Odyssey 2012
The Speaker and Language Recognition Workshop
25-28 June 2012, Singapore

Study on the Effects of Intrinsic Variation using i-Vectors in Text-Independent Speaker Verification

Sheng Chen, Mingxing Xu, Emlyn Pratt

Key Laboratory of Pervasive Computing, Ministry of Education
Tsinghua National Laboratory for Information Science and Technology (TNList)
Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China

Abstract

Speaker verification performance is adversely affected by mismatches between training and testing data in intrinsic variations. This paper explores how recent technologies focused on modeling the total variability behave in addressing the effects of intrinsic variation in speaker verification. The effects of intrinsic variation are investigated from six aspects including speaking style, speaking rate, speaking volume, emotional state, physical status, and speaking language. The speaker and session variability are modeled with the i-vector framework in the total variability space and the cosine similarity is used as the final decision score in the i-vector based speaker verification system. Intrinsic variations are compensated in the i-vector framework with a variety of techniques, specifically Linear Discriminant Analysis (LDA), Within-Class Covariance Normalization (WCCN) and Nuisance Attribute Projection (NAP). Experiments in the intrinsic corpus show that speaker volume has dramatic effects on the results of speaker verification systems and whisper speech brings the largest degradation of speaker verification performance. The best results are obtained by i-vector modeling with the combined compensation of LDA and WCCN in the i-vector based systems. Compared to the GMM-UBM based system, around 36.76% relative improvement in Equal Error Rate (EER) is obtained in the i-Vector+LDA+WCCN system.

1. Introduction

Performances of speaker verification systems are adversely affected by extrinsic variability such as mismatched channels and environmental noise, and intrinsic variability associated with factors from the speaker such as speaking style, emotion, speech volume, state of health and so on. Significant progress has been made in recent years to address the effects of extrinsic variability with a variety of model domain approaches, including eigenchannel compensation [1], joint factor analysis [2] [3] and total variability modeling [4]. A few effective techniques such as feature warping [5], short-time Gaussianization [6] and feature mapping [7] are also proposed in the feature domain to address the problems caused by multi-channels and noise.

Compared to the significant progress in solving extrinsic variation problems, a limited amount of research has been done to address problems caused by intrinsic variation for speaker verification in recent years. Shriberg studied the effects of vocal effort and speaking style in GMM-UBM based speaker verification systems in [8] and found that vocal effort level has a dramatic effect on speaker verification performance. Studies in [9] demonstrate that the performances of the traditional GMM-UBM based speaker verification systems decline sharply due to the effects of device, language and environmental mismatch. Studies in [10] [11] show an important influence of the emotional state upon text-independent speaker recognition. Score normalization is used in [12] for speaking-style variation robust speaker recognition.

The i-vector approach has been successfully applied in [13] to model both speaker and channel variability as state-of-the-art technology. The supervectors from utterances are projected to i-vectors, which represent the hidden variables of speaker and channel factors in the low-dimensional total variability space. The cosine distance score is used to measure the similarity between the enrollment and testing utterances to make the final decision for speaker verification. It is indicated in [14] that techniques originally designed for channel compensation are indeed modeling the intrinsic variation represented in the data.

In this paper, we intend to study the effects of intrinsic variation using an i-vector approach from six aspects: speaking style, speaking rate, speaking volume, emotional state, physical status and speaking language. An intrinsic variation corpus is designed to support our study on the effects of a variety of intrinsic variations, including reading, fast, slow, loud, soft, whispered, angry, happy, denasalized, mumbled, and English. A GMM-UBM based speaker verification system is used as the baseline system in measuring speaker verification performance. Several i-vector based speaker verification systems are built up to address the effects of intrinsic variation for robust speaker verification. A variety of compensation techniques, including LDA, WCCN and NAP are applied to improve speaker discrimination in the intrinsic variation corpus.

This paper is organized as follows. The intrinsic variation corpus used in our experiments is described in the section 2. In section 3, we will introduce the i-vector framework for the intrinsic variation modeling and a variety of intersession compensation techniques, including LDA, WCCN and NAP. Experimental setup is described in section 4 and results on the intrinsic variation corpus are discussed in section 5. Finally, section 6 concludes the paper.

2. Intrinsic Variation Corpus

An intrinsic variation corpus has been designed and collected to study the effects of intrinsic variation in speaker verification. The extrinsic factors such as mismatched channels and environmental noise are excluded in the intrinsic variation corpus design. Effects of voice change due to aging process are also out of research domain in this paper. In the following sections, we will introduce twelve intrinsic variation forms from six aspects of common intrinsic variations and describe the speech data and recording conditions in the intrinsic variation corpus.

2.1. Intrinsic Variation Forms

Considering that various kinds of intrinsic variations exist in realistic scenarios, the intrinsic variation corpus is designed from six different aspects including speaking style, speaking rate, speaking volume, emotional state, physical status and speaking language. The six intrinsic factors are considered separately and each variation form can be represented by a six dimension tuple of form <style, rate, volume, emotion, physical, language>.

To obtain different forms of intrinsic variation, we define the intrinsic variation form of the neutral spontaneous speech at normal rate and volume in Chinese as the base form. Taking into account the six factors of intrinsic variation described above, eleven different variation forms are derived from the base form. Figure 1 shows the derivation process of the eleven variation forms, which are simply noted as reading, fast, slow, loud, soft, whispered, angry, happy, denasalized, mumbled, and English, respectively. Including the base form, referred to as spontaneous, a total of twelve intrinsic variation forms are obtained for the design of the intrinsic variation corpus.

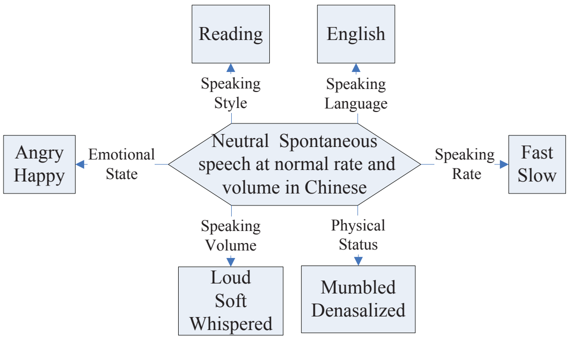

Figure 1: Derivation process of the twelve intrinsic variation forms.

The eleven derivation forms are obtained from the spontaneous base form by changing just one dimension of the six aspects while the other aspects remains unchanged. Since the emotional state is closely connected with the speech rate and volume, the six aspects are not absolutely independent for each other.

2.2. Speech Data and Recording Conditions

For the intrinsic variation corpus, we collected speech data from 110 (46 males and 64 females) native Chinese-speaking university students whose ages ranged from 18 to 29 years old. Given the 12 intrinsic variation forms described above, each subject speaks with each variation form for about 3 minutes. Utterances with 12 variation forms from 110 subjects are recorded with a sample rate of 8k, a resolution of 8 bits and mono channel into the intrinsic variation corpus for the study of effects of intrinsic variation on robust speaker verification.

The whole recording process was completed in a quiet office acoustically insolated from the surrounding environment. Each subject spoke with a wireless headset in the office and the speech data was then transmitted and recorded into a laptop in another room, where staff operated the recording software. The separated environments help the subject to focus on their speaking process with different variation forms and avoid distractions caused by the recording staff. This recording condition environment ensures that the effects of channels and environmental noise are minimal and that the collected speech data can be used to analyze the effects of the intrinsic variation for speaker verification. More details about the intrinsic variation corpus can be found in [15].

3. i-Vector Framework for Intrinsic Variation Modeling

The i-vector framework has been successfully used to model the speaker and channel factors for speaker verification. Based on the assumption that intersession compensation techniques originally designed for channel compensation can indeed model the intrinsic variation, the i-vector framework is used to model the speaker variability and intrinsic variation with a variety of intersession compensation techniques. In the following sections, we will discuss the total variability modeling and several intrinsic variation compensation techniques including LDA, WCCN and NAP.

3.1. Total Variability

A total variability space is defined to model the speaker and session variability simultaneously in the i-vector framework. The total variability modeling assumes that an utterance can be represented by the speaker- and session-dependent Gaussian Mixture Model (GMM) supervector defined by the following equation

$$\boldsymbol{M} = m + \boldsymbol{T}w \tag{1}$$

where m is the speaker- and session-independent supervector which can be obtained by the Universal Background Model (UBM) training, $\boldsymbol{T}$ is rectangular matrix of low rank representing the total variability space and w represents the total variability factors and is referred to as the identity vector or i-vector. The i-vector model can be seen as a method of factor analysis which projects a speech utterance into the low-dimensional total variability space. After the i-vectors are obtained, the process of compensation and scoring becomes considerably more computationally efficient in the i-vector space compared to the supervector space.

The i-vector w, used to model the speaker and session variability stands for the hidden variables of the total factors, which have a standard normal distribution $N(0, I)$. In order to extract the i-vector w, the Baum-Welch statistics need to be calculated for a given speech utterance with the UBM composed of C Mixture components defined in the feature space of dimension F. The Baum-Welch zero order statistics $\boldsymbol{N}$ and first order statistics $\boldsymbol{F}$ are defined for a given speaker s and acoustic features $\{x_1, x_2 \cdots x_L\}$ for each mixture component c by the following equations:

$$\boldsymbol{N}_c(s) = \sum_{t=1}^{L} \gamma_t(c) \tag{2}$$

$$\boldsymbol{F}_c(s) = \sum_{t=1}^{L} \gamma_t(c) x_t \tag{3}$$

where $c = 1, 2, \cdots, C$ is the Gaussian component index and $\gamma_t(c)$ corresponds to the posterior possibility of components c generating the acoustic feature x_t. The first order Baum-Welch statistics are then centered by the following equation

$$\tilde{\boldsymbol{F}}_c(s) = \sum_{t=1}^{L} \gamma_t(c)(x_t - m_c) \tag{4}$$

where m_c is the mean vector of UBM component c. Given an utterance, the corresponding i-vector w is can be calculated in the following equation

$$w = (\boldsymbol{I} + \boldsymbol{T}^t\Sigma^{-1}\boldsymbol{N}(s)\boldsymbol{T})^{-1}\boldsymbol{T}^t\Sigma^{-1}\tilde{\boldsymbol{F}}(s) \tag{5}$$

where $\boldsymbol{I}$ is the $CF * CF$ identity matrix, $\boldsymbol{N}(s)$ is the $CF * CF$ diagonal matrix composed of F blocks of $\boldsymbol{N}_c(s)\boldsymbol{I}(c = 1, 2, \cdots, C)$ and Σ is the UBM diagonal covariance matrix of dimension $CF * CF$. The supervector $\tilde{\boldsymbol{F}}(s)$ is formed by concatenating the centered first order statistics.

3.2. Cosine Similarity Scoring

After i-vectors from utterances are obtained in the low dimensional total variability space, cosine distance score is used to measure the similarity between the enrollment and testing utterances for making the speaker verification decision. Given the target speaker i-vector w_{target} from known speaker and the testing i-vector w_{test} from unknown speaker, the cosine distance score between them is defined by the following equation

$$score(w_{target}, w_{test}) = \frac{\langle w_{target}, w_{test}\rangle}{\|w_{target}\| \, \|w_{test}\|} \tag{6}$$

Since the magnitude of the i-vector is easily affected by the session variability, only the angle between the target and testing i-vectors is used as the decision score in the scoring process to improve the robustness of the i-vector system while the factor of the i-vector magnitude is not considered.

3.3. Intrinsic Variation Compensation

Extracted i-vectors in their raw forms are not optimized for speaker discrimination due to the effects of intrinsic variation. After the low-dimensional i-vectors are obtained in the total variability space, it is straight forward and computationally efficient to apply intrinsic variation compensation techniques to the i-vectors prior to classification. A number of existing approaches borrowed from SVM speaker verification such as LDA, WCCN and NAP are used as the intersession compensation techniques to remove the nuisance effects.

3.3.1. Linear Discriminant Analysis

LDA is widely used in the field of pattern recognition as a technique for dimension reduction. In order to compensate for the intersession variability, LDA is used in the context of the i-vector framework to enhance discrimination between i-vectors of different speakers. LDA aims to find a reduced set of axes that minimize the within-speaker variability observed in the i-vectors while simultaneously maximizing the between-speaker variability. This is accomplished by defining a projection matrix $\boldsymbol{A}$ formed as the subset of the eigenvectors of the general eigenvalue equation as follows:

$$\boldsymbol{S}_B v = \lambda \boldsymbol{S}_W v \tag{7}$$

where the between speaker covariance matrix $\boldsymbol{S}_B$ and the within speaker covariance matrix $\boldsymbol{S}_W$ are respectively defined as

$$\boldsymbol{S}_B = \sum_{s=1}^{S} (\mu_s - \mu)(\mu_s - \mu)^t \tag{8}$$

$$\boldsymbol{S}_W = \sum_{s=1}^{S} \frac{1}{N_s} \sum_{i=1}^{N_s} (w_{i,s} - \mu_s)(w_{i,s} - \mu_s)^t \tag{9}$$

Here, $\mu_s = \frac{1}{N_s}\sum_{i=1}^{N_s} w_{i,s}$ is the mean of i-vectors for speaker s, S is the number of speakers, and N_s is the number of utterances for each speaker s. The speaker population mean vector μ is equal to the zero vector due to the factor analysis assumption that the i-vectors have a standard distribution $N(0, I)$.

After the projection matrix $\boldsymbol{A}$ is obtained from LDA, the new cosine distance scoring between two i-vectors w_1 and w_2 is calculated as

$$score(w_1, w_2) = \frac{(\boldsymbol{A}^t w_1)^t(\boldsymbol{A}^t w_2)}{\sqrt{(\boldsymbol{A}^t w_1)^t(\boldsymbol{A}^t w_1)}\sqrt{(\boldsymbol{A}^t w_2)^t(\boldsymbol{A}^t w_2)}} \tag{10}$$

3.3.2. Within Class Covariance Normalization

WCCN was introduced by Andrew Hatch in [16] and has been successfully applied in SVM modeling based on linear separation between target speaker and imposters using a one-versus-all decision. The idea behind WCCN is to deemphasize the direction of high intra-speaker variability in i-vector comparisons by scaling the total variability space inversely proportional to an estimate of the within-class covariance matrix. We assume that all utterances from the same speaker belong to one class and the within-class covariance matrix computed over all the imposters in the training background as follows

$$\boldsymbol{W} = \frac{1}{S}\sum_{s=1}^{S} \frac{1}{N_s} \sum_{i=1}^{N_s} (w_{i,s} - \mu_s)(w_{i,s} - \mu_s)^t \tag{11}$$

The number of speakers is S, each speaker provides N_s utterances in the training data and $\mu_s = \frac{1}{N_s}\sum_{i=1}^{N_s} w_{i,s}$ is the mean of i-vectors for speaker s. The i-vectors are normalized by the inverse of the within-speaker covariance matrix, which is equivalent to scaling the i-vector space with the projection matrix $\boldsymbol{B}$ which can obtained through Cholesky decomposition of the inverse of the within-speaker covariance matrix as follows

$$\boldsymbol{W}^{-1} = \boldsymbol{B}\boldsymbol{B}^t \tag{12}$$

The new version of cosine distance scoring by WCCN is given by the following equation

$$score(w_1, w_2) = \frac{{w_1}^t \boldsymbol{W}^{-1} w_2}{\sqrt{{w_1}^t \boldsymbol{W}^{-1} w_1}\sqrt{{w_2}^t \boldsymbol{W}^{-1} w_2}} \tag{13}$$

where w_1 and w_2 are the total i-vectors and $\boldsymbol{W}$ is the within-class covariance matrix.

While the WCCN approach focuses on attenuating dimensions of high within-class variability, it can also remove information about the between-class variability contained within the attenuated dimensions. In order to alleviate this problem with the WCCN approach, we can firstly use LDA to project the i-vectors into a new subspace that minimizes the within-class variance and maximizes the between-class variance and then apply WCCN to the transformed i-vectors to improve the classification performance.

3.3.3. Nuisance Attribute Projection

The nuisance attribute projection algorithm is presented in [17] to address the effects of variability directly in the SVM kernel. The NAP approach accomplishes this by performing projection in a similar manner to WCCN. However, NAP attempts to find an appropriate projection matrix to remove the nuisance direction rather than weighting the i-vector dimensions by WCCN. The NAP transformation matrix has the form

$$\boldsymbol{P} = \boldsymbol{I} - \boldsymbol{V}\boldsymbol{V}^t \tag{14}$$

where $\boldsymbol{I}$ is the identity matrix and $\boldsymbol{V}$ is a rectangular matrix of low rank which can be obtained by taking the top k eigenvectors having the best eigenvalues of the same within-class covariance matrix as defined in Equation 11.

Given two i-vectors w_1 and w_2, the cosine similarity based on the NAP matrix is given as follows

$$score(w_1, w_2) = \frac{(\boldsymbol{P}w_1)^t(\boldsymbol{P}w_2)}{\sqrt{(\boldsymbol{P}w_1)^t(\boldsymbol{P}w_1)}\sqrt{(\boldsymbol{P}w_2)^t(\boldsymbol{P}w_2)}} \tag{15}$$

4. Experimental Setup

Experiments are performed in the intrinsic variation corpus and a NIST SRE-like task is created to study the effects of intrinsic variations. GMM-UBM based speaker verification system is chosen as the baseline system and four i-vector based systems, namely i-Vector+LDA, i-Vector+WCCN, i-Vector+NAP and i-Vector+LDA+WCCN are built up to address intrinsic variations.

4.1. Experiment data

The intrinsic variation corpus introduced in section 2 is used in our experiments for the intrinsic variation robust speaker verification. The duration of each enrollment utterance and testing utterance is about 18 seconds. Table 1 presents the data partitions in the intrinsic variation corpus. The gender independent UBM for the GMM-UBM baseline system and i-vector based speaker verification systems is trained using data from 30 speakers (15 males and 15 females) lasting for 18 hours with all 12 intrinsic variations. The total variability space matrix is trained with data from another group of 30 speakers, of the same total duration and gender proportion. Speech data with 12 intrinsic variations from 20 speakers is used to train the projection matrixes of LDA, WCCN and NAP for the intrinsic variation compensation. The testing data set is composed of 2400 utterances with 12 intrinsic variations from 20 speakers (10 males and 10 females).

Table 1: Data partitions in the intrinsic variation corpus

Function	Source	Description
UBM traing data	30 speakers	18 hours 12 variation forms
Training data used for total variability space	30 speakers	18 hours 12 variation forms
Training data used for LDA,WCCN and NAP	20 speakers	12 hours 12 variation forms
Testing data	20 speakers	2400 utterances 12 variation forms

Mel Frequency Cepstral Coefficients (MFCC) with 32ms window length and 16ms frame rate are extracted as features to be modeled in the speaker verification systems in our experiments. The MFCC features are composed of 12 cepstral coefficients and energy, adding derivation of first and second order to produce 39 dimensional feature vectors.

4.2. Task and Performance Measure

A NIST SRE-like task is created in the intrinsic variation corpus to investigate the performances of different speaker verification systems with a variety of intrinsic conditions. Each recording of the subject is used as an enrollment utterance to train the target speaker model. In the testing process, all the other recordings of the same subject are used to create target trials and all recordings of the other subjects are used to create imposter trials. The content of each recording are different from each other to ensure that the speaker verification task is text-independent.

The performance of a speaker verification system is usually measured by false acceptance rate (FAR) and false rejection rate (FRR). False acceptance occurs when the system incorrectly accepts an imposter and false rejection occurs when the system incorrectly rejects the target speaker. With different score threshold settings, FAR and FRR are presented in the DET plot and the equal error rate (EER) is obtained when FAR equals FRR. The EER and DET curve are used to measure the performances of speaker verification systems in our experiments.

4.3. System Description

4.3.1. GMM-UBM Baseline System

The baseline system is based on the GMM-UBM model paradigm in which a speaker model of Gaussian mixture model (GMM) is adapted from a universal background model (UBM) with the adaption of Maximum a posteriori (MAP). GMM is used for modeling the probability density function of a multi-dimensional feature vector. Given a speech feature vector x, the probability density of x in a Gaussian mixture model λ can be defined as

$$P(x|\lambda) = \sum_{i=1}^{M} \omega_i g(x, \mu_i, \Sigma_i) \tag{16}$$

where there is the additional constraint of $\sum_{i=1}^{M} \omega_i = 1$ and $g(x, \mu_i, \Sigma_i)$ is the probability density function of single Gaussian model with mean vector μ_i and variance matrix Σ_i.

The decision score is calculated by testing the candidate speech utterance U against the adapted target speaker model and the UBM in the following equation

$$S(U) = \log P(U|\lambda_{TAR}) - \log P(U|\lambda_{UBM}) \tag{17}$$

where λ_{TAR} is the target speaker model and λ_{UBM} is the universal background model. Speaker verification is achieved by comparing the decision score against a threshold with the result of acceptance or rejection.

The gender independent UBM used in the baseline system is composed of 512 Gaussian mixtures. Given the enrollment utterance with duration about 18 seconds, the target speaker model is obtained by adapting the UBM with the MAP approach. Utterances with the same duration from unknown speaker are evaluated by the UBM and target speaker model in the testing process and the likelihood ratio score obtained in Equation 17 is used to compare with the threshold to make the final verification decision.

4.3.2. i-Vector based Speaker Verification Systems

Four i-vector based speaker verification systems, namely i-Vector+LDA, i-Vector+WCCN, i-Vector+NAP and i-Vector+LDA+WCCN are built up to address the effects of the intrinsic variation in our experiments. A gender independent UBM with 512 gaussian mixtures is trained to be used as the speaker and session independent model in the i-vector based speaker verification systems. The i-vector w is made of 200 total factors defined in the total variability space. The total variability matrix $\boldsymbol{T}$ is trained using speech data with the 12 intrinsic variations from 30 speakers.

After i-vectors are obtained in the total variability space, several techniques including LDA, WCCN and NAP are applied to the i-vectors to compensate for the intrinsic variations. Speech data from 20 speakers with duration of 12 hours are used as the developing data set to train the LDA projection matrix $\boldsymbol{A}$, WCCN projection matrix $\boldsymbol{B}$ and NAP projection matrix $\boldsymbol{P}$. The LDA projection matrix $\boldsymbol{A}$ is composed of the 25 best eigenvectors from the Equation 7. The WCCN transformation matrix $\boldsymbol{B}$ is obtained with the Cholesky decomposition in Equation 12. The NAP projection matrix $\boldsymbol{P}$ is calculated in Equation 15 in which matrix $\boldsymbol{V}$ is composed of top 100 eigenvectors having the best eigenvalues of the whin-class covariance matrix. Following the LDA compensation method, WCCN is used to further compensate for intrinsic variability in the i-Vector+LDA+WCCN system. After the compensated i-vectors are obtained from enrollment and testing utterances, the cosine similarity is used as the decision score for speaker verification.

5. Experimental Results and Discussion

In the following experiments, we investigated the effects of intrinsic variation in matched and mismatched conditions. A variety of techniques including LDA, WCCN and NAP are applied in the i-vector framework to compensate for intrinsic variation. Performances of i-vector based systems in the intrinsic corpus and effects of different intrinsic variations on speaker verification performances are presented in the following sections.

5.1. Performances of i-vector based systems in intrinsic variation conditions

The GMM-UBM based system is used as the baseline system. Performances of four i-vector based systems, namely i-Vector+LDA, i-Vector+WCCN, i-Vector+NAP and i-Vector+LDA+WCCN are investigated in the intrinsic variation corpus in our experiments. The speaker verification performances are investigated for each enrollment condition by testing the speaker verification systems using utterances with all the twelve variation forms. The experiments are performed on 2400 utterances from 20 speakers for each enrollment condition. There are 12 conditions in the intrinsic variations and every subject has 10 utterances for each condition. For each subject, 1 utterance of the enrollment condition is used for training and the other 119 utterances are used to create target trials. All the 2280 utterances from the other subjects are used to create imposter trials. The training and testing procedures are repeated 10 times by choosing another enrollment utterance with the enrollment condition.

Table 2 presents the speaker verification results obtained in the GMM-UBM baseline system and four i-vector based systems. The best results of the five speaker verification systems are formatted in bold for each row. It is obvious that i-vector based systems outperform the GMM-UBM based system in modeling intrinsic variation for each enrollment condition. The whisper enrollment condition causes the largest degradation of performance for each speaker verification system. The significant improvement of EER in the whisper enrollment condition is still very impressive from 46.93% in the GMM-UBM based system to 25.88% in the i-Vector+LDA+WCCN system.

For all the enrollment conditions, the overall EERs are calculated for each speaker verification system and the results are presented in Table 3. Significant relative improvements of EER around 21.89%, 27.65%, 23.32% and 36.76% are obtained respectively in the i-vector based systems of i-Vector+LDA, i-Vector+WCCN, i-Vector+NAP and i-Vector+LDA+WCCN compared to the GMM-UBM baseline system. Figure 2 shows the corresponding DET curve for the speaker verification systems. It is obvious that the best results are obtained with the combination of LDA and WCCN for the intrinsic variation compensation. These results demonstrate the effectiveness of i-vector framework in modeling intrinsic variations.

5.2. Effects of intrinsic variation in matched and mismatched conditions between training and testing

We study the effects of intrinsic variation on speaker verification performances in matched and mismatched conditions between enrollment and testing. Each subject has 10 utterances for each variation form. In order to evaluate the performances of speaker verification systems in matched conditions, 1 utter-

Table 2: EERs(%) for each enrollment condition when testing is done by using utterances with all the twelve variation forms.

Speech Variation	Variation Form	GMM-UBM	i-Vector+LDA	i-Vector+WCCN	i-Vector+NAP	i-Vector+LDA+WCCN
Base Case	Spontaneous	21.39	18.66	14.41	16.89	**13.57**
Speaking Style	Reading	25.13	20.25	17.39	20.46	**14.92**
Speaking Volume	Loud	27.10	20.08	21.22	19.96	**17.06**
	Soft	31.51	22.69	17.82	20.63	**16.68**
	Whispered	46.93	32.69	33.24	32.02	**25.88**
Speaking Rate	Fast	27.49	23.19	20.80	22.52	**19.92**
	Slow	23.03	19.58	19.12	18.15	**16.93**
Emotional State	Angry	26.60	23.28	21.43	23.36	**19.41**
	Happy	23.49	18.49	16.43	18.24	**15.42**
Physical Status	Denasalized	20.71	17.94	16.39	19.75	**14.41**
	Mumbled	22.52	18.49	16.34	18.07	**15.25**
Speaking Language	English	18.57	18.24	15.55	18.49	**14.83**

Table 3: Performances of i-vector based systems

System	EER(%)	Relative Reduction(%)
GMM-UBM(baseline)	27.33	
i-Vector+LDA	21.35	21.89
i-Vector+WCCN	19.78	27.65
i-Vector+NAP	20.96	23.32
i-Vector+LDA+WCCN	**17.29**	**36.76**

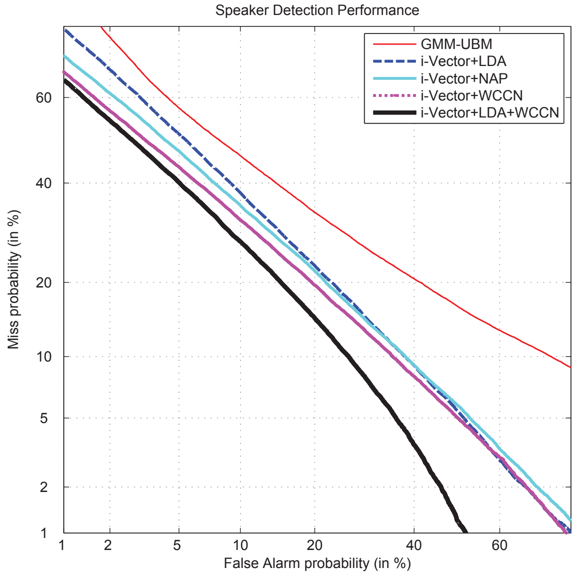

Figure 2: DET curve for GMM-UBM based system and four i-vector based systems.

ance is used for enrollment while the other 9 utterances are used for target trials. 190 utterances from other 19 speakers are used for imposter trials. The enrollment and testing procedure is repeated 10 times for each subject by choosing another enrollment utterance. The effects of mismatched conditions are investigated with the same procedure in the GMM-UBM system and i-Vector+LDA+WCCN system except that 119 utterances and 2280 utterances with mismatched variation forms are used for target trials and imposter trails, respectively.

The results of matched and mismatched conditions are presented in figure 3 which shows the EERs obtained in the GMM-UBM system and the i-Vector+LDA+WCCN system for each enrollment form. It is noted that mismatches between training and testing data in intrinsic variations cause sharp degradation in speaker verification performance. Effects of low volume level are more obvious than that of other variations in matched conditions with EER 13.89% and 17.22% for soft and whisper condition respectively in the GMM-UBM system. The i-Vector+LAD+WCCN system is much more robust for the effects of low vocal effects with EER 3.89% and 6.11% in the corresponding conditions.

The i-Vector+LDA+WCCN system performs much better than the GMM-UBM system for each enrollment form in mismatched conditions. The effect of speech volume is the most significant in all the variations. The speaker verification performance declines sharply when the speech volume level becomes lower. The worst performance results are obtained in the whisper situation with EER 46.93% in the GMM-UBM system and EER 25.88% in the i-Vector+LDA+WCCN system. Considering the speaking rate, slow rate utterances bring better performance with EER 16.93% compared to fast rate utterances with EER 19.92% in the i-Vector+LDA+WCCN system. In emotional conditions, the angry state with EER 19.41% presents more difficult situation than the happy state with EER 15.42% in the i-Vector+LDA+WCCN system when they are used for enrollment.

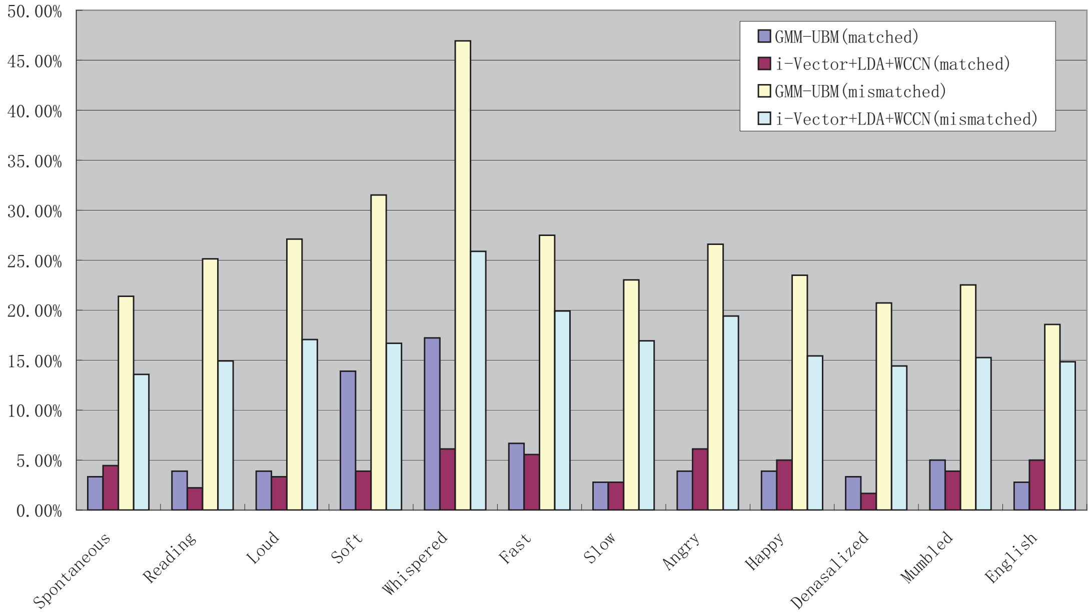

Figure 3: Comparison of EERs between the GMM-UBM baseline system and i-Vector+LDA+WCCN system in matched and mismatched conditions between enrollment and testing.

5.3. Effects of spontaneous and whisper conditions on speaker verification performances

In this paper, the base speech form is defined as the form of neutral spontaneous speech at normal rate and volume in Chinese, which is the most general way that people speak. We use the spontaneous utterances for enrollment and investigate the effects of the twelve variations by testing speaker verification systems in each variation condition. The experiments are accomplished in the GMM-UBM system and i-vector based systems including i-Vector+LDA, i-Vector+WCCN, i-Vector+NAP and i-Vector+LDA+WCCN. These speaker verification systems are tested with 10 target trials and 190 imposter trials for each testing condition. The training and testing procedure is repeated 10 times by choosing another enrollment utterance with the spontaneous condition for each subject.

As the whisper condition causes the most significant degradation in speaker verification performance, we intend to study the effects of whispering as an enrollment condition when testing condition changes in the twelve variation forms. The experiment process is the same with the above experiment except that we use whisper utterances for the enrollment phase.

The results of spontaneous and whisper utterances for enrollment are presented respectively in Table 4 and Table 5. The best results are formatted in bold for each row. It is obvious that speaker verification performances with spontaneous utterances for enrollment are much better than that with whisper condition. It is demonstrated that WCCN performs better than LDA and NAP for the compensation of intrinsic variations. Table 4 shows that the best result, with EER 1.67%, is obtained in the i-Vector+WCCN system when both enrollment and testing condition is spontaneous. The whisper testing data brings the worst performance with EER 24.00% of the i-Vector+LDA+WCCN system in the twelve testing conditions. The effects of speaking rate, speaking volume, and emotional state are more obvious than physical status and speaking language.

When whisper utterances are used for enrollment, the performances of speaker verification systems declines sharply. The only exception to that is when testing data uses whisper utterances, giving an EER of 5.00% in the i-Vector+WCCN system. Table 5 shows that the i-vector based systems perform much better than the GMM-UBM based system in the whisper conditions and the best results are obtained in the i-Vector+LDA+WCCN system for each mismatched form of testing data. Since the whisper condition is very different from the other variation forms in acoustic characteristics, feature com-

Table 4: EERs(%) for each testing condition when spontaneous utterances are used for enrollment.

Speech Variation	Variation Form	GMM-UBM	i-Vector+LDA	i-Vector+WCCN	i-Vector+NAP	i-Vector+LDA+WCCN
Base Case	Spontaneous	3.33	7.78	**1.67**	6.67	4.44
Speaking Style	Reading	13.00	14.00	**8.50**	12.50	11.00
Speaking Volume	Loud	27.00	14.00	14.50	14.00	**12.50**
	Soft	16.50	17.00	**10.00**	15.00	11.50
	Whispered	35.00	37.00	29.00	32.50	**24.00**
Speaking Rate	Fast	15.00	23.00	15.00	18.50	**14.00**
	Slow	28.00	15.50	16.00	16.00	**14.00**
Emotional State	Angry	17.50	24.00	**15.50**	20.50	**15.50**
	Happy	33.00	18.50	**11.50**	15.00	13.00
Physical Status	Denasalized	13.50	15.00	**8.00**	11.50	9.50
	Mumbled	22.00	17.00	14.00	15.00	**12.50**
Speaking Language	English	14.00	17.00	**10.50**	16.50	11.00

Table 5: EERs(%) for each testing condition when whisper utterances are used for enrollment.

Speech Variation	Variation Form	GMM-UBM	i-Vector+LDA	i-Vector+WCCN	i-Vector+NAP	i-Vector+LDA+WCCN
Base Case	Spontaneous	47.50	37.00	29.50	34.50	**23.00**
Speaking Style	Reading	46.50	37.50	31.00	34.50	**28.50**
Speaking Volume	Loud	45.00	35.00	27.00	33.50	**23.00**
	Soft	43.50	32.00	30.00	29.50	**27.50**
	Whispered	17.22	11.67	**5.00**	6.67	6.11
Speaking Rate	Fast	47.00	35.00	34.00	36.50	**28.50**
	Slow	45.00	35.00	36.50	32.50	**27.50**
Emotional State	Angry	47.50	31.00	36.00	34.50	**26.50**
	Happy	41.50	30.50	31.00	32.50	**24.00**
Physical Status	Denasalized	47.50	32.00	33.00	31.00	**24.00**
	Mumbled	45.50	30.00	30.50	28.50	**25.00**
Speaking Language	English	49.50	34.50	36.50	36.50	**26.50**

pensation techniques will be attempted before modeling the whisper variations to address the problems caused by whisper utterances.

6. Conclusions

This paper investigates the effects of intrinsic variation using i-vector framework from six aspects, namely speaking style, speaking rate, speaking volume, emotional state, physical status and speaking language. Performances of speaker verification systems decrease sharply due to mismatches between training and testing data in intrinsic variations. A variety of techniques are used to compensate for the intrinsic variations, including LDA, WCCN and NAP. Experimental results on the intrinsic variation corpus demonstrate that the i-vector based speaker verification systems perform better than the GMM-UBM based system and the best results are obtained using the i-vector framework with a combination of LDA and WCCN. Compared to the GMM-UBM baseline system, relative improvement in EER of around 36.76% is obtained in the i-Vector+LDA+WCCN system in the intrinsic variation corpus.

Spontaneous utterances used for enrollment bring better performances compared to other enrollment conditions when testing is done by using utterances with different intrinsic variations. The effects of speaking volume are the most significant in the six variation aspects. Speaker verification performances decline when the volume level becomes lower. Whisper testing data bring the largest degradation of speaker verification performances in the twelve variation forms when spontaneous utterances are used for enrollment. The speaker verification performances decrease sharply for each mismatched testing condition when the whisper recordings are used for enrollment. Although whisper represents the most difficult situations, the improvement in speaker verification performance is still very impressive in the i-Vector+LDA+WCCN system compared to the GMM-UBM based system.

Problems caused by intrinsic variation remain an important challenge for speaker verification systems. Our future work will attempt more techniques for the compensation of intrinsic variations in the i-vector framework for robust speaker verification. As whisper utterances cause the largest degradation in speaker verification performance, more research will be done to address the effects of whisper variation.

7. Acknowledgment

The work in this paper is partially supported by the National Natural Science Foundation of China (No. 61171116) and National Basic Research Program of China (973 Program, No. 2012CB316400).

8. References

[1] A. Solomonoff, W. Campbell, and I. Boardman, "Advances in channel compensation for SVM speaker recognition," in *Proc. IEEE ICASSP'05*, vol. 1. IEEE, 2005, pp. 629–632.

[2] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of interspeaker variability in speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 5, pp. 980–988, 2008.

[3] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Speaker and session variability in GMM-based speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1448–1460, 2007.

[4] N. Dehak, R. Dehak, P. Kenny, N. Brümmer, P. Ouellet, and P. Dumouchel, "Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification," in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[5] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Odyssey: The Speaker and Language Recognition Workshop*, 2001, pp. 213–218.

[6] B. Xiang, U. Chaudhari, J. Navratil, G. Ramaswamy, and R. Gopinath, "Short-time gaussianization for robust speaker verification," in *Proc. IEEE ICASSP'02*, vol. 1. IEEE, 2002, pp. 681–684.

[7] D. Reynolds, "Channel robust speaker verification via feature mapping," in *Proc. IEEE ICASSP'03*, vol. 2. IEEE, 2003, pp. 53–56.

[8] E. Shriberg, M. Graciarena, H. Bratt, A. Kathol, S. Kajarekar, H. Jameel, C. Richey, and F. Goodman, "Effects of vocal effort and speaking style on text-independent speaker verification," in *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[9] B. Ma, H. Meng, and M. Mak, "Effects of device mismatch, language mismatch and environmental mismatch on speaker verification," in *Proc. IEEE ICASSP'07*, vol. 4. IEEE, 2007, pp. 301–304.

[10] H. Bao, M. Xu, and T. Zheng, "Emotion attribute projection for speaker recognition on emotional speech," in *Eighth Annual Conference of the International Speech Communication Association*, 2007.

[11] M. Ghiurcau, C. Rusu, and J. Astola, "A study of the effect of emotional state upon text-independent speaker identification," in *Proc. IEEE ICASSP'11*. IEEE, 2011, pp. 4944–4947.

[12] M. Xu, L. Zhang, and L. Wang, "Score normalization-based speaking-style variation robust speaker recognition," in *Journal of Tsinghua University(Science and Technology)*, 2009, pp. 1278–1282.

[13] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[14] E. Shriberg, S. Kajarekar, and N. Scheffer, "Does session variability compensation in speaker recognition model intrinsic variation under mismatched conditions?" in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[15] M. Xu, L. Zhang, and L. Wang, "Database collection for study on speech variation robust speaker recognition," in *Proc of O-COCOSDA2008. Kyoto*, 2008.

[16] A. Hatch, S. Kajarekar, and A. Stolcke, "Within-class covariance normalization for SVM-based speaker recognition," in *Ninth International Conference on Spoken Language Processing*, 2006.

[17] W. Campbell, D. Sturim, D. Reynolds, and A. Solomonoff, "SVM based speaker verification using a GMM supervector kernel and NAP variability compensation," in *Proc. IEEE ICASSP'06*, vol. 1. IEEE, 2006, pp. 97–100.