A Principle Solution for Enroll-Test Mismatch in Speaker Recognition

Lantian Li, Dong Wang, Jiawen Kang, Renyu Wang, Jing Wu, Zhendong Gao, Xiao Chen

***Abstract*—Mismatch between enrollment and test conditions causes serious performance degradation on speaker recognition systems. This paper presents a statistics decomposition (SD) approach to solve this problem. This approach is based on the normalized likelihood (NL) scoring framework, and is theoretically optimal if the statistics on both the enrollment and test conditions are accurate. A comprehensive experimental study was conducted on three datasets with different types of mismatch: (1) physical channel mismatch, (2) speaking style mismatch, (3) near-far recording condition mismatch. The results demonstrated that the proposed SD approach is highly effective, and outperforms the ad-hoc multi-condition training approach that is commonly adopted but not optimal in theory.**

***Keywords*—Speaker Recognition; Deep Speaker Embedding; Condition Mismatch**

I. Introduction

Speaker Recognition has been emerging as an applicable technique after decades of research [1], [2], [3]. Traditional speaker recognition methods are based on statistical models, in particular the Gaussian mixture model-universal background model (GMM-UBM) [4] and the subspace alternatives including joint factor analysis (JFA) [5] and the i-vector model [6]. Recently, deep learning methods have demonstrated significant progress with regard to speaker recognition [7]. Early research learns frame-level features and derive utterance-level representations by simple average [8], [9], and the following research focuses on deriving utterance-level representations directly [10], [11]. The utterance-level representations are often called **speaker vectors**, and the process that maps a variant-length speech signal to a fixed-dimension speaker vector is often called **speaker embedding**. As far, the most popular deep learning architecture for speaker embedding is the x-vector model, proposed by Snyder et al. [10]. Compared to an alternative deep learning approach that employs an end-to-end architecture and training scheme [12], [13], [14], the embedding approach is easier to train [15] and the derived speaker vectors support various speaker-related tasks. Recent progress on the deep speaker embedding approach includes more comprehensive architecture [16], [17], improved pooling methods [11], [18], [19], [20], better training criteria [21], [22], [23], [24], [25], [26], and better training schemes [27], [28], [29].

In spite of the tremendous progress, the present speaker recognition performance yet cannot meet the request of real applications. A particular problem is that the performance of a model trained with data from one domain will be substantially reduced when applied to another domain. Numerous studies have been carried on to deal with the domain-mismatch problem [30], [31], [32], [33], [34], [35], [36], [37], [38]. Most of the research focuses on domain drift, and the central idea is to build statistical models suitable for the target domain, by using the knowledge of data or model of the source domain. These techniques are suitable for solving the mismatch between the training data and the deployment environment. For example, they can adapt a model trained with a large reading speech database so as to make it applicable in a new scenario of online meetings.

In this paper, we focus on another mismatch issue that seems more serious but rarely studied: the mismatch between the enrollment and test conditions. During enrollment, the condition is often controlled: people tend to speak in the same acoustic environment and in the same style, and the recording device does not change. During test, the condition tends to be much more uncontrollable and unpredictable. It is often significantly different from the enrollment condition, and varies from one test to another. Significant performance reduction is often observed with this mismatch [39], [40], [41], [42]. Some typical scenarios that involve serious enrollment-test mismatch are:

- Cross-Channel test, where the enrollment uses one device and the test uses another device. This leads to clear mismatch on the recording devices.
- Time-Variance test, where the test is conducted in a few weeks or months after the enrollment. This leads to significant mismatch on the speaking styles.
- Near-Far field test, where the enrollment uses a near-filed microphone while the test uses a far-field microphone, or vice versa. This leads to clear mismatch on the recording conditions.

More complex situations may involve more multiple mismatches. For example, in the multi-genre scenario, the speaker may enroll in reading speech with a near electret microphone, while the test could be in chatting or singing, with a far-field dynamic microphone.

The essential problem associated with this enrollment-test mismatch is that the statistical properties (e.g., between-

This work was supported by the National Natural Science Foundation of China (NSFC) under the project No.61633013, and also the Huawei Innovation Research Program under Cross-device Speaker Recognition Project Contract No.YBN2019125091. Dong Wang is the corresponding author.

L. Li, D. Wang and J. Kang are with the Center for Speech and Language Technologies (CSLT), BNRist at Tsinghua University, Beijing 100084, China (E-mail: {lilt,kangjw}@cslt.org, wangdong99@mails.tsinghua.edu.cn).

R. Wang, X. Chen are with the Noah's Ark Lab, Huawei, at Hong Kong Science Park, Pak Shek kok, Shatin, New Territories, Hong Kong (E-mail: {wangrenyu1,chen.xiao2}@huawei.com).

J. Wu, Z. Gao are with Huawei Technologies CO. Limited, Pudong, Shanghai, China (E-mail: wujing14@huawei.com, gaozhendong@hisilicon.com).

speaker variance and within-speaker variance) of the enrollment data and the test data are different, which we call **statistical incoherence**. This incoherence fails most of the scoring models, e.g., the famous probabilistic linear discriminant (PLDA) [43], [44], for which coherent statistics between enrollment and test are prerequisite.

We highlight that the incoherence on statistics is unique for the enrollment-test mismatch and does not exist in the domain-mismatch scenario. In that scenario, although the model does not fully match the data in the target domain, the statistics for enrollment and test are coherent, and so if sufficient data are collected from the target domain, the mismatch between the model and the data will be eventually resolved. In the enrollment-test mismatch scenario, however, the statistical incoherence exits no matter how many data are collected from either the enrollment condition or the test condition. Therefore, any solution that does not address this incoherence is theoretically not optimal. For example, the commonly adopted multi-condition training (MCT) approach that pools data from both enrollment and test condition for PLDA training [42]. By this approach, the resultant PLDA is just an empirical compromise between the enrollment condition and the test condition, but is not optimal for either of them.

This paper will present a principle solution for the enrollment-test mismatch problem. The solution is based on the normalized likelihood (NL) scoring framework we proposed recently [45]. Although equivalent to PLDA when the statistics are coherent, the NL formulation allows decoupling the entire scoring process into three phase: enrollment, prediction, and normalization, and each phase can employ its own statistical model. We call this approach **statistics decomposition (SD)**. It is a principle solution for the enrollment-test mismatch problem, as the resultant NL score is optimal if the statistics on the enrollment and test conditions are both accurate, even if they are not coherent.

The rest of the paper is organized as follows. Section II will present a theoretical background for the normalized likelihood (NL) scoring framework. Section III presents the SD approach for the enrollment-test mismatch problem based on a simple linear condition transform. Experiments are reported in Section V, and finally the paper is concluded in Section VI.

II. Theory background

This section revisits the NL scoring framework presented in [45]. For simplicity, we will only present the case where the between-speaker and within-speaker are all isotropic Gaussians. The derivation can be easily extended to the case with Gaussians in general forms, as described in [45].

A. Preliminary

We shall assume a simple linear Gaussian as follows:

$$p(\boldsymbol{\mu}) = N(\boldsymbol{\mu}; \mathbf{0}, \epsilon\mathbf{I}) \tag{1}$$

$$p(\boldsymbol{x}|\boldsymbol{\mu}) = N(\boldsymbol{x}; \boldsymbol{\mu}, \sigma\mathbf{I}). \tag{2}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{x}$ represent speaker means and speaker vectors, respectively; ϵ and σ are the between-speaker variance and the within-speaker variance, respectively.

With this model, it is easy to derive the marginal probability $p(\boldsymbol{x})$ and the posterior probability $p(\boldsymbol{\mu}|\boldsymbol{x})$ as follows:

$$p(\boldsymbol{x}) = N(\boldsymbol{x}; \mathbf{0}, (\epsilon + \sigma)\mathbf{I}) \tag{3}$$

$$p(\boldsymbol{\mu}|\boldsymbol{x}) = N(\boldsymbol{\mu}; \frac{\epsilon}{\epsilon + \sigma}\boldsymbol{x}, \frac{\epsilon\sigma}{\epsilon + \sigma}\mathbf{I}) \tag{4}$$

If the observations are more than one, the posterior probability is:

$$p(\boldsymbol{\mu}|\boldsymbol{x}_1, ..., \boldsymbol{x}_n) = N(\boldsymbol{\mu}; \frac{n\epsilon}{n\epsilon + \sigma}\bar{\boldsymbol{x}}, \frac{\epsilon\sigma}{n\epsilon + \sigma}\mathbf{I}), \tag{5}$$

where $\bar{\boldsymbol{x}}$ is the average of the observations $\{\boldsymbol{x}_1, ..., \boldsymbol{x}_n\}$.

B. Normalized likelihood

The task of speaker recognition is to test the following two outcomes regarding to a speaker vector $\boldsymbol{x}$ and check which one is more probable: $\{$ H_0: $\boldsymbol{x}$ belongs to class k; H_1: $\boldsymbol{x}$ belongs to any class other than k $\}$. According to the decision theory, if the risk for each choice is equal, the decisions incurring minimum Bayes risk (MBR) should be based on the posterior probabilities of the choices. Assuming an equal prior, the posterior for H_0 is:

$$p(H_0|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|H_0)}{p(\boldsymbol{x}|H_0) + p(\boldsymbol{x}|H_1)}.$$

Note that $p(\boldsymbol{x}|H_0)$ is the likelihood of $\boldsymbol{x}$ generated by class k, denoted by:

$$p_k(\boldsymbol{x}) = N(\boldsymbol{x}; \boldsymbol{\mu}_k, \sigma\mathbf{I}), \tag{6}$$

and $p(\boldsymbol{x}|H_1)$ summarizes the likelihood that $\boldsymbol{x}$ is generated by all possible classes, which is exactly the marginal distribution, or evidence, $p(\boldsymbol{x})$. By setting an appropriate threshold on the posterior, one can deal with different priors on H_0 and H_1 and different trade-off between the risk on them.

Note that what matters in the posterior is the ratio $p(\boldsymbol{x}|H_0)/p(\boldsymbol{x}|H_1)$, which is essentially the likelihood $p_k(\boldsymbol{x})$ normalized by the evidence $p(\boldsymbol{x})$. We therefore define the normalized likelihood (NL) as:

$$NL(\boldsymbol{x}|k) = \frac{p(\boldsymbol{x}|H_0)}{p(\boldsymbol{x}|H_1)} = \frac{p_k(\boldsymbol{x})}{p(\boldsymbol{x})}. \tag{7}$$

With the linear Gaussian model formulated in the previous section, $p_k(\boldsymbol{x})$ and $p(\boldsymbol{x})$ can be computed according to Eq. (6) and Eq. (3) respectively.

In practice, the class mean $\boldsymbol{\mu}_k$ has to be estimated from the enrollment utterances. Denote the enrollment samples by $\{\boldsymbol{x}_1^k, ..., \boldsymbol{x}_n^k\}$ and their average by $\bar{\boldsymbol{x}}_k$, the posterior probability for the mean $\boldsymbol{\mu}_k$ can be derived as follows according to Eq. (5):

$$p(\boldsymbol{\mu}_k|\boldsymbol{x}_1^k, ..., \boldsymbol{x}_n^k) = N(\boldsymbol{\mu}_k; \frac{n_k\epsilon}{n_k\epsilon + \sigma}\bar{\boldsymbol{x}}_k, \frac{\epsilon\sigma}{n_k\epsilon + \sigma}\mathbf{I}). \tag{8}$$

The likelihood $p_k(\boldsymbol{x})$ can therefore be computed by marginalizing over $\boldsymbol{\mu}_k$, according to Eq. (8). Following Eq. (3), we have:

$$\begin{aligned} p_k(\boldsymbol{x}) &= p(\boldsymbol{x}|\boldsymbol{x}_1^k, ..., \boldsymbol{x}_n^k) \\ &= \int p(\boldsymbol{x}|\boldsymbol{\mu}_k) p(\boldsymbol{\mu}_k|\boldsymbol{x}_1^k, ..., \boldsymbol{x}_n^k) \mathrm{d}\boldsymbol{\mu}_\mathrm{k} \\ &= N(\boldsymbol{x}; \frac{n_k\epsilon}{n_k\epsilon+\sigma}\bar{\boldsymbol{x}}_k, (\sigma + \frac{\epsilon\sigma}{n_k\epsilon+\sigma})\mathbf{I}) \end{aligned} \tag{9}$$

A simple computation shows that:

$$\begin{aligned} \log NL(\boldsymbol{x}|k) &= \log p_k(\boldsymbol{x}) - \log p(\boldsymbol{x}) \\ &\propto -\frac{1}{\sigma + \frac{\epsilon\sigma}{n_k\epsilon+\sigma}} ||\boldsymbol{x} - \tilde{\boldsymbol{\mu}}_k||^2 + \frac{1}{\epsilon+\sigma}||\boldsymbol{x}||^2 \end{aligned} \tag{10}$$

where we have defined:

$$\tilde{\boldsymbol{\mu}}_k = \frac{n_k\epsilon}{n_k\epsilon+\sigma}\bar{\boldsymbol{x}}_k.$$

We highlight that the NL formulation of Eq.(7) is a general form, and both the likelihood $p_k(\boldsymbol{x})$ and the normalizer $p(\boldsymbol{x})$ can be based on any model. The above derivation based on the linear Gaussian model is a simple special case and will be used in this paper. In fact, this linear Gaussian NL is equivalent to the PLDA likelihood ratio, evidenced by the following equation:

$$\frac{p(\boldsymbol{x}|\boldsymbol{x}_1, ..., \boldsymbol{x}_n)}{p(\boldsymbol{x})} = \frac{p(\boldsymbol{x}, \boldsymbol{x}_1, ..., \boldsymbol{x}_n)}{p(\boldsymbol{x})p(\boldsymbol{x}_1, ..., \boldsymbol{x}_n)},$$

where the LHS is the NL form and the RHS is the PLDA form.

Note that the NL form under the linear Gaussian assumption has been proposed and discussed by several authors, e.g., Kenny et al. [46], Borgström et al. [47] and McCree et al. [48]. All these authors derived the NL form from the PLDA likelihood ration (LR), and therefore the equivalence between the NL and PLDA is a natural result. Our derivation for NL starts from the MBR decision theory, which explicitly affirms the optimum of the NL score.

III. Methods

PLDA treats enrollment and test as a one-step process, as evidenced by the joint likelihood $p(\boldsymbol{x}, \boldsymbol{x}_1, ..., \boldsymbol{x}_n)$. This leads to a fundamental problem when dealing with enrollment-test mismatch, where the enrollment and test cannot be modeled by a single statistical model. In contrast, the NL formulation can decouple the scoring process into several phases and each phase can use its own statistical model. This offers a principle solution for the enrollment-test mismatch problem, as we will show in this section.

A. General form

Under the linear Gaussian assumption, the statistical model of each condition is represented by the between-speaker and within-speaker covariances. We suppose the model can be well estimated for both the enrollment and the test conditions. Assume a linear discriminant analysis (LDA) transform on each condition, the covariances of both conditions will be transformed to be diagonal. Note that this transform does not change the NL score [45]. For a clear presentation, we will assume these covariances are isotropic and are denoted by $\{\epsilon\mathbf{I}, \sigma\mathbf{I}\}$ for the enrollment condition and $\{\hat{\epsilon}\mathbf{I}, \hat{\sigma}\mathbf{I}\}$ for the test condition, respectively. The derivation for the general non-isotropic forms can be derived in exactly the same way, by appropriate variable expansion.

Rewrite the NL formulation as below:

$$\log NL(\boldsymbol{x}) = \log \int p(\boldsymbol{\mu}|\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n) p(\boldsymbol{x}|\boldsymbol{\mu}) \mathrm{d}\boldsymbol{\mu} - \log \mathrm{p}(\boldsymbol{x}). \tag{11}$$

It shows that the scoring process can be separated to three phases:

- **Enrollment**: Compute the posterior $p(\boldsymbol{\mu}|\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n)$.
- **Prediction**: Compute the likelihood $p_k(\boldsymbol{x}) = \int p(\boldsymbol{\mu}|\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n) p(\boldsymbol{x}|\boldsymbol{\mu}) \mathrm{d}\boldsymbol{\mu}$.
- **Normalization**: Compute the normalization term $p(\boldsymbol{x})$ and use it normalize the likelihood by $\log p_k(\boldsymbol{x}) - \log p(\boldsymbol{x})$.

Usually these three phases use the same statistical model, under the assumption that the statistics of the enrollment and test conditions are coherent. However, with enrollment-test condition mismatch, the statistics of the enrollment and test conditions are incoherent. In this case, using the statistical model of neither condition is optimal for all these three phases. A principle solution is to use the respective statistical model for each phase, which we call **statistics decomposition (SD)**.

We will propose a simple and effective SD implementation based on the linear Gaussian assumption. Firstly, notice that for the enrollment phase, the model of the enrollment condition $\{\epsilon\mathbf{I}, \sigma\mathbf{I}\}$ should be used, and for the normalization phase, the model of the test condition $\{\hat{\epsilon}\mathbf{I}, \hat{\sigma}\mathbf{I}\}$ should be used. The prediction phase is more complex. According to the NL formulation, it marginalizes the conditional probability $p(\boldsymbol{x}|\boldsymbol{\mu})$ with respect to the posterior $p(\boldsymbol{\mu}|\boldsymbol{x}_1, ..., \boldsymbol{x}_n)$. However, $p(\boldsymbol{\mu}|\boldsymbol{x}_1, ..., \boldsymbol{x}_n)$ is computed based on the statistical model of the enrollment condition, and so is not a correct posterior for marginalizing the conditional probability $p(\boldsymbol{x}|\boldsymbol{\mu})$ where $\boldsymbol{\mu}$ and $\boldsymbol{x}$ are quantities in the test condition.

In order to solve this problem, a connection must be established between the quantities in the enrollment condition and the test condition. A possible solution is to design a mapping function that transforms $\boldsymbol{\mu}$ from the enrollment condition to the test condition, or vice versa, transforms $\boldsymbol{x}$ from the test condition to the enrollment condition. In this study, we will choose the transform on $\boldsymbol{x}$, as it results in a relatively simpler form. For simplicity, we shall assume that the transform here is a linear:

$$\boldsymbol{x} = \mathrm{M}\hat{\boldsymbol{x}} + \boldsymbol{b}, \tag{12}$$

where $\hat{\boldsymbol{x}}$ represents the observation in the test condition, and $\boldsymbol{x}$ is the transformed data in the enrollment condition. If we assume that the transformed data can be well represented by the statistical model of the enrollment condition, the NL score can be derived.

Firstly, according to Eq. (8), the posterior $p(\boldsymbol{\mu}_k|\boldsymbol{x}_1^k, ..., \boldsymbol{x}_n^k)$ is computed using the model of the enrollment condition:

$$p(\boldsymbol{\mu}_k|\boldsymbol{x}_1^k, ..., \boldsymbol{x}_n^k) = N(\boldsymbol{\mu}_k; \frac{n_k\epsilon}{n_k\epsilon+\sigma}\bar{\boldsymbol{x}}_k, \frac{\epsilon\sigma}{n_k\epsilon+\sigma}\mathbf{I}).$$

Secondly, transform the test sample $\hat{\boldsymbol{x}}$ by the linear transform, and perform the prediction using the model of the enrollment condition:

$$\begin{aligned} p_k(\hat{\boldsymbol{x}}; \mathbf{M}, \boldsymbol{b}) &= p(\mathbf{M}\hat{\boldsymbol{x}} + \boldsymbol{b}|\boldsymbol{x}_1^k, ..., \boldsymbol{x}_n^k) \\ &= \int p(\mathbf{M}\hat{\boldsymbol{x}} + \boldsymbol{b}|\boldsymbol{\mu}_k)p(\boldsymbol{\mu}_k|\boldsymbol{x}_1^k, ..., \boldsymbol{x}_n^k)\mathrm{d}\boldsymbol{\mu}_\mathrm{k} \\ &= N(\mathbf{M}\hat{\boldsymbol{x}} + \boldsymbol{b}; \frac{n_k\epsilon}{n_k\epsilon+\sigma}\bar{\boldsymbol{x}}_k, (\sigma + \frac{\epsilon\sigma}{n_k\epsilon+\sigma})\mathbf{I}) \end{aligned} \quad (13)$$

According to Eq. (3), the normalization term $p(\hat{\boldsymbol{x}})$ is computed based the model of the test condition:

$$p(\hat{\boldsymbol{x}}) = N(\hat{\boldsymbol{x}}; \mathbf{0}, (\hat{\epsilon} + \hat{\sigma})\mathbf{I}). \quad (14)$$

The NL score is then computed as follows, by using the results of Eq. (13) and Eq. (14):

$$\log NL(\hat{\boldsymbol{x}}|k) \propto -\frac{1}{\sigma + \frac{\epsilon\sigma}{n_k\epsilon+\sigma}}||\mathbf{M}\hat{\boldsymbol{x}} + \boldsymbol{b} - \tilde{\boldsymbol{\mu}}_k||^2 + \frac{1}{\hat{\epsilon}+\hat{\sigma}}||\hat{\boldsymbol{x}}||^2, \quad (15)$$

where we have defined:

$$\tilde{\boldsymbol{\mu}}_k = \frac{n_k\epsilon}{n_k\epsilon+\sigma}\boldsymbol{x}_k. \quad (16)$$

The optimal parameters $\{\mathbf{M}, \boldsymbol{b}\}$ for the linear transform can be estimated by maximum likelihood (ML) training [49], which maximizes $p_k(\hat{\boldsymbol{x}}; \mathbf{M}, \mathbf{b})$ with respect to $\mathbf{M}$ and $\mathbf{b}$, using samples in the test condition and the samples of the same speaker in the enrollment condition. It is essentially an maximum likelihood linear regression (MLLR) task [50], and the objective function for the optimization can be written by:

$$\mathcal{L}(\mathbf{M}, \boldsymbol{b}) = \sum_{k=1}^{K}\sum_{i=1}^{N} p_k(\hat{\boldsymbol{x}}_i; \mathbf{M}, \boldsymbol{b}) \quad (17)$$

where K denotes the number of speakers, and N denotes the number of test samples in each speaker.

Note that with the ML training, we essentially minimize the KL divergence between the within-speaker distribution of the enrollment condition and the one of the test condition after the linear transform. If the two within-speaker distributions are perfectly matched with the linear transform in Eq.(12), the NL score computed as above will be MBR optimal. This approach will be called **statistics decomposition with linear transform (SD/LT)** in the paper.

We highlight that if the condition mismatch is very complex, a more complicated transform is required. In any case, if the within-speaker distributions in the two conditions can be perfectly matched with the transform, the NL score will be MBR optimal. We therefore obtain a principle solution for the enrollment-test mismatch problem. As a preliminary study, we focus on SD/LT approach in this paper and leave more complicated transforms (e.g., those based on deep neural nets) for future work.

B. Special case 1: global shift compensation

A special case of the enrollment-test mismatch is that the behavior of the speakers do not change much (i.e. assume the between- and within-speaker variances do not change between the enrollment and test conditions), but the speaker vectors are globally shifted by a constant vector $\boldsymbol{b}$. This often happens in channel mismatch, where a displacement is often observed between the speaker vectors recorded from devices with different physical characteristics.

In this special case, we assume the between-speaker and within-speaker covariance does not change, therefore a simple global shift $\boldsymbol{b}$ on the speaker vectors of the test condition is sufficient to obtain the optimal NL score. Formally, the SD/LT is simplified to be:

$$\boldsymbol{x} = \hat{\boldsymbol{x}} + \boldsymbol{b},$$

and the NL is derived by:

$$\log NL(\hat{\boldsymbol{x}}|k) \propto -\frac{1}{\sigma + \frac{\epsilon\sigma}{n_k\epsilon+\sigma}}||\hat{\boldsymbol{x}} + \boldsymbol{b} - \tilde{\boldsymbol{\mu}}_k||^2 + \frac{1}{\epsilon+\sigma}||\hat{\boldsymbol{x}} + \boldsymbol{b}||^2, \quad (18)$$

where $\tilde{\boldsymbol{\mu}}_k$ is defined by Eq.(16).

C. Special case 2: within-speaker variance adaptation

As the second special case, considering that people tend to speak with a large variation during test than during enrollment. This often happens when a registered person performs time-variance test (e.g., several days or weeks later) and the test is performed multiple times. In this case, the speaking style at each test tends to be different, thus leading to a large within-speaker variation.

For this special case, we will assume that the within-speaker variance is the only changed statistical quantity[1]. With this assumption, it is easy to compute the NL score for a test sample $\boldsymbol{x}$ by using σ in the enrollment phase, and $\hat{\sigma}$ in the prediction phase and normalization phase:

$$\log NL(\boldsymbol{x}|k) \propto -\frac{1}{\hat{\sigma} + \frac{\epsilon\sigma}{n_k\epsilon+\sigma}}||\boldsymbol{x} - \tilde{\boldsymbol{\mu}}_k||^2 + \frac{1}{\epsilon+\hat{\sigma}}||\boldsymbol{x}||^2, \quad (19)$$

where $\tilde{\boldsymbol{\mu}}_k$ is defined by Eq.(16).

[1]One may argue that the speaking style at each test time tends to be the same, and so the variation of each time is not large. This can therefore be treated as a global shift as in recording device mismatch. However this is not applicable because the speaking style at each test time cannot be determined, and what we can obtain is only the long-term change on the within-speaker variation.

IV. Related work

Research dedicated to enrollment-test condition mismatch is rare. In practice, people often rely on algorithms developed for addressing training-deployment domain mismatch. These algorithms can be categorized to *data theme* [31], [30], *model theme* [51], [52], [53], [40] and *scoring theme* [35], [36], [37], [38], [32], [33], [34], [54], working on data/features, embedding models and scoring models respectively. For each theme, the basic idea is either normalization or adaptation: the former aims to make the data or model domain-independent [31], [51], [52], [53], [35], [36], [37], [38], while the latter aims to make them more suitable for the target domain [30], [40], [32], [33], [34], [54]. The normalization methods can be readily applied to alleviate the enrollment-test mismatch problem, though the adaptation methods can not as the enrollment and test are in different conditions.

The multi-condition training (MCT) [42] can be regarded as a special normalization approach, belonging to the scoring theme. It pools the data from both enrollment and test conditions and trains a multi-conditional PLDA. Intuitively, it emphasizes on the most speaker-relevant variation and attenuates other speaker-irrelevant variations, and so can alleviate impact from condition mismatch. This approach works well in many situations, however it does not offer an optimal score, as the statistical model it is based on is neither accurate for the enrollment nor for the test.

Our approach (SD/LT as the first and simplest case) belongs to the scoring theme, but it is fundamentally different from the normalization methods that pursue a condition-insensitive model as in IDVC [35] or MCT [42]. Instead, it admits the discrepancy between enrollment and test, and models the statistics of speaker vectors in the two conditions. If the two models are accurate, the NL formulation ensures that the resultant scores are MBR optimal.

V. Experiments

We will conduct experiments on three datasets to verify the proposed SD/LT scoring approach. The three tests reflect different types of enrollment-test mismatch, as detailed below:

- **Cross-Channel test**: Conducted on the AIShell-1 dataset, where the enrollment and test data are recorded at the same time but with different recording devices. It tests the performance with physical channel mismatch.
- **Time-Variance test**: Conducted on the CSLT-Chronos dataset, where the test data are recorded several times, and weeks or months later than the enrollment data. The data are recorded with the same device and in the same acoustic environment. It tests the performance with long-term speaking style variation.
- **Near-Far test**: Conducted on the HI-MIA dataset, where the enrollment and test data are recorded at the same time, by devices of the same brand, but with different recording distances. It tests the performance with recording condition mismatch.

A. Data

Four datasets were used in our experiments: VoxCeleb [16], [55], AIShell-1 [56], CSLT-Chronos [57], and HI-MIA [58]. VoxCeleb was used to train the speaker embedding model, which is the x-vector model in our experiment. The other three datasets were used for performance evaluation under different enrollment-test mismatch conditions. More information about these datasets is presented below.

VoxCeleb[2]: This is a large-scale audiovisual speaker dataset collected by the University of Oxford [16]. The entire database consists of *VoxCeleb1* and *VoxCeleb2*. All the speech signals were collected from open-source media channels and therefore involve rich variations in channel, style, and ambient noise. This dataset was used to train the x-vector embedding model. The entire dataset contains $2,000$+ hours of speech signals from $7,000$+ speakers. Data augmentation was applied to improve robustness, with the MUSAN corpus [59] used to generate noisy utterances, and the room impulse responses (RIRS) corpus [60] was used to generate reverberant utterances.

AIShell-1[3]: This is an open-source multi-channel Chinese Mandarin speech dataset published by AISHELL [56]. All the speech utterances are recorded in parallel via three categories of devices, including high fidelity Microphones (Mic), Android phones (AND) and Apple iPhones (iOS). This dataset is used for the **Cross-Channel test** in our experiment. The entire dataset consists of two parts: *AIShell-1.Train*, which covers 3 devices and involves $360,897$ utterances from 340 speakers, was used to implement both commonly used domain adaptation methods and our proposed NL-based scoring methods. *AIShell-1.Eval*, which also covers 3 devices and involves $64,495$ utterances from 60 speakers, was used for performance evaluation on enrollment-test mismatch conditions.

CSLT-Chronos[4]: This is a time-varying speaker recognition dataset collected by Tsinghua University [57]. For commercial usage, only the x-vectors are published for research usage. There are 8 recording sessions, which have the same acoustic environment (speakers, prompt texts and recording devices), are selected to evaluate the time-varying effect. The first 4 sessions are recorded in an interval of approximately one week, and the following 4 sessions are recorded in an interval of approximately one month. This dataset is used for the **Time-Variance test** in our experiment. In each session, it involves 57 speakers, and each speaker has 150 utterances. For each speaker, the first 20 utterances are selected for enrollment, and the rest are used for test.

HI-MIA[5]: This is an open-source far-field text-dependent speaker recognition dataset published by AISHELL [58]. This dataset is used for AISHELL Speaker Verification Challenge 2019[6]. All the speech utterances are recorded with the wake-up words 'Hi-MIA' from 1m, 3m and 5m fields in parallel. This dataset is used for the **Near-Far test** in our experiment. Specially, the recordings in the normal speaking speed from

[2]http://www.robots.ox.ac.uk/ vgg/data/voxceleb/
[3]http://openslr.org/33/
[4]http://data.cslt.org/
[5]http://openslr.org/85/
[6]http://challenge.aishelltech.com/

the clean environments are selected in our experiments. Finally, we build two subsets: *HI-MIA.Train*, which covers 3 fields and involves $15,186$ utterances from 254 speakers, was used to implement different adaptation and scoring methods. *HI-MIA.Eval*, which also covers 3 fields and involves $5,048$ utterances from 86 speakers, was used for performance evaluation.

B. Embedding models

We built the state-of-the-art x-vector embedding model based on the TDNN architecture [10]. This x-vector model was created using the Kaldi toolkit [61], following the VoxCeleb recipe. The acoustic features are 40-dimensional Fbanks. The main architecture contains three components. The first component is the feature-learning component, which involves 5 time-delay (TD) layers to learn frame-level speaker features. The slicing parameters for these 5 TD layers are: $\{t\text{-}2,\ t\text{-}1,\ t,\ t\text{+}1,\ t\text{+}2\}$, $\{t\text{-}2,\ t,\ t\text{+}2\}$, $\{t\text{-}3,\ t,\ t\text{+}3\}$, $\{t\}$, $\{t\}$. The second component is the statistical pooling component, which computes the mean and standard deviation of the frame-level features from a speech segment. The final one is the speaker-classification component, which discriminates between different speakers. This component has 2 full-connection (FC) layers and the size of its output is $7,185$, corresponding to the number of speakers in the training set. Once trained, the 512-dimensional activations of the penultimate FC layer are read out as an x-vector.

C. Qualitative analysis

In this section, we investigate the statistics incoherence involved in our three tests. The t-SNE [62] program is employed to project the x-vectors to a 2-dimensional space, by which we can analyze the statistical properties of the within-speaker and between-speaker distributions in the enrollment and test conditions, especially the change of those properties caused by each type of condition mismatch. The t-SNE plots are shown in Figure 1, and detailed analysis is presented below.

1) Cross-Channel test: For the cross-channel test, we select x-vectors of 3 speakers recorded with 3 recording devices, and plot them in Figure 1(a). In this figure, each shape represents a particular speaker, and each color represents a particular recording device (blue for AND, green for iOS, red for Mic). We firstly observe that different speakers are clearly projected to different clusters and are well separated from each other. Based on that, x-vectors that belong to the same speaker but different devices form a small sub-cluster. From this picture, it seems that the between-speaker and within-speaker variances are not substantially different for different devices. The only statistics incoherence caused by enrolling within one device and testing with another is the small shift on the mean vectors of the within-class distribution. Interestingly, the shift between AND and iOS is much smaller than that between Mic and AND/iOS, which is intuitively reasonable as AND/iOS are both hand-set devices. If this shift is globally equal for each speaker (which cannot be justified from the picture due to the nonlinearity of t-SNE), we will meet the the assumption of special case 1.

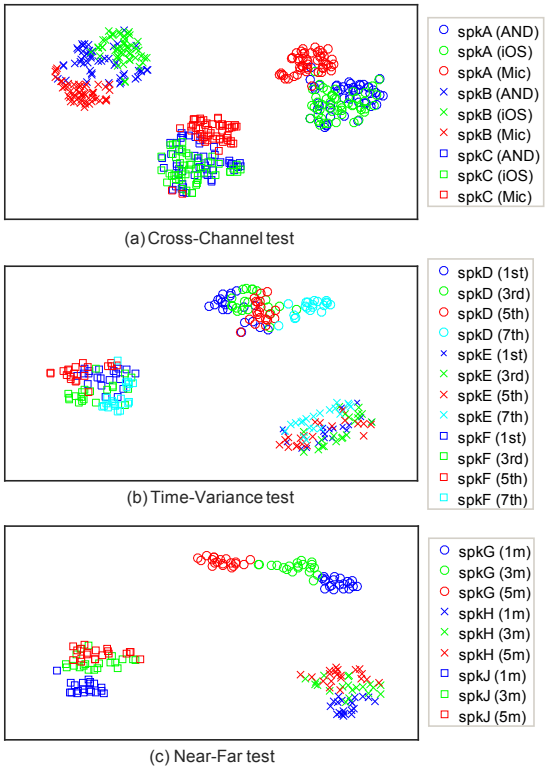

Fig. 1: The distributions of x-vectors in three test scenarios. Each shape represents one speaker and each color represent a condition, which is device in the Cross-Channel test (a), recording session in the Time-Variance test (b), and recording distance in the Near-Far test (c).

2) Time-Variance test: For the time-variance test, we select x-vectors from 3 speakers recorded in 4 sessions, and plot them in Figure 1(b). In this figure, each shape represents one speaker, and each color represents one recording session. It can be seen that the differentiation among speakers is not seriously influenced by the session variation, which we attribute to the different speaking styles of different sessions. Next, the speaking styles of different sessions may be clearly different (e.g., the two speakers on the top-left) or may be roughly the same (e.g., the speaker on the bottom-right). In general, the within-speaker variance tends to be increased gradually with more recording sessions pooled, though the between-speaker variance and the mean vectors of each speaker do not change much. These statistical properties indicate that this test scenario corresponds to the special case 2.

3) Near-Far test: For near-far scenarios, x-vectors from 3 speakers with 3 recording fields are presented in Figure 1(c). In this figure, each shape represents one speaker, and each color represents one recording distance (blue for 1m, green for 3m, red for 5m). Firstly, we observe that with the variation on the recording condition, the x-vectors are still discriminative for speakers. Secondly, the within-speaker variance is clearly different with different recording conditions, and the distribu-

tion with the near-filed condition (1m) looks more compact than the far-field conditions (3m and 5m). Finally, there is a clear shift between the distributions with any two different recording conditions, and the shift is more evident between the near-field condition and the two far-field conditions compared with the shift between the two far-field conditions. All the observations indicate that the Near-Far test involves more complex mismatch compared to the other two tests, and cannot be modeled by the two special cases.

D. Quantitative analysis

In this section, we conduct a quantitative analysis for the statistics incoherence problem associated with the three enrollment-test mismatch tests. Specifically, we will evaluate some statistical quantities for data in different conditions and analyze the differences on the quantities from one condition to another. The tested quantities are explained as follows:

- Between-speaker and within-speaker variances. Since the covariance matrices (either between-speaker or within-speaker) of different conditions may be in different directions, we test the variance in the principle directions of each condition. Specifically, we select one condition and estimate an LDA transform, and then project x-vectors of the test speech in all the conditions (including the enrollment condition itself) to the LDA space. Finally, we plot the between-speaker and within-speaker variances on dimensions that correspond to the 50 leading directions of the LDA transform.
- Angle metric and length metric to measure the global shift. The global mean vectors of x-vectors in the enrollment condition $\boldsymbol{x}_e$ and the test condition $\boldsymbol{x}_t$ are firstly computed, and then the cosine distance and Euclidean distance between them are computed as the angle metric and length metric to measure the global shift. The angle metric is formulated by:

$$\theta = (1 - \frac{\boldsymbol{x}_e^T \boldsymbol{x}_t}{||\boldsymbol{x}_e|| \; ||\boldsymbol{x}_t||}) * 10^3, \tag{20}$$

and the length metric is formulated by:

$$\ell = ||\boldsymbol{x}_e - \boldsymbol{x}_t||^2 * 10^2. \tag{21}$$

1) Cross-Channel test: We use *AIShell-1.Train* to compute the statistics. The dataset involves 340 speakers recorded with 3 recording devices at the same time.

The between-speaker and within-speaker variances of the test data are reported in Figure 2, where each enrollment condition is presented separately. The observations are: (1) The pattern of the between-speaker variances are not significantly different for different enrollment scenarios, and for each enrollment scenario, the between-class variances for test in different conditions are also similar. (2) For each of the enrollment scenario, the pattern of within-speaker variances is different for test data different conditions, but the difference is rather marginal. The within-class variances are in the range of 0.8 to 1.2 with enrollment-test mismatch, which is close to 1.0, the value with matched enrollment and test conditions (e.g., Mic-Mic)[7].

The angle metric and length metric that measure the global shift are reported in Table I, where we report the results with all pairs of mismatched conditions. Note that for global shift the mismatch is symmetric for enrollment and test. It can be observed that the global shift is evident in all the mismatch scenarios, and the shift between AND and iOS is much smaller than that between Mic and AND/iOS.

All the results coincide the observations in the qualitative analysis, and suggest that in the Cross-Channel scenario, the enrollment-test mismatch is mainly reflected in a global shift. Therefore, it can be compensated for by a simple shifting back, as formulated by the first special case of the SD/LT approach.

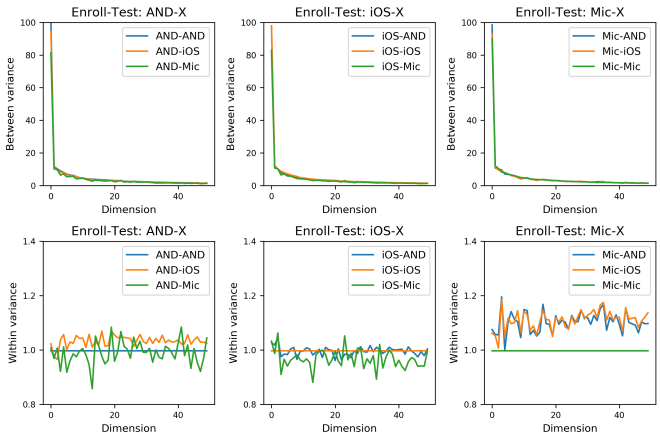

Fig. 2: Between-speaker and within-speaker variances (the first 50 dimensions) in the Cross-Channel test.

TABLE I: Angle metric and length metric in the Cross-Channel test.

Enroll-Test	Angle (θ)	Length (ℓ)
AND-iOS	0.940	4.673
iOS-Mic	5.199	10.789
Mic-AND	5.645	10.867

2) Time-Variance test: We use the *CSLT-Chronos* dataset to measure the statistics incoherence in the Time-Variance test. The dataset involves 57 speakers recorded with 8 recording sessions in the same acoustic environment. The data of the first session is used for enrollment, and all the 8 sessions are used for test. In order to study the incremental change on the statistics, we perform a serial test where the test sessions are gradually added to the test data. For example, the test **1st-[1st:5th]** means the enrollment is based on the first session, and the test is based on the data from the first to the fifth

[7]LDA plays a role of distribution normalization, by which the within-speaker covariance is normalized to an identity matrix $\mathbf{I}$.

sessions. Note that the test **1st-[1st:1st]** is the condition-matched test.

The between-speaker and within-speaker variances of the test data are reported in Figure 3, where the enrollment is based on the first session in all the test cases. It can be observed that when more sessions are involved in the test data, the pattern of the between-speaker variances does not change much, but the within-speaker variances are gradually increased and finally converge to stable values. The angle metric and length metric for the condition-mismatched tests are reported in Table II. Compared to the values in the Cross-Channel test shown in Table I, one can find that the global shift is small in the Time-Variance test.

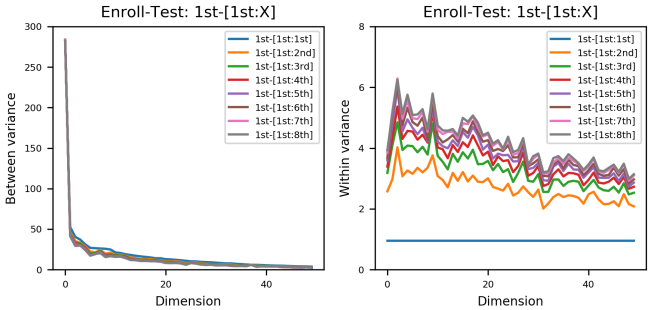

Fig. 3: Between-speaker and within-speaker variances (the first 50 dimensions) in the Time-Variance test.

TABLE II: Angle metric and length metric for global shift in the Time-Variance test.

Enroll-Test	Angle (θ)	Length (ℓ)
1st-[1st:2nd]	0.085	1.615
1st-[1st:3rd]	0.159	2.326
1st-[1st:4th]	0.234	2.654
1st-[1st:5th]	0.269	2.991
1st-[1st:6th]	0.301	3.224
1st-[1st:7th]	0.346	3.439
1st-[1st:8th]	0.393	3.590

All these observations are consistent with the results obtained from the qualitative analysis presented in the previous section. From these observations, we can conclude that in the Time-Variance test, the enrollment-test mismatch mainly impacts the within-speaker variance. Therefore, a simple within-speaker variance adaptation may be a good approach to alleviate the mismatch, as assumed by the second special case of our SD/LT proposal.

3) Near-Far test: The *HI-MIA.Train* dataset is used to compute the statistics in the Near-Far test. It involves 254 speakers recorded with 3 recording distances at the same time.

As in the Cross-Channel test, we examine the between-speaker and within-speaker variances of the test data and present them independently for three enrollment conditions. The results are shown in Figure 4. Firstly, it can be observed that the patterns of the between-speaker variances with the three enrollment conditions are different: enrollment with the near-filed speech (1m) shows relatively large between-class variance. Moreover, in each enrollment scenario, the near-field test speech shows a higher between-class variance. These observations demonstrated that: (1) the near-field speech is more speaker discriminative, used for either enrollment or test; (2) the mismatch between the near-field and far-field speech is reflected in the between-class variance.

Paying attention to the within-speaker variance, it shows that in all the enrollment scenarios, the condition-matched test data (e.g., 3m-3m) show the smallest and theoretically correct within-class variance; all the condition-mismatched test data exhibit a larger within-class variance. On the one hand, it suggests that the within-class variance of speech in different conditions are different; on the other hand, it implies that the directions of the within-class covariance is different for speech in different conditions, since speech in one condition always gets a higher variance in the principle directions of the speech in the other condition.

The angle metric and length metric are reported in Table III. It can be observed that the global shifts between two different conditions are significant, when compared to results in the Cross-Channel test (Table I) and the Time-Variance test (Table II). Moreover, the global shift between the two far-field conditions is much smaller than that between the near-field condition and the two far-field conditions.

As a summary, in the Near-Far test, the statistics incoherence is highly complex. It involves mismatch in the between-speaker and within-speaker variances, as well as the change on the directions of the within-speaker distribution. Additionally, a clear global-shift is observed. Therefore, it cannot be well addressed by the two special cases of the SD/LT approach; the general form of SD/LT should be employed.

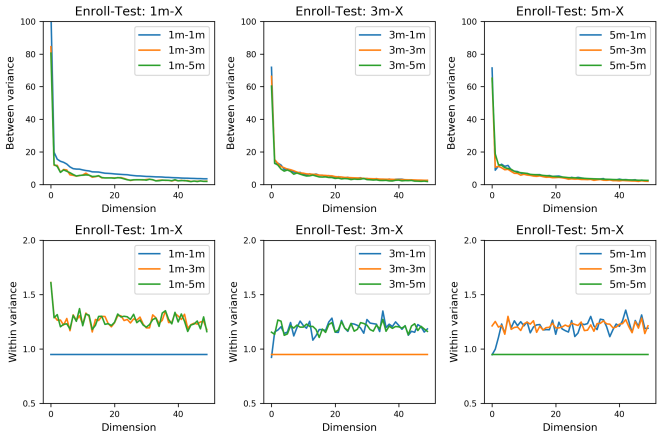

Fig. 4: Between-speaker and within-speaker variances (the first 50 dimensions) in the Near-Far test.

TABLE III: Angle metric and length metric in the Near-Far test.

Enroll-Test	Angle (θ)	Length (ℓ)
1m-3m	4.959	10.745
1m-5m	6.509	11.890
3m-5m	1.703	4.319

E. Scoring methods in comparison

We verify the SD/LT method on the speaker recognition task, by comparing different forms of SD/LT (two special cases and one general form) and several baseline methods. To make the comparison fair, we categorize these methods by the data it requires: the no-data methods do not require any data from the test condition; the unparallel-data methods require data from the test condition but the data are independent from the enrollment data; the parallel-data methods require data from the test condition and the speakers are shared with the enrollment data.

All the methods are based on the NL scoring approach, and the statistical models are trained with the PLDA code [43] implemented in the Kaldi toolkit [61]. The source code for the NL scoring has been published online[8].

1) Non-data methods:

- **Baseline**: The basic NL scoring approach where the statistical models are trained with development set in the enrollment condition, and the model is employed to perform the test by ignoring the enrollment-test mismatch.

2) Nonparallel-data methods:

- **GSC**: Global shift compensation (GSC). It corresponds to the special case 1 of the SD/LT approach, as mentioned in Section III-B, and assumes that the enrollment-test mismatch can be largely compensated for by a global shift, as shown in Eq.(18). Note that the global shift can be easily computed by the nonparallel development data. Due to the previous qualitative and quantitative analysis, we expect that this method is effective in the Cross-Channel test.
- **WVA**: Within variance adaptation (WVA). It corresponds to the special case 2 of the SD/LT approach, as mentioned in Section III-C. It assumes that the enrollment-test mismatch is mainly due to the difference in the within-speaker variance, and so can be compensated for by using the correct within-speaker variance as Eq. (19). Note that the within-speaker variances in the enrollment and test conditions can be obtained using the nonparallel development data. Due to the previous qualitative and quantitative analysis in the previous section, we expect this method is effective in the Time-Variance test.

3) Parallel-data methods:

- **SD/LT**: Linear conditional transform (SD/LT). It corresponds to the general form as mentioned in Section III-A. This model is a linear transform $\boldsymbol{x} = \mathbf{M}\hat{\boldsymbol{x}} + \boldsymbol{b}$ that maps the test sample $\hat{\boldsymbol{x}}$ to the enrollment condition so that it can be represented by the statistical model of the enrollment condition. Importantly, the transform is only used to compute the likelihood term $p_k(\boldsymbol{x})$; and the normalization term $p(\boldsymbol{x})$ is still computed using the statistical model of the test condition, based on the original data $\hat{\boldsymbol{x}}$. We implemented it based on the MLE criterion with the training objective shown in Eq. (17). In our experiment, the Adam optimizer [63] was used to optimize the parameters $\{\mathbf{M}, \boldsymbol{b}\}$.
- **MCT**: Multi-condition training (MCT). A widely used baseline approach to deal with enrollment-test condition mismatch. In this method, data from both the enrollment and test conditions are pooled together, and then are used to train the PLDA model. This model can be regarded as an interpolation of the models for the enrollment and test conditions, respectively. Although not an optimal solution, MCT often delivers good performance in practice. Note that it requires parallel data to discover the speaker subspace shared by different conditions.
- **CAT**: Condition-adaptation training (CAT). Another simple baseline that transforms speaker vectors from the test condition to the enrollment condition and then performs the scoring there. The transform is linear and is the same as in SD/LT, defined by $\boldsymbol{x} = \mathbf{M}\hat{\boldsymbol{x}} + \boldsymbol{b}$, and the parameters $\{\mathbf{M}, \boldsymbol{b}\}$ are estimated using the same ML training as the SD/LT approach. We highlight that the main difference between SD/LT and CAT is that for SD/LT, the normalization term $p(\boldsymbol{x})$ of the NL score is computed using the model of the test condition, and for CAT, it is computed using the model of the enrollment condition. This difference is not trivial: the normalization term computed in the test condition is optimal if the statistical model for the test condition is accurate, so SD/LT will be MBR optimal under that condition, but CAT is not.

F. Speaker recognition results

In this section, experimental results in terms of equal error rate (EER) are presented for the three speaker recognition tasks: the Cross-Channel test, the Time-Variance test and the Near-Far test. We will use 'Base', 'NPD', 'PD' to represent the Baseline method, the nonparallel data methods (including GSC and WAV), and the parallel data methods (including SD/LT, MCT and CAT), respectively.

1) Cross-Channel test: In this test, the statistical models of all the scoring methods are trained with the *AIShell-1.Dev* dataset, and the results on the *AIShell-1.Eval* dataset are reported in Table IV. The observations are as follows.

- The performance with matched enrollment and test conditions is much higher than those with mismatched conditions, which indicates that the enrollment-test mismatch is indeed a serious problem for speaker recognition. Moreover, the performance loss with different mismatch condition pairs is different. For instance, The pair AND-iOS obtains better performance than the pair AND-Mic. This is consistent to the observations in the

[8]https://gitlab.com/csltstu/enroll-test-mismatch

TABLE IV: EER(%) on the Cross-Channel test.

Enroll-Test	Base	NPD		PD		
		GSC	WVA	MCT	CAT	SD/LT
AND-AND	0.797	-	-	-	-	-
AND-Mic	2.146	1.764	2.165	1.151	1.245	**0.981**
AND-iOS	1.425	1.382	1.401	1.161	1.312	**0.623**
Mic-AND	2.175	1.665	2.033	1.161	1.189	**0.712**
Mic-Mic	0.778	-	-	-	-	-
Mic-iOS	2.251	1.892	2.081	1.293	1.481	**0.812**
iOS-AND	1.599	1.430	1.590	1.156	1.184	**0.755**
iOS-Mic	2.216	1.759	2.231	1.137	1.231	**1.052**
iOS-iOS	0.920	-	-	-	-	-

qualitative and quantitative analysis, in particular the results in Figure 2 and Table I.

- For the nonparallel data methods, the GSC method consistently outperforms the baseline on all these mismatched conditions, while WVA does not show much advantage. This observation supports our argument that the main effect of the cross-channel mismatch is a global shift, and a simple GSC can compensate for the mismatch to a large extent.
- The three parallel data methods deliver much better performance than the two nonparallel data methods. That is not surprising as the parallel data methods utilize extra information of cross-channel speakers. Moreover, the proposed SD/LT approach consistently outperforms MCT and CAT, demonstrating that SD/LT is a more effective approach in dealing with enrollment-test mismatch. The comparison between SD/LT and CAT is especially interesting, as the two methods look very similar and only differ in the normalization term. The clear advantage of SD/LT demonstrated the importance of a principle solution based on a solid theory.

2) Time-variance test: The test is conducted with the *CSLT-Chronos* dataset, which consists of 57 speakers and each speaker has 150 utterances. The first 20 utterances of each speaker are selected as the enrollment set, as well as the development set to train the scoring models. The rest 130 utterances are used for evaluation. The EER results are reported in Table V. The observations are as follows.

- With more sessions involved in the test data, the performance of the baseline system is gradually reduced, and finally tends to converge to a stable value. This is consistent with our qualitative and quantitative analysis that with more sessions involved, the within-class variance will be gradually increased, leading to more enrollment-test mismatch.
- Looking at the performance of the two nonparallel methods, it can be found that compared with the baseline, WVA obtains obvious performance improvement in all the test conditions, while GSC does not show much help. This supports our argument that the main effect of the time-variance mismatch is in the within-speaker variance.
- Again, the three parallel data methods attain much better performance than the two nonparallel methods, and SD/LT attains the best performance. Interestingly, when the test data is sufficient large, the SD/LT performance is even better than the condition-matched baseline, i.e., the test **1st-[1st:1st]**. We attribute this good performance to the more reliable estimation for the within-speaker and between-speaker covariances of the test condition. As a comparison, the CAT performance is never better than the condition-matched baseline, as it does not fully use the statistical model of the test condition.

TABLE V: EER(%) on the Time-Variance test.

Enroll-Test	Base	NPD		PD		
		GSC	WVA	MCT	CAT	SD/LT
1st-[1st:1st]	4.799	-	-	-	-	-
1st-[1st:2nd]	6.400	6.346	5.934	5.258	5.549	**4.339**
1st-[1st:3rd]	6.863	6.777	6.156	4.976	5.242	**4.062**
1st-[1st:4th]	6.884	6.810	6.084	4.619	4.885	**3.710**
1st-[1st:5th]	7.108	7.022	6.230	4.348	4.804	**3.675**
1st-[1st:6th]	7.856	7.768	6.938	4.348	4.861	**3.589**
1st-[1st:7th]	7.906	7.825	7.005	4.262	4.903	**3.746**
1st-[1st:8th]	7.881	7.815	6.993	4.300	5.041	**3.939**

3) Near-Far field test: In this test, the *HI-MIA.Dev* dataset is used to train the scoring models, and the results on the *HI-MIA.Eval* dataset are reported in Table VI. The observations are as follows.

TABLE VI: EER(%) on the Near-Far test.

Enroll-Test	Base	NPD		PD		
		GSC	WVA	MCT	CAT	SD/LT
1m-1m	0.620	-	-	-	-	-
1m-3m	3.968	3.725	3.644	**2.510**	3.563	2.996
1m-5m	4.866	4.623	4.623	**3.082**	4.623	3.731
3m-1m	1.938	1.705	1.705	1.318	1.628	**1.085**
3m-3m	0.891	-	-	-	-	-
3m-5m	3.244	3.325	3.082	2.595	3.244	**2.433**
5m-1m	3.566	3.256	3.411	2.326	2.481	**2.171**
5m-3m	2.834	2.591	2.591	**2.105**	2.753	**2.105**
5m-5m	1.135	-	-	-	-	-

- The performance is seriously degraded with the near-far recoding condition mismatch, especially when enroll with near-field speech and test on far-field speech.
- The two nonparallel methods, GSC and WAV, both offer reasonable performance improvement compared to the baseline. This is consistent with our quantitative analysis that this recording condition mismatch incurs both incoherence on within-class variances and global shift.

- Again, the three parallel data methods offers more performance gains, as in the Cross-Channel test and the Time-Variance test. A slight difference is that in this test, SD/LT does not always obtain the test performance. Among the 6 mismatched scenarios, MCT outperforms SD/LT in two scenarios. This is a little surprising as SD/LT is MBR optimal in theory. More careful analysis shows that it is probably because the linear transform nature of SD/LT cannot address the complex incoherence caused by the different recording distances.

VI. Conclusions

This paper investigated the issue of enrollment-test mismacth in speaker recognition, and presented a statistics decomposition (SD) approach to solve the problem. Specifically, we decouple the scoring process to separate phases according to the normalized likelihood (NL) framework, and statistics from different conditions are used in different phases. This SD approach guarantees that the resultant scores are MBR optimal, and therefore is a principle solution for the enrollment-test mismatch problem.

As an initial study, we proposed an implementation of the SD approach based on a linear transform (SD/LT). A comprehensive study was conducted and the experimental results demonstrated that: (1) Different types of mismatch exhibits different statistic incoherence, and so should be treated differently; (2) If the statistics incoherence shows a simple form, a simplified version of the SD/LT algorithm would be effective. An advantage of these simplified algorithms is that they only require nonparallel data, i.e., data without cross-condition speakers; (3) If parallel data are available, the general form SD/LT algorithm is highly effective and obtained the best performance among all the competitive methods in the test.

Future work will investigate the conditional transform based on more complex transform functions, e.g., a deep neural net. This may significantly improve the performance in scenarios with complex condition mismatch, for instance the near-far recording distance mismatch in our experiment.

References

[1] J. P. Campbell, "Speaker recognition: A tutorial," *Proceedings of the IEEE*, vol. 85, no. 9, pp. 1437–1462, 1997.

[2] D. A. Reynolds, "An overview of automatic speaker recognition technology," in *IEEE international conference on Acoustics, speech, and signal processing (ICASSP)*, vol. 4. IEEE, 2002, pp. IV–4072.

[3] J. H. Hansen and T. Hasan, "Speaker recognition by machines and humans: A tutorial review," *IEEE Signal processing magazine*, vol. 32, no. 6, pp. 74–99, 2015.

[4] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital signal processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[5] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1435–1447, 2007.

[6] N. Dehak, P. J. 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.

[7] L. Deng and D. Yu, "Deep learning: methods and applications," *Foundations and trends in signal processing*, vol. 7, no. 3–4, pp. 197–387, 2014.

[8] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 4052–4056.

[9] L. Li, Y. Chen, Y. Shi, Z. Tang, and D. Wang, "Deep speaker feature learning for text-independent speaker verification," in *Proceedings of the Annual Conference of International Speech Communication Association (INTERSPEECH)*, 2017, pp. 1542–1546.

[10] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.

[11] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive statistics pooling for deep speaker embedding," in *Proceedings of the Annual Conference of International Speech Communication Association (INTERSPEECH)*, 2018, pp. 2252–2256.

[12] G. Heigold, I. Moreno, S. Bengio, and N. Shazeer, "End-to-end text-dependent speaker verification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5115–5119.

[13] S.-X. Zhang, Z. Chen, Y. Zhao, J. Li, and Y. Gong, "End-to-end attention based text-dependent speaker verification," in *Spoken Language Technology Workshop (SLT)*. IEEE, 2016, pp. 171–178.

[14] F. R. Rahman Chowdhury, Q. Wang, I. L. Moreno, and L. Wan, "Attention-based models for text-dependent speaker verification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5359–5363.

[15] D. Wang, L. Li, Z. Tang, and T. F. Zheng, "Deep speaker verification: Do we need end to end?" in *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2017, pp. 177–181.

[16] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep speaker recognition," in *Proceedings of the Annual Conference of International Speech Communication Association (INTERSPEECH)*, 2018, pp. 1086–1090.

[17] J. weon Jung, H.-S. Heo, J. ho Kim, H. jin Shim, and H.-J. Yu, "RawNet: Advanced end-to-end deep neural network using raw waveforms for text-independent speaker verification," in *Proceedings of the Annual Conference of International Speech Communication Association (INTERSPEECH)*, 2019, pp. 1268–1272.

[18] W. Cai, J. Chen, and M. Li, "Exploring the encoding layer and loss function in end-to-end speaker and language recognition system," in *Proceedings of Odyssey: The Speaker and Language Recognition Workshop*, 2018, pp. 74–81.

[19] W. Xie, A. Nagrani, J. S. Chung, and A. Zisserman, "Utterance-level aggregation for speaker recognition in the wild," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 5791–579.

[20] N. Chen, J. Villalba, and N. Dehak, "Tied mixture of factor analyzers layer to combine frame level representations in neural speaker embeddings," in *Proceedings of the Annual Conference of International Speech Communication Association (INTERSPEECH)*, 2019, pp. 2948–2952.

[21] L. Li, D. Wang, C. Xing, and T. F. Zheng, "Max-margin metric learning for speaker recognition," in *10th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 2016, pp. 1–4.

[22] W. Ding and L. He, "MTGAN: Speaker verification through multi-tasking triplet generative adversarial networks," in *Proceedings of the Annual Conference of International Speech Communication Association (INTERSPEECH)*, 2018, pp. 3633–3637.

[23] J. Wang, K.-C. Wang, M. T. Law, F. Rudzicz, and M. Brudno, "Centroid-based deep metric learning for speaker recognition," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 3652–3656.

[24] Z. Bai, X.-L. Zhang, and J. Chen, "Partial AUC optimization based deep speaker embeddings with class-center learning for text-independent speaker verification," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6819–6823.

[25] Z. Gao, Y. Song, I. McLoughlin, P. Li, Y. Jiang, and L.-R. Dai, "Improving aggregation and loss function for better embedding learning in end-to-end speaker verification system," in *Proceedings of the Annual Conference of International Speech Communication Association (INTERSPEECH)*, 2019, pp. 361–365.

[26] J. Zhou, T. Jiang, Z. Li, L. Li, and Q. Hong, "Deep speaker embedding extraction with channel-wise feature responses and additive supervision softmax loss function," in *Proceedings of the Annual Conference of International Speech Communication Association (INTERSPEECH)*, 2019, pp. 2883–2887.

[27] R. Li, N. Li, D. Tuo, M. Yu, D. Su, and D. Yu, "Boundary discriminative large margin cosine loss for text-independent speaker verification," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6321–6325.

[28] S. Wang, J. Rohdin, L. Burget, O. Plchot, Y. Qian, K. Yu, and J. Cernocky, "On the usage of phonetic information for text-independent speaker embedding extraction," in *Proceedings of the Annual Conference of International Speech Communication Association (INTERSPEECH)*, 2019, pp. 1148–1152.

[29] T. Stafylakis, J. Rohdin, O. Plchot, P. Mizera, and L. Burget, "Self-supervised speaker embeddings," in *Proceedings of the Annual Conference of International Speech Communication Association (INTERSPEECH)*, 2019, pp. 2863–2867.

[30] S. Shon, S. Mun, W. Kim, and H. Ko, "Autoencoder based domain adaptation for speaker recognition under insufficient channel information," in *Proceedings of the Annual Conference of International Speech Communication Association (INTERSPEECH)*, 2017, pp. 1014–1018.

[31] Q. Wang, W. Rao, S. Sun, L. Xie, E. S. Chng, and H. Li, "Unsupervised domain adaptation via domain adversarial training for speaker recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4889–4893.

[32] J. Villalba and E. Lleida, "Bayesian adaptation of PLDA based speaker recognition to domains with scarce development data," in *Proceedings of Odyssey: The Speaker and Language Recognition Workshop*, 2012, pp. 47–54.

[33] D. Garcia-Romero and A. McCree, "Supervised domain adaptation for i-vector based speaker recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 4047–4051.

[34] D. Garcia-Romero, A. McCree, S. Shum, and C. Vaquero, "Unsupervised domain adaptation for i-vector speaker recognition," in *Proceedings of Odyssey: The Speaker and Language Recognition Workshop*, 2014.

[35] H. Aronowitz, "Inter dataset variability compensation for speaker recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 4002–4006.

[36] H. Aronowitz, "Compensating inter-dataset variability in PLDA hyper-parameters for robust speaker recognition." in *Proceedings of Odyssey: The Speaker and Language Recognition Workshop*, 2014, pp. 280–286.

[37] A. Kanagasundaram, D. Dean, and S. Sridharan, "Improving out-domain PLDA speaker verification using unsupervised inter-dataset variability compensation approach," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 4654–4658.

[38] M. H. Rahman, A. Kanagasundaram, I. Himawan, D. Dean, and S. Sridharan, "Improving PLDA speaker verification performance using domain mismatch compensation techniques," *Computer Speech & Language*, vol. 47, pp. 240–258, 2018.

[39] E. Shriberg, M. Graciarena, H. Bratt, A. Kathol, S. 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.

[40] C. Zhang, S. Ranjan, and J. H. Hansen, "An analysis of transfer learning for domain mismatched text-independent speaker verification," in *Odyssey*, 2018, pp. 181–186.

[41] S. J. Park, C. Sigouin, J. Kreiman, P. A. Keating, J. Guo, G. Yeung, F.-Y. Kuo, and A. Alwan, "Speaker identity and voice quality: Modeling human responses and automatic speaker recognition." in *Interspeech*, 2016, pp. 1044–1048.

[42] A. Afshan, J. Guo, S. J. Park, V. Ravi, A. McCree, and A. Alwan, "Variable frame rate-based data augmentation to handle speaking-style variability for automatic speaker verification," *arXiv preprint arXiv:2008.03616*, 2020.

[43] S. Ioffe, "Probabilistic linear discriminant analysis," in *European Conference on Computer Vision (ECCV)*. Springer, 2006, pp. 531–542.

[44] S. J. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *2007 IEEE 11th International Conference on Computer Vision*. IEEE, 2007, pp. 1–8.

[45] D. Wang, "Remarks on optimal scores for speaker recognition," CSLT@Tsinghua University, 2020. [Online]. Available: http://wangd.cslt.org/public/pdf/nl.pdf

[46] P. Kenny, T. Stafylakis, P. Ouellet, M. J. Alam, and P. Dumouchel, "PLDA for speaker verification with utterances of arbitrary duration," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7649–7653.

[47] B. J. Borgström and A. McCree, "Discriminatively trained bayesian speaker comparison of i-vectors," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7659–7662.

[48] A. McCree, G. Sell, and D. Garcia-Romero, "Extended variability modeling and unsupervised adaptation for plda speaker recognition." in *INTERSPEECH*, 2017, pp. 1552–1556.

[49] I. J. Myung, "Tutorial on maximum likelihood estimation," *Journal of mathematical Psychology*, vol. 47, no. 1, pp. 90–100, 2003.

[50] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer speech & language*, vol. 9, no. 2, pp. 171–185, 1995.

[51] J. Williams and S. King, "Disentangling style factors from speaker representations," in *Proc. Interspeech 2019*, 2019, pp. 3945–3949. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-1769

[52] G. Bhattacharya, J. Alam, and P. Kenny, "Adapting end-to-end neural speaker verification to new languages and recording conditions with adversarial training," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6041–6045.

[53] W. H. Kang, S. H. Mun, M. H. Han, and N. S. Kim, "Disentangled speaker and nuisance attribute embedding for robust speaker verification," *IEEE Access*, 2020.

[54] K. A. Lee, Q. Wang, and T. Koshinaka, "The coral+ algorithm for unsupervised domain adaptation of plda," 2019.

[55] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: a large-scale speaker identification dataset," *arXiv preprint arXiv:1706.08612*, 2017.

[56] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "AIShell-1: An open-source mandarin speech corpus and a speech recognition baseline," in *Oriental COCOSDA 2017*, 2017, p. Submitted.

[57] L. Wang and T. F. Zheng, "Creation of time-varying voiceprint database," in *Proc. Oriental-COCOSDA*, 2010.

[58] X. Qin, H. Bu, and M. Li, "HI-MIA: A far-field text-dependent speaker verification database and the baselines," 2019.

[59] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[60] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech

recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5220–5224.

[61] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The Kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. EPFL-CONF-192584. IEEE Signal Processing Society, 2011.

[62] L. v. d. Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[63] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

Lantian Li received the B.Sc. degree from China University of Mining and Technology, Beijing in 2013. He received the Ph.D. degree from the Department of Computer Science, Tsinghua University in 2018. Since 2018, he has been with the Center for Speech and Language Technology (CSLT), Tsinghua University as a postdoctoral fellow. His research interest is speaker recognition with machine learning methods.

Dong Wang received the B.Sc. and M.Sc. degrees in computer science from Tsinghua University in 1999 and 2002. He received the Ph.D. degree (supported by a Marie Curie fellowship) from CSTR, University of Edinburgh, in 2010. He was employed with Oracle China during 2002 to 2004, and IBM China during 2004 to 2006. He joined CSTR, University of Edinburgh, in 2006 as a Research Fellow. From 2010 to 2011, he was with EURECOM as a Postdoctoral Fellow, and from 2011 to 2012, was a Senior Research Scientist with Nuance. He is now an Associate Professor with Tsinghua University, Beijing, China.

Jiawen Kang received the B.Sc. degree from Jilin University in 2019. Since 2019, he has been a research intern at Center for Speech and Language Technology (CSLT), Tsinghua University.

Renyu Wang received the B.Sc. degree in computer science and M.Sc. degree in linguistic science from Jiangsu Normal University in 2014 and 2017. He was employed with Samsung R&D Institute China during 2017 to 2019. He is now a researcher at Noah's Ark Lab, Huawei, China.

Jing Wu received the B.Sc degree from Nanjing University of Science and Technology in 2015. Since 2015, he has been a researcher at Huawei, Shanghai, China.

Zhengdong Gao received the Ph.D. degree from Shanghai Jiaotong University in 2007. Since 2008, he has been a researcher at Huawei, Shanghai, China.

Xiao Chen received the Ph.D. degree from City University of Hong Kong in 2012. He was employed with Sogou-inc during 2012 to 2014. He is now a principal researcher at Noah's Ark Lab, Huawei, China.