\ldots(R-k+2)}{2^R(k-1)\ldots 1}$$

- $k = R$, the selected combination contains R non-target speakers or $R-1$ non-target speakers and the target speaker,

$$p(s=R) = \frac{(R+1)}{2^R}$$

Hence we can derive the posterior probability of a given combination containing k speakers using Bayes theorem [7]

$$p(s=k|\boldsymbol{i}) = \frac{p(\boldsymbol{i}|s=k)p(s=k)}{\sum_j p(\boldsymbol{i}|s=j)p(s=j)} \tag{5}$$

Define

$$\mathcal{F}_{post}(\boldsymbol{i}) = p(s=1|\boldsymbol{i}) \tag{6}$$

in which the $\boldsymbol{i}$ maximizing $\mathcal{F}(\boldsymbol{i})$ is the combination that has the max posterior probability of containing only one speaker among all candidate combinations. Similar to (4), we modify (6) as follows

$$\mathcal{F}_{\overline{post}}(\boldsymbol{i}) = p(s=1|\boldsymbol{i}) \times p(s=R|\boldsymbol{j}), i_r \oplus j_r = 1 \tag{7}$$

4. SUMMED-CHANNEL SCORING

After speaker diarization and target speaker clustering, we obtain speech segments of the target speaker. Speaker modeling techniques for two-channel speaker recognition can be used for enrollment. But in the summed-channel testing stage, there are still extra works compared to two-channel condition.

Generally, likelihood ratio is used for confidence measure on two-channel condition [3]. For instance, in iVector-based PLDA model system, given a collection of enrolling vectors of the target speaker $\boldsymbol{W} = \{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_R\}$ and a test vector $\boldsymbol{w}_x$, We wish to test the hypothesis $\mathcal{H}_s$ that they come from the same speaker against the hypothesis $\mathcal{H}_d$ that they come from different speakers. The likelihood ratio for this hypothesis test is

$$S(\boldsymbol{W}, \boldsymbol{w}_x) = \frac{p(\boldsymbol{W}, \boldsymbol{w}_x|\mathcal{H}_s)}{p(\boldsymbol{W}, \boldsymbol{w}_x|\mathcal{H}_d)} \tag{8}$$

On summed-channel condition, we obtain two test vectors $\boldsymbol{w}_{x0}$ and $\boldsymbol{w}_{x1}$ from a test excerpt after speaker diarization. (8) needs to be modified for this condition. Consider all possible hypotheses

[1] A vector of fixed dimension which contains most of the relevant information about the speaker identity.

1. $\boldsymbol{w}_{x0}$ and $\boldsymbol{W}$ are from the same speaker, while $\boldsymbol{w}_{x1}$ from another one
2. $\boldsymbol{w}_{x1}$ and $\boldsymbol{W}$ are from the same speaker, while $\boldsymbol{w}_{x0}$ from another one
3. $\boldsymbol{w}_{x1}$, $\boldsymbol{w}_{x0}$ and $\boldsymbol{W}$ are from three different speakers

The first two hypotheses are target tests, so the likelihood ratio is written as

$$S(\boldsymbol{W}, \boldsymbol{w}_{x0}, \boldsymbol{w}_{x1}) = \frac{p(\boldsymbol{W}, \boldsymbol{w}_{x0}|\mathcal{H}_s)p(\boldsymbol{w}_{x1}) + p(\boldsymbol{W}, \boldsymbol{w}_{x1}|\mathcal{H}_s)p(\boldsymbol{w}_{x0})}{p(\boldsymbol{W}, \boldsymbol{w}_{x0}, \boldsymbol{w}_{x1}|\mathcal{H}_d)} = \frac{p(\boldsymbol{W}, \boldsymbol{w}_{x0}|\mathcal{H}_s)}{p(\boldsymbol{W}, \boldsymbol{w}_{x0}|\mathcal{H}_d)} + \frac{p(\boldsymbol{W}, \boldsymbol{w}_{x1}|\mathcal{H}_s)}{p(\boldsymbol{W}, \boldsymbol{w}_{x1}|\mathcal{H}_d)} \tag{9}$$

Note that the likelihood ratio on summed-channel test condition equals to the sum of likelihood ratios of two speech segments calculated with training excerpts respectively. Another confidence measure formula can be given by using the max operator instead of sum operator.

$$S(\boldsymbol{W}, \boldsymbol{w}_{x0}, \boldsymbol{w}_{x1}) = \max\left(\frac{p(\boldsymbol{W}, \boldsymbol{w}_{x0}|\mathcal{H}_s)}{p(\boldsymbol{W}, \boldsymbol{w}_{x0}|\mathcal{H}_d)}, \frac{p(\boldsymbol{W}, \boldsymbol{w}_{x1}|\mathcal{H}_s)}{p(\boldsymbol{W}, \boldsymbol{w}_{x1}|\mathcal{H}_d)}\right) \tag{10}$$

For the cosine distance criterion, we choose the larger score directly as the distance has no sense of probability.

$$S(\boldsymbol{W}, \boldsymbol{w}_{x0}, \boldsymbol{w}_{x1}) = \max\Big(\sum_r \frac{\boldsymbol{w}_r^{\mathsf{T}}\boldsymbol{w}_{x0}}{\|\boldsymbol{w}_r\|\|\boldsymbol{w}_{x0}\|}, \sum_r \frac{\boldsymbol{w}_r^{\mathsf{T}}\boldsymbol{w}_{x1}}{\|\boldsymbol{w}_r\|\|\boldsymbol{w}_{x1}\|}\Big) \tag{11}$$

5. EXPERIMENTS

5.1. Datasets and configurations

5.1.1. Datasets

The summed-channel speaker recognition experiments are performed on the 3summed-summed subtask of the NIST SRE 2008 (SRE08) [8]. Here 3summed means that there are three summed-channel excerpts for each target speaker enrollment, and summed refers to the summed-channel test excerpt. We focus on female telephone data only, on which the state-of-the-art performance is worse than the one on male data. The meaning of female data here refers to the gender of target speaker, while there may be conservations of cross-gender in enrolling and test excerpts. In this case, we use gender-independent models for speaker diarization and gender-dependent models for speaker clustering and verification.

The corresponding subtask of 3summed-summed on two-channel condition is named 3conv-short3 in SRE08. We can obtain the trials mapping relationships between those two tasks using the information (e.g. speaker id) provided by NIST. We performed speaker recognition on 3conv-short3 task as the upper bound of the summed task. Finally, the female part of 3summed-summed task comprises 709 target speakers, $709 \times 3 = 2127$ enrolling files, 1503 test files and 14891 trails.

5.1.2. Configurations

Diarization system: operated on a 20-dimensional MFCC feature, gender-independent UBM with 1024 Gaussians, iVector extractor of dimension 100.

Clustering system and verification system: operated on a 60-dimensional feature which is formed by 20-dimensional MFCC appended with the first and second order derivatives, gender-dependent UBM with 2048 Gaussians, iVector extractor of dimension 800, LDA and heavy-tailed PLDA (HT-PLDA) models with speaker factor of dimension 200.

Table 1. *Corpora used to estimate the UBM, total variability matrix ($\boldsymbol{T}$), LDA and PLDA models.*

	Diarization		clustering and Verification			
	UBM	$\boldsymbol{T}$	UBM	$\boldsymbol{T}$	LDA	PLDA
Fisher	√	√	√	√		
SwitchBoard	√	√	√	√	√	√
NIST2004	√	√	√	√	√	√
NIST2005	√	√	√	√	√	√
NIST2006	√	√	√	√	√	√

Table 1 summarizes the development corpora of all sub systems. Finally, no score normalization technique is applied to any of the systems.

5.2. Results

5.2.1. Target speaker clustering results

The goal of the target speaker clustering is to find 3 target segments from 6 separated segments in the case of 3summed training condition. There are 8 possible combinations. To begin with, we separate each summed-channel enrolling excerpt into two segments according to the reference diarization answers provided by NIST. Although the speaker clustering accuracy can be evaluated by the performance of subsequent speaker recognition, we count the false speech segments clustered under objective functions described in Section 3 in order to contrast and analyze those objective functions.

Table 2 reports the counts of different error numbers for the 709 speakers on 3summed training condition. The first column shows the counts of speakers clustered without error (error = 0) and the second column shows the counts of speakers clustered with a false segment and two true segments (error = 1), and so on. The performance of those objective functions is similar except $\mathcal{F}_{like}$, with the best performance under $\mathcal{F}_{\overline{like}}$.

Table 2. *Counts of different error numbers on 3summed training condition under different objective functions.*

error	0	1	2	3
$\mathcal{F}_{cos}$	688	14	0	7
$\mathcal{F}_{like}$	525	95	41	48
$\mathcal{F}_{\overline{like}}$	697	6	0	6
$\mathcal{F}_{post}$	693	7	0	9
$\mathcal{F}_{\overline{post}}$	691	12	0	6

In the case of $R = 3$, the binary sequence $\boldsymbol{i}$ has 8 possible combinations. The calculations of objective functions are tractable for small R. As R increasing, the number of possible combinations grows exponentially. We can partition the excerpts into groups so that each group contains $3 \sim 5$ excerpts only.

5.2.2. Summed-channel speaker recognition results

To evaluate the benefit from speaker diarization and clustering, we first conduct speaker recognition experiments on summed-channel speech data without diarization, as the lower bound of this task.

Table 3. *Speaker recognition performance on 3summed-summed condition without speaker diarization.*

	EER(%)	minDCF
LDA	19.3	0.0874
HT-PLDA	18.8	0.0736

Table 4. *Speaker recognition performance on 3summed-summed condition with different objective functions for target speaker clustering.*

	LDA		HT-PLDA	
	EER(%)	minDCF	EER(%)	minDCF
$\mathcal{F}_{cos}$	**6.460**	**0.0316**	4.109	0.0204
$\mathcal{F}_{like}$	11.429	0.0436	8.387	0.0355
$\mathcal{F}_{\overline{like}}$	6.521	0.0317	**4.052**	**0.0199**
$\mathcal{F}_{post}$	6.582	0.0320	4.216	0.0209
$\mathcal{F}_{\overline{post}}$	6.575	0.0319	4.098	0.0201

Then speaker diarization and clustering are applied to the summed-channal data. Finally, we compare the performance of different combinations between summed-channel and two-channel speech data.

Table 3 shows the poor performance of system without speaker diarization because of the impurity of the speech data. Table 4 reports the speaker recognition performance based on LDA and HT-PLDA with different objective functions for target speaker clustering. It can be seen that speaker diarization and clustering significantly improve the summed-channel speaker recognition performance in terms of both EER and minDCF. The performance of five proposed objective functions is similar except $\mathcal{F}_{like}$, which is consistent with the results in Table 2. It indicates that the cosine distance criterion works as well as the more complex PLDA approach. Furthermore, the sum and max operation in HT-PLDA confidence measure (Formula (9) and (10)) are equivalent in our task.

In order to evaluate the upper bound of summed-channel speaker recognition under proposed system, we also report the performance on 3conv-short3 subtask in SRE08, which is the corresponding task of 3summed-summed on two-channel condition, and different combinations between their training and test conditions in Table 5. 3conv and short3 refer to the two-channel training and test condition respectively. 3summed$_1$ and summed$_1$ denote the reference diarization which is assumed to have no diarization error. 3summed$_2$ and summed$_2$ denote the diarization described in Section 2. $\mathcal{F}_{\overline{like}}$ is used as the objective function for all summed-channel training conditions.

As the results shown in Table 5 and Fig.2, system performance degrades obviously from short3 to summed test conditions due to the max (or sum) operation in each summed-channel trials. However, the impact from 3conv to 3summed training conditions is less. It indicates that the target speaker clustering approach is effective with reliable front-end speaker diarization.

In summary, the performance loss from two-channel condition to summed-channel condition in our work can be attributed to the following three aspects. Firstly, the speaker diarization on summed-channel condition is not accurate enough. Secondly, the objective function of target speaker clustering in training phase needs to be developed further. The accurate rate of objective function with best performance is $697/709 = 98.30\%$ even if the speaker diarization is supposed to have no error. The performance degrades further with the increasing of diarization error rate. Finally, the decision between two confidence scores incurs additional risks in false alarm.

Table 5. *Speaker recognition performance of different combinations between training and test conditions in terms of EER(%) based on HT-PLDA model.*

training \ test	short3	summed$_1$	summed$_2$
3conv	3.082	3.601	4.037
3summed$_1$	3.270	3.788	4.209
3summed$_2$	3.585	3.972	4.052

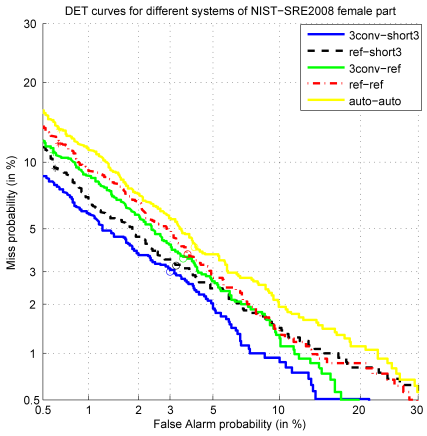

Fig. 2. Detection error tradeoff (DET) curves of different combinations between training and test conditions based on HT-PLDA model

6. CONCLUSION AND RELATION TO PRIOR WORK

This paper studies about the NIST speaker recognition tasks on summed-channel condition. The proposed summed-channel speaker recognition system includes three major stages: speaker diarization, target speaker clustering and speaker verification. Several objective functions for target speaker clustering are presented and the best one among them achieves 98.30% accurate rate with reference diarization on the female part of 3summed training dataset. We also give the confidence measures for summed-channel trials scoring. Experiments on NIST SRE 2008 show that the system works well on summed-channel condition. We achieve an EER of 4.05% on female telephone trials of 3summed-summed task based on heavy-tailed PLDA model, which is only 1% absolute EER performance loss compared to the result obtained on two-channel condition under same approach.

Previous works on summed-channel evaluation tasks can be found in [9, 10, 11, 12]. The works reported in [9, 10] considered speaker recognition with summed-channel data on test condition, while the enrolling data were two-channel recorded. This paper presents a summed-channel speaker recognition system applied on the task whose enrolling and test data are both recorded in summed-channels. [11, 12] mainly studied about cross-show speaker diarization, which had some similarities with the target speaker clustering presented in this paper. However, we are concerned about the target speaker only in the framework of speaker recognition, which makes us analysis the problem from different perspectives. Several target speaker clustering methods for multi-session enrollment are proposed and achieve remarkable performance in our work.

7. REFERENCES

[1] Douglas A Reynolds, Thomas F Quatieri, and Robert B Dunn, "Speaker verification using adapted gaussian mixture models," *Digital signal processing*, vol. 10, no. 1, pp. 19–41, 2000.

[2] Najim Dehak, Patrick J Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-end factor analysis for speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 4, pp. 788–798, 2011.

[3] Patrick Kenny, "Bayesian speaker verification with heavy tailed priors," in *Speaker and Language Recognition Workshop (IEEE Odyssey)*, 2010.

[4] Patrick Kenny, Douglas Reynolds, and Fabio Castaldo, "Diarization of telephone conversations using factor analysis," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 4, no. 6, pp. 1059–1070, 2010.

[5] Patrick Kenny, "Bayesian analysis of speaker diarization with eigenvoice priors," *CRIM, Montreal, Technical Report*, 2008.

[6] Rong Zheng, Ce Zhang, Shanshan Zhang, and Bo Xu, "Variational bayes based i-vector for speaker diarization of telephone conversations," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014.

[7] C.M. Bishop et al., *Pattern recognition and machine learning*, vol. 4, NY:springer Science+Business Media, LLC, 2006.

[8] NIST-SRE 2008, "The NIST year 2008 speaker recognition evaluation plan," `http://www.itl.nist.gov/iad/mig/tests/sre/2008/index.html`, 2008.

[9] Hagai Aronowitz and Yosef Solewicz, "Speaker recognition in two wire test sessions," in *INTERSPEECH*, 2008, p. 865 C 868.

[10] Hanwu Sun, Bin Ma, Chien-Lin Huang, Trung Hieu Nguyen, and Haizhou Li, "The iir nist sre 2008 and 2010 summed channel speaker recognition systems.," in *INTERSPEECH*, 2010, pp. 366–369.

[11] Qian Yang, Qin Jin, and Tanja Schultz, "Investigation of cross-show speaker diarization.," in *INTERSPEECH*, 2011, pp. 2925–2928.

[12] Viet-Anh Tran, Viet Bac Le, Claude Barras, and Lori Lamel, "Comparing multi-stage approaches for cross-show speaker diarization.," in *INTERSPEECH*, 2011, pp. 1053–1056.