

RESEARCH ARTICLE

Reflection of Conditional Independence Structure to Noise Variability for Noise Robust Text Dependent Speaker Verification

SUNGHYUN YOON , (Member, IEEE)

Department of Artificial Intelligence, Kongju National University, Cheonan 31080, South Korea

e-mail: syoon@kongju.ac.kr


This work was supported in part by the Research Grant from Kongju National University in 2022; and in part by the “Regional Innovation Strategy (RIS)” through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (MOE) (2022D-02-06-07-008).

ABSTRACT In the field of speaker verification (SV), the development of noise-robust systems is a challenge for their deployment in real-world environments. Noise variability compensation is a common strategy for increasing the robustness to noise variations. The performance of noise compensation depends on how well the noise variability, which is inherent in within-class variability, is estimated. However, to date, there is no information about the true noise variability that could reduce the gap between the empirical and true statistics. Most studies assume that true noise covariates are independent. This study aims to demonstrate the assumption that true noise variability has a conditional independence structure rather than an independence structure. This assumption was motivated by our previous findings, which revealed that optimal within-class variability has a conditional independence structure in text-dependent speaker verification (TD-SV) in clean environments. This indicates that the optima of all the variabilities in within-class variability, except noise variability, has a conditional independence structure; however, it is unknown whether this is also true for optimal noise variability. Our assumption was supported by the experimental results obtained under noisy TD-SV trials using systems built with graphical least absolute shrinking and selection operator-based probabilistic linear discriminant analysis, which achieved up to 10% relative equal error rate improvements.

INDEX TERMS Background noise, conditional independence, graphical least absolute shrinking and selection operator (GLASSO), probabilistic linear discriminant analysis (PLDA), text-dependent speaker verification.

I. INTRODUCTION

Automatic speaker verification (ASV) is a biometric technique that is used to verify the identity of a user by voice. When a user speaks with the claim that he/she is the same person as a known target speaker (whose reference utterance is pre-enrolled), the ASV compares the user's utterance (corresponding to the test utterance) with the target speaker's reference utterance (corresponding to the enrollment utterance). An identity claim is accepted if the similarity between the enrollment and test utterances exceeds a pre-defined threshold; otherwise, it is rejected.

The associate editor coordinating the review of this manuscript and approving it for publication was Filbert Juwono .

Depending on whether the phrase of utterance is constrained, there are two categories of ASV systems: text-independent speaker verification (TI-SV) and text-dependent speaker verification (TD-SV). This study focuses on TD-SV. In contrast to TI-SV, where no constraint exists for the phrase, TD-SV restricts speakers to the phrases in a fixed lexicon, speaking exactly the same phrase as one of those in a fixed lexicon. The identity claim is rejected if the enrollment and test utterances have different phrases, even if both are from the same speaker. This phrase constraint makes TD-SV less flexible than TI-SV. Nonetheless, TD-SV can achieve higher verification accuracy than TI-SV even with short utterances because the uncertainty of the phrase, which is a factor that degrades the accuracy of TI-SV, is suppressed and thus easier to handle. In addition, users do not need to speak at length

with TD-SV, which makes it more convenient to use. Owing to the above advantages, TD-SV has been widely used in various authentic applications that require both high accuracy and convenience, such as voice assistants [1], [2].

We focus on clean-enrollment and noisy-test conditions for TD-SV, which is the most common condition in real environments. Throughout this paper, we use the term ‘noise’ is used to mean ‘background noise in utterance, not ‘statistical uncertainty’, unless otherwise noted. In real environments, there are various types of background noise, which is one of the major factors that degrades the verification accuracy. TD-SV is generally more sensitive to background noise because of its shorter length. However, it is impossible to completely remove the noise from a noisy utterance [3]. Furthermore, it is unrealistic to constrain users to use ASV systems only in clean environments during the test phase. Consequently, noise-robust ASV systems have gained considerable importance.

Recently, many studies have been conducted to develop various types of noise-robust ASV systems [4], [5], [6], [7], [8]. We address noise robustness based on probabilistic linear discriminant analysis (PLDA) [9], [10]. PLDA determines a more discriminative subspace by probabilistically modeling between- and within-class variabilities. The similarity score between two embeddings (e.g., feature vector from utterance) is computed on the PLDA subspace rather than the original embedding space. End-to-end approaches, where the front-end embedding extraction and back-end scoring modules are jointly optimized as a single module, have emerged as one of the noteworthy methods in ASV over the last several years [11], [12], [13], [14]. However, separate modeling of the front- and back-end modules, especially with the PLDA backend [15], [16], [17], [18], [19], [20], [21], has achieved compelling performances. Thus, PLDA has become a popular back-end scoring method for ASV.

Because the true statistics (e.g., the between- and within-class variabilities in the PLDA) are unknown, we use empirical statistics from the training dataset instead. It is reasonable that better performances are achieved if the empirical statistics are closer to the true statistics. In our previous work [23], we extended PLDA by applying the graphical least absolute shrinking and selection operator (GLASSO) [24], [25], [26], dubbed GLASSO-PLDA. The GLASSO-PLDA is based on the following statements: (i) the empirical statistics contain estimation errors that should be reduced and (ii) the true within-class variability is assumed to have a conditional independence structure (i.e., the true within-class precision matrix is sparse, but not diagonal) in TD-SV. GLASSO-PLDA addresses these by making a within-class precision matrix (namely, the inverse of the covariance matrix) sparse using GLASSO. Using GLASSO-PLDA, we achieved significant performance improvements in TD-SV under clean conditions and confirmed that our assumption holds.

In [23], we considered only clean conditions for both the enrollment and test phases. It is not yet clear whether the true noise variability has a conditional independence structure.

Building noise-robust ASV systems would become easier if the true noise variability had a certain structure and the bias toward the true structure could be utilized during the training step. However, no prior information about noise variability is available, which makes their construction difficult [8], [27]. Typically, noise covariance is assumed to be isotropic (i.e., true covariates are independent and the true covariance matrix is diagonal) to simplify the development, as in [28]. However, the isotropic assumption is extremely restrictive. Moreover, to the best of our knowledge, no studies on ASV have considered any specific structure beyond simple independence for noise variability. In this study, we assumed that the true noise variability has a conditional independence structure (i.e., the true precision matrix is sparse but not diagonal). We confirmed this assumption by evaluating GLASSO-PLDA in noisy test environments. If our assumption is valid, GLASSO-PLDA would result in performance improvements under noisy conditions. The contributions of this study are threefold.

- 1) We formulate the assumption of true noise variability with a conditional independence structure (i.e., sparse structure in the true noise precision matrix).
- 2) We propose a method to reflect the conditional independence structure to noise variability using the proposed GLASSO-PLDA.
- 3) We demonstrate that the true noise variability has a conditional independence structure by evaluating the performance of the proposed method under various noisy conditions.

The remainder of this paper is organized as follows. Section II outlines the background of this study. Section III introduces the proposed method. Section IV describes the experiments and discusses their results. Finally, Section V concludes the paper.

II. BACKGROUND

A. VARIABILITIES IN TD-SV

In this section, we describe TD-SV in terms of variability. In TD-SV, the class (i.e., speaker-phrase pair) identity depends on both the speaker information and the phrase information, owing to the restriction on the available lexicon. Therefore, the between-class variability Σ_b , which explains the variations across different classes, consists of the speaker variability Σ_{spk} and phrase variability Σ_{phr} , namely $\Sigma_b = \Sigma_{spk} + \Sigma_{phr}$. The within-class variability Σ_w describes the uncertainty within a class. Under clean conditions, Σ_w can be decomposed into the session variability Σ_{sess} and residual variability (i.e., all the other unexplained variabilities) Σ_ϵ , i.e., $\Sigma_w = \Sigma_{sess} + \Sigma_\epsilon$. Note that we use the term session variability to describe all possible variabilities within a class in clean condition, such as the variabilities of transmission channel, reverberation, and distance from the microphone. Under noisy conditions, where the noise variability Σ_{noise} is inherent in Σ_w , Σ_w can be decomposed as $\Sigma_w = \Sigma_{sess} + \Sigma_{noise} + \Sigma_\epsilon$.

The main purpose of this study is to confirm the conditional independence structure of the optimum of Σ_{noise} . However, it is unrealistic to directly treat Σ_{noise} because it is difficult to perfectly disentangle Σ_{noise} from $\Sigma_w = \Sigma_{sess} + \Sigma_{noise} + \Sigma_\epsilon$. Therefore, we treat Σ_w instead. Reflecting the conditional independence structure to Σ_w corresponds to reflecting the structure to all the variabilities in Σ_w , namely Σ_{sess} , Σ_{noise} , and Σ_ϵ . In our previous work [23], we confirmed that the optima of Σ_{ch} and Σ_ϵ possess the conditional independence structure. Therefore, it can be inferred that the optimum of Σ_{noise} also possess the conditional independence structure if the performance under noisy conditions improves when reflecting the structure to $\Sigma_w = \Sigma_{sess} + \Sigma_{noise} + \Sigma_\epsilon$.

The above approach (i.e., reflecting the structure to Σ_w to confirm the true structure of Σ_{noise}) is reasonable only when Σ_{noise} comprises a sufficient proportion of Σ_w . When the scale of Σ_{noise} is negligible compared to that of $\Sigma_{sess} + \Sigma_\epsilon$, it is difficult to claim that the true Σ_{noise} has a conditional independence structure, even if reflecting the structure to Σ_w improves the performance. Therefore, the scale of Σ_{noise} relative to $\Sigma_w = \Sigma_{sess} + \Sigma_{noise} + \Sigma_\epsilon$ should be checked first.

B. PLDA

PLDA [9], [10], [29], [30], [31] is a generative probabilistic model in which the between- and within-class variabilities are modeled using latent variables. The PLDA aims to find a discriminative subspace where the between-class variability is maximized, and simultaneously, the within-class variability is minimized. Among some variants of PLDA, we used the two-covariance PLDA [30] based on [9] (i.e., Kaldi [32] PLDA).

Let $\mathbf{x} \in \mathbb{R}^D$ be speaker-phrase embedding. PLDA models \mathbf{x} as follows:

$$\mathbf{x} = \boldsymbol{\mu} + \mathbf{A}\mathbf{u} \quad (1)$$

$$\mathbf{u}|\mathbf{v} \sim N(\mathbf{v}, \mathbf{I}) \quad (2)$$

$$\mathbf{v} \sim N(\mathbf{0}, \boldsymbol{\Psi}) \quad (3)$$

where $\boldsymbol{\mu} \in \mathbb{R}^D$ is the mean embedding in the original space, $\mathbf{A} \in \mathbb{R}^{D \times D}$ is the PLDA projection matrix, $\mathbf{v} \in \mathbb{R}^D$ and $\mathbf{u} \in \mathbb{R}^D$ represent the class and an example of that class in the projected space, respectively, $N(\cdot)$ denotes the normal distribution, and $\boldsymbol{\Psi} \in \mathbb{R}^{D \times D}$ is the between-class diagonal covariance in the projected space. The dimensionality of the subspace was the same as that of the original space. The three parameters of the PLDA model, $\boldsymbol{\mu}$, \mathbf{A} , and $\boldsymbol{\Psi}$, can be computed using the following eigenproblem:

$$\Phi_w^{-1} \Phi_b \mathbf{A} = \mathbf{A} \boldsymbol{\Psi} \quad (4)$$

where $\Phi_b \in \mathbb{R}^{D \times D}$ and $\Phi_w^{-1} \in \mathbb{R}^{D \times D}$ represent the between-class covariance and within-class precision matrices, respectively. In PLDA, Φ_b and Φ_w^{-1} are estimated using the expectation-maximization (EM) algorithm [33] starting from the initial between-class covariance matrix $\Phi_b^{(0)} \in \mathbb{R}^{D \times D}$ and within-class covariance matrix $\Phi_w^{(0)} \in \mathbb{R}^{D \times D}$,

respectively. Both $\Phi_b^{(0)}$ and $\Phi_w^{(0)}$ are directly computed from the training dataset as follows:

$$\Phi_b^{(0)} = \sum_{c=1}^C (\boldsymbol{\mu}_c - \boldsymbol{\mu})(\boldsymbol{\mu}_c - \boldsymbol{\mu})^T \quad (5)$$

$$\Phi_w^{(0)} = \sum_{c=1}^C \frac{1}{N_c} \sum_{n=1}^{N_c} (\mathbf{x}_{c,i} - \boldsymbol{\mu}_c)(\mathbf{x}_{c,i} - \boldsymbol{\mu}_c)^T \quad (6)$$

where C is the number of classes (i.e., the number of different speaker-phrase pairs), and $\boldsymbol{\mu}_c \in \mathbb{R}^D$, N_c , and $\mathbf{x}_{c,i} \in \mathbb{R}^D$ represent the mean embedding, number of embeddings, and i -th embedding, respectively, for the class c .

Using the PLDA model, the log-likelihood ratio between two embeddings $\mathbf{x}_e \in \mathbb{R}^D$ and $\mathbf{x}_t \in \mathbb{R}^D$ (i.e., from the enrollment and test utterances, respectively) is computed as follows:

$$\log N\left(\mathbf{u}_e \middle| \frac{\boldsymbol{\Psi}}{\boldsymbol{\Psi} + \mathbf{I}} \mathbf{u}_t, \mathbf{I} + \frac{\boldsymbol{\Psi}}{\boldsymbol{\Psi} + \mathbf{I}}\right) - \log N(\mathbf{u}_e | \mathbf{0}, \mathbf{I} + \boldsymbol{\Psi}) \quad (7)$$

where $\mathbf{u}_e \in \mathbb{R}^D$ and $\mathbf{u}_t \in \mathbb{R}^D$ are the projected embeddings obtained using the transform of $\mathbf{u} = \mathbf{A}^T (\mathbf{x} - \boldsymbol{\mu})$ from \mathbf{x}_e and \mathbf{x}_t , respectively.

C. GLASSO

1) GAUSSIAN MARKOV RANDOM FIELD

Let $\mathbf{x} = [x_1, \dots, x_D]^T$ be a D -dimensional random vector (e.g., an embedding) that follows a multivariate normal distribution $N(\boldsymbol{\mu}, \Theta^{-1})$ with mean $\boldsymbol{\mu} \in \mathbb{R}^D$ and precision matrix $\Theta \in \mathbb{R}^{D \times D}$. The random vector \mathbf{x} is a Gaussian Markov random field (GMRF) if it satisfies Markov properties, related to conditional independence [34].

GMRF is an undirected graphical model. Let $G = (V, E)$ be an undirected graph, where V and E represent a set of vertices and set of edges, respectively. Each vertex $x_i \in V$ corresponds to a variable in \mathbf{x} . Each edge $e_{i,j} = e_{j,i} \in E$ represents the connection between distinct vertices (variables) x_i and x_j such that $i \neq j$. The set of edges E define the conditional dependencies of the vertices V . There is no edge $e_{i,j} \notin E$ if and only if variables x_i and x_j are conditionally independent $x_i \perp x_j | \mathbf{x}_{-ij}$ given all other variables \mathbf{x}_{-ij} , known as the pairwise Markov property [35].

In GMRF, E can be represented as Θ . The element Θ_{ij} in row i and column j of Θ corresponds to $e_{i,j}$. The zero value $\Theta_{ij} = 0$ corresponds to the absence of edge $e_{i,j} \notin E$, which means $x_i \perp x_j | \mathbf{x}_{-ij}$. It means that Θ contains the information about the covariances between x_i and x_j , conditioned on all other variables, called the partial covariances [36]. Therefore, the sparsity of Θ implies conditional independence of the variables. In other words, it is able to reflect the conditional independence structure to the variables by making Θ sparse [34].

2) GLASSO

GLASSO is a variable selection method based on LASSO L_1 regularization. Let $\Theta = \Sigma^{-1} \in \mathbb{R}^{D \times D}$ be the true

precision matrix and $\mathbf{S} \in \mathbb{R}^{D \times D}$ be the empirical covariance matrix estimated from samples (e.g., embeddings from the training dataset) that follow a multivariate normal distribution $N(\boldsymbol{\mu}, \boldsymbol{\Theta}^{-1})$. GLASSO estimates the sparse precision matrix $\hat{\boldsymbol{\Theta}}$ by iteratively maximizing the following L_1 -penalized Gaussian log-likelihood [25]:

$$\hat{\boldsymbol{\Theta}} = \underset{\boldsymbol{\Theta}}{\operatorname{argmax}} \log \det \boldsymbol{\Theta} - \operatorname{tr} \mathbf{S} \boldsymbol{\Theta} - \rho \|\boldsymbol{\Theta}\|_1 \quad (8)$$

where $\det(\cdot)$ and $\operatorname{tr}(\cdot)$ are the determinant and trace, respectively, $\rho \in \mathbb{R}^+$ is a hyperparameter for regularization, and $\|\cdot\|_1$ is the L_1 norm (i.e., the sum of the absolute values of the elements). As ρ is higher (i.e., more regularization), $\hat{\boldsymbol{\Theta}}$ is sparser, which means a lower variance (i.e., estimation error) and simultaneously a higher bias toward zero in $\hat{\boldsymbol{\Theta}}$. To summarize, besides reducing the estimation error, GLASSO can reflect the conditional independence structure in the underlying model.

To achieve the associated improvement with GLASSO regularization, the following points must be noted: (i) the true covariates have a conditional independence structure (i.e., the true precision matrix is actually sparse), and (ii) the empirical covariates are not severely dependent (i.e., the empirical covariance matrix \mathbf{S} and the accompanying precision matrix \mathbf{S}^{-1} should be close to diagonal). The first statement is based on the fact that GLASSO pursues a sparse structure for the precision matrix. The second statement is based on asymptotic properties such as selection consistency [37], [38], which ensure stable convergence [39].

III. GLASSO-PLDA FOR CONDITIONALLY INDEPENDENT NOISE VARIABILITY

We proposed GLASSO-PLDA in [23], which is an extension of the conventional PLDA obtained by applying GLASSO. The only difference between GLASSO-PLDA and conventional PLDA is the within-class precision matrix. In conventional PLDA, the model parameters are estimated using the empirical within-class precision matrix $\boldsymbol{\Phi}_w^{-1} \in \mathbb{R}^{D \times D}$ and between-class covariance matrix $\boldsymbol{\Phi}_b \in \mathbb{R}^{D \times D}$ (see (4)). In GLASSO-PLDA, the regularized within-class precision matrix $\hat{\boldsymbol{\Phi}}_w^{-1} \in \mathbb{R}^{D \times D}$ is used to estimate the parameters, rather than $\boldsymbol{\Phi}_w^{-1}$, as follows:

$$\hat{\boldsymbol{\Phi}}_w^{-1} \boldsymbol{\Phi}_b \mathbf{A} = \mathbf{A} \boldsymbol{\Psi} \quad (9)$$

where $\hat{\boldsymbol{\Phi}}_w^{-1}$ is the GLASSO regularization of $\boldsymbol{\Phi}_w^{-1}$, obtained as follows (see (8)):

$$\hat{\boldsymbol{\Phi}}_w^{-1} = \underset{\boldsymbol{\Theta}}{\operatorname{argmax}} \log \det \boldsymbol{\Theta} - \operatorname{tr} \boldsymbol{\Phi}_w \boldsymbol{\Theta} - \rho \|\boldsymbol{\Theta}\|_1. \quad (10)$$

Here, $\boldsymbol{\Phi}_w$ is the within-class covariance matrix estimated from the initial within-class covariance matrix $\boldsymbol{\Phi}_w^{(0)}$ (see (6)). By converting $\boldsymbol{\Phi}_w^{-1}$ into sparse $\hat{\boldsymbol{\Phi}}_w^{-1}$, GLASSO-PLDA reduces the estimation error in $\boldsymbol{\Phi}_w^{-1}$ and reflects the conditional independence structure to within-class variability.

To achieve better discriminative power with GLASSO-PLDA, two preconditions, which stem from those for GLASSO (see Section II-C-2), must be satisfied. The first

is the sparse assumption of the true within-class precision matrix. GLASSO-PLDA regularizes $\boldsymbol{\Phi}_w^{-1}$ to be sparse, based on the assumption that the true within-class precision matrix is sparse (corresponding to the true within-class variability with the conditional independence structure). This means that the performance of TD-SV in noisy environments would be improved with GLASSO-PLDA if the optima of all the variabilities (i.e., the session variability $\boldsymbol{\Sigma}_{sess}$, noise variability $\boldsymbol{\Sigma}_{noise}$, and residual variability $\boldsymbol{\Sigma}_\epsilon$; see Section II-A) in the within-class variability $\boldsymbol{\Sigma}_w$ have a conditional independence structure. As confirmed in [23], the optima of $\boldsymbol{\Sigma}_{ch}$ and $\boldsymbol{\Sigma}_\epsilon$ have the conditional independence structure. Therefore, we can determine whether the optimum of $\boldsymbol{\Sigma}_{noise}$ has a conditional independence structure based on whether GLASSO-PLDA improves the performance of TD-SV in noisy conditions.

The other precondition is that $\boldsymbol{\Phi}_w$ and the accompanying $\boldsymbol{\Phi}_w^{-1}$ are close to the diagonal, which is consistent with the empirical covariates not being significantly dependent. Some types of embeddings, such as the i-vector [40] designed with the assumption of standard normality, satisfied the precondition, although most neural network-based embeddings (e.g., d-vector [11], r-vector [15], and x-vector [41]) do not. The total covariance matrices of these embeddings are far from the diagonal, and so are the corresponding $\boldsymbol{\Phi}_w$ and $\boldsymbol{\Phi}_w^{-1}$. Because GLASSO is prone to failure in convergence with a far-from-diagonal covariance matrix, GLASSO-PLDA cannot achieve performance improvement with embeddings whose covariates are quite dependent, even if the sparse assumption of the true within-class precision matrix holds. The close-to-diagonal precondition can be satisfied by orthogonalizing embeddings using the principal component analysis (PCA) transform. The PCA transform diagonalizes the total covariance matrix, which makes $\boldsymbol{\Phi}_w$ and $\boldsymbol{\Phi}_w^{-1}$ close to the diagonal and facilitates the stable convergence of GLASSO.

IV. EXPERIMENTS

A. DATABASE

We used parts 1 and 2 of the robust speaker recognition (RSR) 2015 dataset [42], designed for TD-SV. Parts 1 and 2 have 30 kinds of short sentences (3.2 s on average including silence) and 30 kinds of keywords (1.99 s on average including silence). Both parts comprise 300 speakers and are divided into background (50 male and 47 female speakers), development (50 male and 47 female speakers), and evaluation (57 male and 49 female speakers) subsets, without speaker overlap. Each speaker spoke 30 sentences/keywords in nine different sessions, so there are 81,000 utterances in each part.

To simulate noisy conditions, we collected six types of noise sounds from Freesound [43]: *babble*, *metro*, *station*, *subway*, *bus*, and *cafe*. The noises of *babble*, *metro*, and *station* were added to all utterances of the background subset with signal-to-noise ratios (SNRs) of 0, 5, and 10 dB, which provides nine variations for each utterance. The *subway* noise with an SNR of 5 DB was added to the test utterances of the

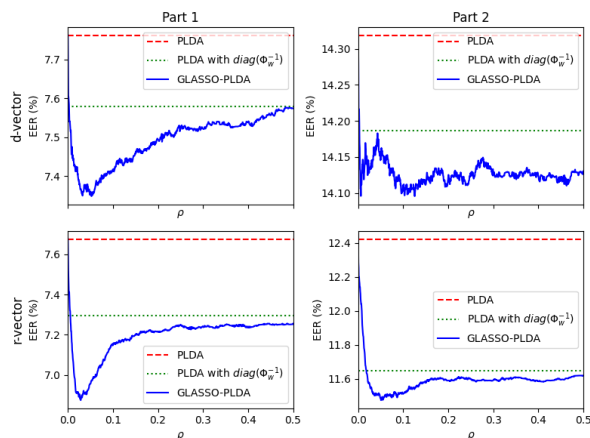


FIGURE 1. EERs of the PLDA (red dashed line), PLDA with $\text{diag}(\Phi_w^{-1})$ (green dotted line), and GLASSO-PLDA (blue solid line) on the development trials according to ρ . The top and bottom correspond to the d-vector and r-vector, respectively. The left and right correspond to the parts 1 and 2, respectively.

development subset. The *bus* and *cafe* noises were added to the test utterances of the evaluation subsets with SNRs of 0, 5, and 10 dB, which constitute six variations for each utterance.

The experiments for each part were conducted individually. The background subsets were used to train the gender-independent TD-SV system (i.e., speaker-phrase embedding extractor and the PLDA model). Development subsets were used to validate the systems in gender-independent trials. The performance of the systems was evaluated in gender-dependent trials in the evaluation subsets.

B. EXPERIMENTAL SETUP

As an acoustic feature, a sequence of 40-dimensional log mel filterbank coefficients was extracted from each utterance, with a 25 ms Hamming window at an interval of 10 ms and 512-point fast Fourier transform (FFT). For each sequence, we applied cepstral mean normalization (CMN) with a 300-frame sliding window followed by energy-based voice activity detection (VAD). A sequence of acoustic features was used to extract speaker-and-phrase embedding.

We used the d-vector [11] and r-vector [15] for speaker-and-phrase embedding. These were extracted from a model that comprised a backbone network followed by a single layer for multi-class classification. Since we empirically found that margin-based softmax loss (e.g., AAM-softmax [44]) is inferior to conventional softmax loss in our task of TD-SV, we used conventional softmax loss rather than margin-based one. For the d-vector, a 2-layer long short-term memory (LSTM) [45] was used as the backbone. The variable-length sequences in each mini-batch were handled using the method proposed in [46]. The d-vector corresponded to the weighted sum of the output sequence of the LSTM along the time axis, where the weights were computed using a self-attention mechanism [47]. For the r-vector, a 34-layer squeeze-and-excitation [48] residual network [49] (SE-ResNet) was used as the backbone, as in [15]. All input features were padded or truncated along the time axis to reserve lengths of 250 and

150 for parts 1 and 2, respectively. The r-vector corresponded to the output of the statistical pooling layer at the top of the backbone. To minimize the cross-entropy loss of each model, AMSGrad [50] was adopted as the optimizer, with a learning rate of 0.001. Each model was trained for 100 epochs using only clean utterances of the background subset, and the model with the lowest equal error rate (EER) on development trials was selected for evaluation.

The empirical statistics (i.e., the between-class covariance and within-class precision matrices) for both PLDA and GLASSO-PLDA were estimated for 10 iterations using clean and noisy (i.e., *babble*, *metro*, and *station*) utterances of the background subset. For GLASSO-PLDA, the within-class precision matrix was regularized using the GLASSO with different values of the regularization hyperparameter ρ in the range of 0 to 0.5, at intervals of 0.0005. We set the maximum number of iterations and tolerance for convergence to 100 and 0.0001, respectively.

C. RESULTS AND DISCUSSION

1) DEMONSTRATION OF THE CONDITIONAL INDEPENDENCE ASSUMPTION OF NOISE VARIABILITY

As mentioned in Section II-A, we first checked the relative scale of the noise variability to the within-class variability. Let $\Phi_{w_clean} = \Phi_{sess} + \Phi_{\epsilon}$ be the within-class covariance matrix for clean condition, and $\Phi_{w_noise} = \Phi_{sess} + \Phi_{noise} + \Phi_{\epsilon}$ be the matrix for noisy condition, where Φ_{sess} , Φ_{noise} , and Φ_{ϵ} are the covariance matrices for the session, noise, and residual variabilities, respectively. We defined the relative scale γ of Φ_{noise} as follows:

$$\gamma = 1 - \frac{\|\Phi_{w_clean}\|_1}{\|\Phi_{w_noise}\|_1} = \frac{\|\Phi_{noise}\|_1}{\|\Phi_{w_noise}\|_1}. \quad (11)$$

In Part 1, the values of γ are 0.3540 and 0.3386 for the d-vector and r-vector, respectively. In Part 2, the values of γ are 0.5086 and 0.3458 for the d-vector and r-vector, respectively. Because Φ_{noise} accounts for a significant proportion of Φ_w , it is reasonable to confirm the true structure of the noise variability by reflecting the structure to within-class variability.

Figure 1 illustrates the EERs of PLDA, PLDA with diagonalized empirical within-class precision matrix Φ_w^{-1} (referred to as $\text{diag}(\Phi_w^{-1})$), and GLASSO-PLDA on the development trials. The enrollment utterances are clean whereas the test utterances have *subway* noise, according to the regularization hyperparameter ρ . Because the GLASSO is a biased estimator that shrinks all the nonzero elements of the precision matrix to zero, the EER of the GLASSO-PLDA (blue solid line) starts at the EER of the original PLDA (red dashed line); corresponding to $\rho = 0$ and converges at the EER of the PLDA with $\text{diag}(\Phi_w^{-1})$ (green dotted line). In our experiments, Φ_w^{-1} was not diagonalized for the interval of $\rho \leq 0.5$.

Overall, the EER of GLASSO-PLDA sharply decreased initially, but then gradually increased and converged. In the

TABLE 1. EERs of the PLDA and GLASSO-PLDA with $\hat{\rho}_{dev} = 0.0525$ with the d-vector on the development and evaluation trials for Part 1.

Trial	Noise	EER (%)		
		PLDA	GLASSO-PLDA	
Dev.	Subway	5 dB	7.7620	7.3488
		0 dB	10.0841	9.7131
		5 dB	5.0867	4.9589
Male	Bus	10 dB	2.3984	2.9364
		0 dB	15.6775	14.2421
		5 dB	7.1347	6.7351
Eval.	Cafe	10 dB	3.6407	3.4742
		0 dB	10.0841	9.9731
		5 dB	5.3139	5.2139
Female	Bus	10 dB	3.0359	2.8142
		0 dB	14.8974	14.3155
		5 dB	6.9429	6.5791
		10 dB	3.3482	3.2126

TABLE 2. EERs of the PLDA and GLASSO-PLDA with $\hat{\rho}_{dev} = 0.0285$ with the r-vector on the development and evaluation trials for Part 1.

Trial	Noise	EER (%)		
		PLDA	GLASSO-PLDA	
Dev.	Subway	5 dB	7.6751	6.8754
		0 dB	11.8416	10.9962
		5 dB	6.5504	6.1456
Male	Bus	10 dB	3.9446	3.6706
		0 dB	17.5920	15.8686
		5 dB	9.6033	8.6990
Eval.	Cafe	10 dB	5.1247	4.8318
		0 dB	10.4694	9.7624
		5 dB	5.8690	5.5051
Female	Bus	10 dB	3.8814	3.6260
		0 dB	16.3031	14.8371
		5 dB	8.5846	7.6938
		10 dB	4.6410	4.3467

development trials of both parts, GLASSO-PLDA outperformed the baseline regardless of ρ . In Part 1, the optimal values of ρ were 0.0525 and 0.0285 for the d-vector and r-vector, respectively. The relative reduction in EER with the optimal GLASSO-PLDA (the lowest point on the blue solid line) against the baseline (the original PLDA) was approximately 5.32% with the d-vector (reduced 7.7620% to 7.3488%; shown in Table 1) and 10.42% with the r-vector (reduced 7.6751% to 6.8754%; shown in Table 2). In Part 2, the optimal values of ρ were 0.1255 and 0.0505 for the d-vector and r-vector, respectively. The relative reduction in EER was approximately 1.56% with the d-vector (reduced from 14.3190% to 14.0955%; presented in Table 3) and 7.65% with the r-vector (reduced from 12.4228% to 11.1730%; presented in Table 4). These results indicated that a sparse structure in the within-class precision matrix was closer to the true statistics in noisy TD-SV than in a dense structure.

However, the above results are insufficient for supporting our assumption that true noise variability has a conditional independence structure. To support this assumption, the GLASSO-PLDA (i.e., reflecting the conditional independence structure) should outperform the PLDA with $\text{diag}(\Phi_w^{-1})$ (i.e., reflecting the independence structure), unless it would be more reasonable to assume that the true noise variability has an independence structure (i.e., the true

TABLE 3. EERs of the PLDA and GLASSO-PLDA with $\hat{\rho}_{dev} = 0.1255$ with the d-vector on the development and evaluation trials for Part 2.

Trial	Noise	EER (%)		
		PLDA	GLASSO-PLDA	
Dev.	Subway	5 dB	14.3190	14.0955
		0 dB	21.6338	20.8418
		5 dB	14.4754	14.2420
Male	Bus	10 dB	9.5507	9.5270
		0 dB	23.7155	22.8050
		5 dB	15.8318	15.4996
Eval.	Cafe	10 dB	9.9860	9.9917
		0 dB	20.8142	20.1418
		5 dB	14.1615	14.1110
Female	Bus	10 dB	9.1876	9.1688
		0 dB	22.6530	22.0061
		5 dB	14.9031	14.7182
		10 dB	9.4530	9.2044

TABLE 4. EERs of the PLDA and GLASSO-PLDA with $\hat{\rho}_{dev} = 0.0505$ with the r-vector on the development and evaluation trials for Part 2.

Trial	Noise	EER (%)		
		PLDA	GLASSO-PLDA	
Dev.	Subway	5 dB	12.4228	11.1730
		0 dB	16.0855	15.3010
		5 dB	10.3531	9.8560
Male	Bus	10 dB	7.3836	6.8672
		0 dB	21.5459	19.5626
		5 dB	13.6377	12.5879
Eval.	Cafe	10 dB	9.0032	8.3181
		0 dB	15.5509	14.4583
		5 dB	10.2117	9.4526
Female	Bus	10 dB	6.9898	6.5363
		0 dB	20.2148	18.5808
		5 dB	12.7801	11.5547
		10 dB	8.2830	7.4625

noise covariance and accompanying precision matrices are diagonal), as assumed in many studies.

In practice, PLDA with $\text{diag}(\Phi_w^{-1})$ outperformed the baseline, but not as much as GLASSO-PLDA, and did not achieve the optimal EER under all the conditions. The relative EER reduction with the PLDA with $\text{diag}(\Phi_w^{-1})$ against the baseline was approximately 2.36% and 4.93% with the d-vector and r-vector, respectively, in Part 1, and 0.92% and 6.23% with the d-vector and r-vector, respectively, in Part 2. These results indicated that the optimal noise variability had a conditional independence structure, rather than an independence structure, as assumed.

2) VISUALIZATION OF THE SPARSITY IN THE OPTIMAL PRECISION MATRIX

Figures 2 and 3 depict the empirical within-class precision matrix Φ_w^{-1} and optimum of its regularization $\hat{\Phi}_w^{-1}$ in the parts 1 and 2, respectively. In both parts, Φ_w^{-1} has no zero entries. In Part 1 (Figure 2), the optimal $\hat{\Phi}_w^{-1}$ for the d-vector (top right in Figure 2; $\rho = 0.0525$) has 4,398 non-zero elements out of 261,632 ($= 512^2 - 512$) off-diagonal elements. The optimal $\hat{\Phi}_w^{-1}$ for the r-vector (bottom right in Figure 2; $\rho = 0.0285$) has 3,126 non-zero elements out of 65,280 ($= 256^2 - 256$) off-diagonal elements. In Part 2 (Figure 3), the optimal $\hat{\Phi}_w^{-1}$ for the d-vector (top right in Figure 3; $\rho = 0.1255$) has 2,650 non-zero elements out

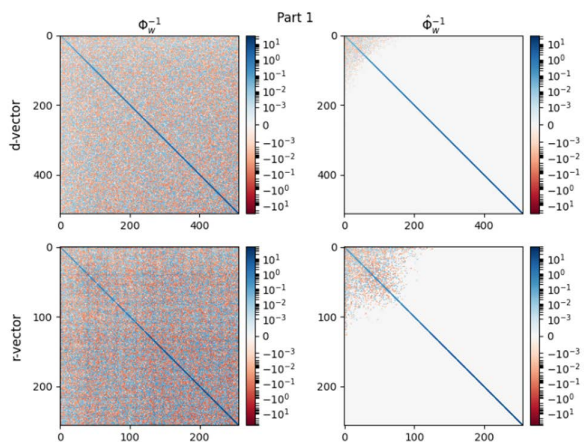


FIGURE 2. (Left) Empirical within-class precision matrix Φ_w^{-1} and (right) its regularization $\hat{\Phi}_w^{-1}$ of (top) d-vectors and (bottom) r-vectors, in Part 1.

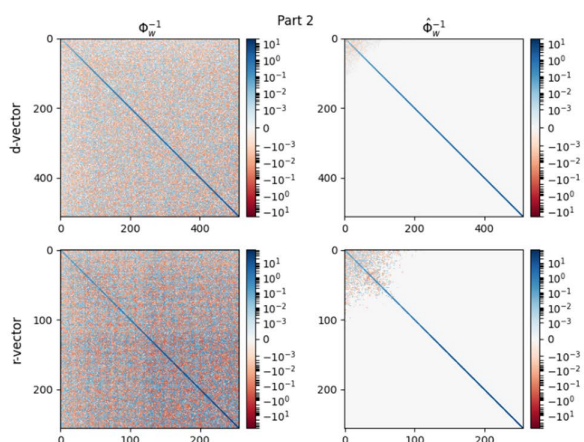


FIGURE 3. (Left) Empirical within-class precision matrix Φ_w^{-1} and (right) its regularization $\hat{\Phi}_w^{-1}$ of (top) d-vectors and (bottom) r-vectors, in Part 2.

of 261,632 off-diagonal elements. The optimal $\hat{\Phi}_w^{-1}$ for the r-vector (bottom right in Figure 3; $\rho = 0.0505$) has 1,870 nonzero elements out of 65,280 off-diagonal elements.

The optimal $\hat{\Phi}_w^{-1}$ in Part 2 (right in Figure 3) was closer to the diagonal than that in Part 1 (right in Figure 2). This is connected with the result in Figure 1 that the relative EER gap between the optimal GLASSO-PLDA and PLDA with $\text{diag}(\Phi_w^{-1})$ is lower in Part 2 (right in Figure 1) than Part 1 (left in Figure 1). However, this difference is not well explained by only the noise variability because the same kinds of noise were used for both the parts in our experiments. Given that the duration of the utterance is the main difference between parts 1 and 2, it may explain this gap. However, this is beyond the scope of this study and is left for future work.

3) EVALUATION OF THE GLASSO-PLDA ON VARIOUS NOISY CONDITIONS

Figures 4, 5, 6, and 7 illustrate the EERs of the PLDA and GLASSO-PLDA on the evaluation trials, where the enrollment utterances were clean, whereas the test utterances had one of six conditions (i.e., two noise types: *bus* and *cafe* \times three SNRs: 0, 5, and 10 dB). The black dotted vertical

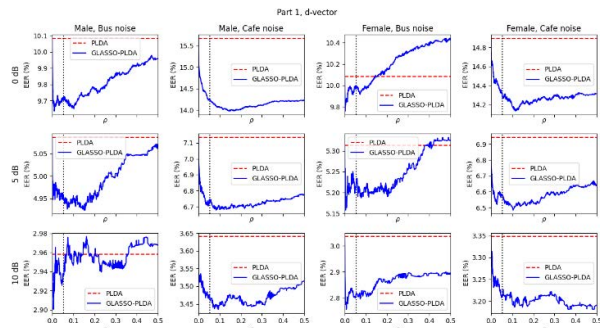


FIGURE 4. EERs of the PLDA (red dashed line) and GLASSO-PLDA (blue solid line) with the d-vector in the evaluation trials for Part 1 according to ρ .

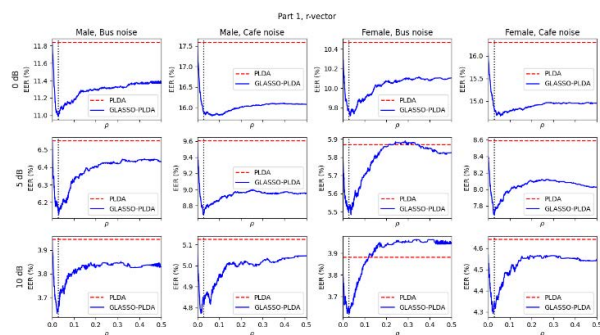


FIGURE 5. EERs of the PLDA (red dashed line) and GLASSO-PLDA (blue solid line) with the r-vector in the evaluation trials for Part 1 according to ρ .

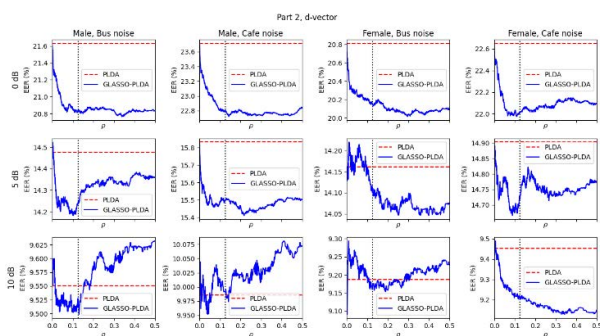


FIGURE 6. EERs of the PLDA (red dashed line) and GLASSO-PLDA (blue solid line) with the d-vector in the evaluation trials for Part 2 according to ρ .

line indicates the position of the optimal ρ for the development trials (denoted by $\hat{\rho}_{dev}$; mentioned in Section IV-C-1). Figures 4 and 5 correspond to the d-vector and r-vector, respectively, for Part 1. Figures 6 and 7 correspond to the d-vector and r-vector, respectively, for Part 2. Each row of the figures corresponds to the SNR while each column corresponds to male trials in *bus* and *cafe* noises, and female trials in *bus* and *cafe* noises, in sequence. Tables 1, 2, 3, and 4 show the EERs of PLDA and GLASSO-PLDA with $\hat{\rho}_{dev}$ for the evaluation trials, and summarize the results in Figures 4, 5, 6, and 7, respectively.

In Part 1, the average relative EER reductions in the evaluation trials were 4.1450% (0.74% to 9.16%) with the d-vector (see Table 1) and 7.5367% (5.71% to 10.38%) with

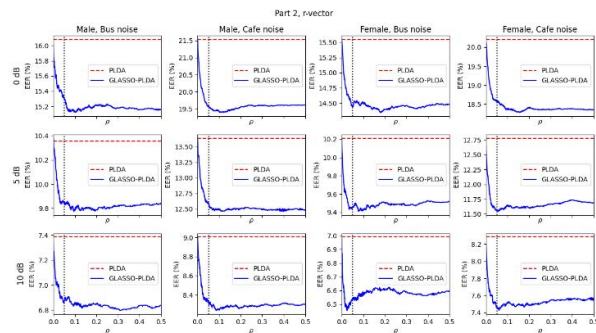


FIGURE 7. EERs of the PLDA (red dashed line) and GLASSO-PLDA (blue solid line) with the r-vector in the evaluation trials for Part 2 according to ρ .

the r-vector (see Table 2). In Part 2, these were 1.8267% (-0.06% to 3.84%) with the d-vector (see Table 3) and 7.4758% (4.8% to 9.91%) with the r-vector (see Table 4).

The EERs in the evaluation trials exhibited trends similar to those in the development trials. Except for one case, GLASSO-PLDA with $\hat{\rho}_{dev}$ also outperformed the baseline in the evaluation trials, although $\hat{\rho}_{dev}$ was not optimal for the evaluation trials. The exception case occurred where the d-vector was used in the male evaluation trials of Part 2 under *cafe* 10 dB noise (see 3rd row and 2nd column of Figure 6, and 7th row of Table 3). In this case, the EER of GLASSO-PLDA with $\hat{\rho}_{dev}$ was higher than but very close to the baseline EER. The relative difference was only 0.06% (an increase from 9.9860% to 9.9917%).

Different tendencies were observed between the d-vector and r-vector. With the r-vector, the EERs of GLASSO-PLDA (see Figures 5 and 7) converged stably. GLASSO-PLDA almost always showed lower EERs than the baseline, regardless of ρ . The EERs of the GLASSO-PLDA were slightly higher within certain intervals of ρ only in two cases, in the female trials of Part 1 under *bus* 5 dB and 10 dB noises (see 3rd column of Figure 5). The optimal ρ for the evaluation trials (denoted by $\hat{\rho}_{eval}$) was close to $\hat{\rho}_{dev}$ in all cases.

Meanwhile, with the d-vector, the EERs of GLASSO-PLDA (see Figures 4 and 6) oscillated locally, with degrees greater than those in the development trials (see the left of Figure 1). In nine out of 24 cases, the GLASSO-PLDA exhibited higher EERs than the baseline within certain intervals of ρ . EERs with $\hat{\rho}_{dev}$ and $\hat{\rho}_{eval}$ usually show non-negligible gaps. In particular, the average performance gain with the d-vector (2.9858%) is significantly lower than that with the r-vector (7.5063%). This problem is probably because of the higher number of parameters in the GLASSO-PLDA built with the d-vectors than that built with the r-vectors. The dimensionality of the d-vector (i.e., 512) was double that of the r-vector (i.e., 256) in our experiments, whereas the number of training utterances was the same. Therefore, the uncertainty of the parameters for the GLASSO-PLDA built with the d-vectors is expected to be higher in this case and appears to be primarily responsible for the problems with the d-vector mentioned above.

V. CONCLUSION

This study sheds light on a certain structure of true noise variability for noise-robust ASV systems. It assumes that true noise variability has a conditional independence structure, rather than simply an independence structure. This assumption was corroborated by evaluating the performance of GLASSO-PLDA-based TD-SV systems under various noisy conditions. The GLASSO-PLDA is an extension of the PLDA that can reflect conditional independence structure to within-class variability $\Sigma_w = \Sigma_{sess} + \Sigma_{noise} + \Sigma_\epsilon$, by making the within-class precision matrix Φ_w^{-1} sparse using GLASSO. Since the true structures of both session variability Σ_{sess} and residual variability Σ_ϵ are conditionally independent (as demonstrated in our previous work), GLASSO-PLDA outperforms PLDA under noisy environments if the true structure of noise variability Σ_{noise} is also conditionally independent. Our findings reveal that the optimal noise variability has a conditional independence structure, which is evident from the experimental results where GLASSO-PLDA surpassed both the original PLDA and PLDA with diagonal Φ_w^{-1} in the noisy TD-SV task. In conclusion, the reflection of the sparse structure on Φ_w^{-1} is informative for building noise-robust TD-SV systems.

Some issues remain to be addressed in future research. First, it is not confirmed whether a specific pattern of sparsity improves the performance. We reflected the sparse structure on Φ_w^{-1} using only GLASSO, a likelihood-based estimator. There may be alternatives to GLASSO; however, not all methods for estimating sparse precision achieve the optimal structure of Σ_{noise} . For example, matrix banding can also estimate a sparse matrix that confines non-zero elements to a diagonal band. However, we found in [23] that PLDA with banded Φ_w^{-1} was inferior to both GLASSO-PLDA and PLDA with diagonal Φ_w^{-1} . Hence, the relationship between the sparsity pattern and performance should be investigated in the future to improve performance. Second, the usefulness of reflecting a sparse structure on Φ_w^{-1} was evaluated only in PLDA-based systems. This does not indicate that reflection is possible only with PLDA. However, there is still no method that reflects the sparse structure in other types of models. Future studies should develop methods that can reflect the structure on other models. For example, an extension of neural PLDA [51] based on sparse Φ_w^{-1} , and/or a loss function for end-to-end networks that push the within-class precision of embeddings to be sparse could be investigated in the future.

REFERENCES

- [1] *Access the Google Assistant With Your Voice*. Accessed: Jul. 17, 2022. [Online]. Available: <https://support.google.com/assistant/answer/7394306>
- [2] *Talk to Bixby Using Voice Wake-up*. Accessed: Jul. 17, 2022. [Online]. Available: <https://www.samsung.com/us/support/answer/ANS00080448>
- [3] L. Lei and S. Kun, "Speaker recognition using wavelet packet entropy, l-vector, and cosine distance scoring," *J. Electr. Comput. Eng.*, vol. 2017, pp. 1–9, May 2017.
- [4] F. Zhao, H. Li, and X. Zhang, "A robust text-independent speaker verification method based on speech separation and deep speaker," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 6101–6105.

- [5] C. Zhang, M. Yu, C. Weng, and D. Yu, "Towards robust speaker verification with target speaker enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 6693–6697.
- [6] M. Jung, Y. Jung, J. Goo, and H. Kim, "Multi-task network for noise-robust keyword spotting and speaker verification using CTC-based soft VAD and global query attention," in *Proc. Interspeech*, Oct. 2020, pp. 931–935.
- [7] W. Chen, J. Huang, and T. Bocklet, "Length- and noise-aware training techniques for short-utterance speaker recognition," in *Proc. Interspeech*, Oct. 2020, pp. 3835–3839.
- [8] J. Li, J. Han, and H. Song, "Gradient regularization for noise-robust speaker verification," in *Proc. Interspeech*, Aug. 2021, pp. 1074–1078.
- [9] S. Ioffe, "Probabilistic linear discriminant analysis," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, May 2006, pp. 531–542.
- [10] S. J. D. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, Oct. 2007, pp. 1–8.
- [11] G. Heigold, I. Moreno, S. Bengio, and N. Shazeer, "End-to-end text-dependent speaker verification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 5115–5119.
- [12] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 4879–4883.
- [13] A. Hajavi and A. Etamad, "Siamese capsule network for end-to-end speaker recognition in the wild," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 7203–7207.
- [14] J. Peng, X. Qu, R. Gu, J. Wang, J. Xiao, L. Burget, and J. Černocký, "Effective phase encoding for end-to-end speaker verification," in *Proc. Interspeech*, Aug. 2021, pp. 2366–2370.
- [15] H. Zeinali, S. Wang, A. Silnova, P. Matějka, and O. Plchot, "BUT system description to VoxCeleb speaker recognition challenge 2019," 2019, *arXiv:1910.12592*.
- [16] Z. Chen and Y. Lin, "Improving X-vector and PLDA for text-dependent speaker verification," in *Proc. Interspeech*, Oct. 2020, pp. 726–730.
- [17] H. Zeinali, K. A. Lee, J. Alam, and L. Burget, "SdSV challenge 2020: Large-scale evaluation of short-duration speaker verification," in *Proc. Interspeech*, Oct. 2020, pp. 731–735.
- [18] J. Villalba, N. Chen, D. Snyder, D. Garcia-Romero, A. McCree, G. Sell, J. Borgstrom, L. P. Garcia-Perera, F. Richardson, R. Dehak, P. A. Torres-Carrasquillo, and N. Dehak, "State-of-the-art speaker recognition with neural network embeddings in NIST SRE18 and speakers in the wild evaluations," *Comput. Speech Lang.*, vol. 60, Mar. 2020, Art. no. 101026.
- [19] B. J. Borgström, "Unsupervised Bayesian adaptation of PLDA for speaker verification," in *Proc. Interspeech*, Aug. 2021, pp. 1039–1043.
- [20] M. Mohammadi and H. R. S. Mohammadi, "Weighted X-vectors for robust text-independent speaker verification with multiple enrollment utterances," *Circuits Syst. Signal Process.*, vol. 41, pp. 1–20, Jan. 2022.
- [21] L. Ferrer, M. McLaren, and N. Brümmer, "A speaker verification backend with robust performance across conditions," *Comput. Speech Lang.*, vol. 71, Jan. 2022, Art. no. 101258.
- [22] K. A. Lee, Q. Wang, and T. Koshinaka, "Xi-vector embedding for speaker recognition," *IEEE Signal Process. Lett.*, vol. 28, pp. 1385–1389, 2021.
- [23] S.-H. Yoon, J.-J. Jeon, and H.-J. Yu, "Regularized within-class precision matrix based PLDA in text-dependent speaker verification," *Appl. Sci.*, vol. 10, no. 18, p. 6571, Sep. 2020.
- [24] M. Yuan and Y. Lin, "Model selection and estimation in the Gaussian graphical model," *Biometrika*, vol. 94, no. 1, pp. 19–35, 2007.
- [25] J. Friedman, T. Hastie, and R. Tibshirani, "Sparse inverse covariance estimation with the graphical lasso," *Biostatistics*, vol. 9, no. 3, pp. 432–441, Dec. 2007.
- [26] D. M. Witten, J. H. Friedman, and N. Simon, "New insights and faster computations for the graphical lasso," *J. Comput. Graph. Statist.*, vol. 20, no. 4, pp. 892–900, 2011.
- [27] T. F. Zheng and L. Li, *Robustness-Related Issues in Speaker Recognition*. Singapore: Springer, 2017.
- [28] T. Hasan and J. H. L. Hansen, "Maximum likelihood acoustic factor analysis models for robust speaker verification in noise," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 2, pp. 381–391, Feb. 2014.
- [29] P. Kenny, "Bayesian speaker verification with heavy-tailed priors," in *Proc. Odyssey Speaker Lang. Recognit. Workshop*, 2010, p. 14. [Online]. Available: https://www.isca-speech.org/archive_open/odyssey_2010/od10_014.html
- [30] N. Brummer and E. de Villiers, "The speaker partitioning problem," in *Proc. Odyssey Speaker Lang. Recognit. Workshop*, 2010.
- [31] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of I-vector length normalization in speaker recognition systems," in *Proc. Interspeech*, Aug. 2011, pp. 249–252.
- [32] M. Ravanelli, T. Parcollet, and Y. Bengio, "The PyTorch-Kaldi speech recognition toolkit," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 6465–6469.
- [33] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via EM algorithm," *J. Roy. Stat. Soc., B, Methodol.*, vol. 39, no. 1, pp. 1–22, 1977.
- [34] H. Rue and L. Held, *Gaussian Markov Random Fields: Theory and Applications*. Boca Raton, FL, USA: CRC Press, 2005.
- [35] K. P. Murphy, "Undirected graphical models (Markov random fields)," in *Machine Learning: A Probabilistic Perspective*. Cambridge, MA, USA: MIT Press, 2012, pp. 661–705.
- [36] T. Hastie, R. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, vol. 2. New York, NY, USA: Springer, 2009. [Online]. Available: https://hastie.su.domains/ElemStatLearn/printings/ESLII_print12_toc.pdf
- [37] N. Meinshausen and P. Bühlmann, "High-dimensional graphs and variable selection with the Lasso," *Ann. Statist.*, vol. 34, no. 3, pp. 1436–1462, 2006.
- [38] P. Zhao and B. Yu, "On model selection consistency of Lasso," *J. Mach. Learn. Res.*, vol. 7, pp. 2541–2563, Nov. 2006.
- [39] J. C. Spall, "Cyclic seesaw process for optimization and identification," *J. Optim. Theory Appl.*, vol. 154, no. 1, pp. 187–208, Jul. 2012.
- [40] N. Dehak, P. J. Kenny, R. Dehak, D. Pierre, and O. Pierre, "Front-end factor analysis for speaker verification," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 4, pp. 788–798, May 2011.
- [41] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 5329–5333.
- [42] A. Larcher, K. A. Lee, B. Ma, and H. Li, "Text-dependent speaker verification: Classifiers, databases and RSR2015," *Speech Commun.*, vol. 60, pp. 56–77, May 2014.
- [43] E. Fonseca, J. Pons, X. Favory, F. Font, D. Bogdanov, A. Ferraro, S. Oramas, A. Porter, and X. Serra, "Freesound datasets: A platform for the creation of open audio datasets," in *Proc. Int. Soc. Music Inf. Retr. (ISMIR)*, Oct. 2017, pp. 486–493.
- [44] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4690–4699.
- [45] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [46] S.-H. Yoon and H.-J. Yu, "A simple distortion-free method to handle variable length sequences for recurrent neural networks in text dependent speaker verification," *Appl. Sci.*, vol. 10, no. 12, p. 4092, Jun. 2020.
- [47] Z. Lin, M. Feng, C. S. Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio, "A structured self-attentive sentence embedding," in *Proc. Int. Conf. Learn. Represent.*, Apr. 2017, pp. 1–15.
- [48] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.
- [49] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
- [50] S. J. Reddi, S. Kale, and S. Kumar, "On the convergence of ADAM and beyond," in *Proc. IEEE Int. Conf. Learn. Represent. (ICLR)*, May 2018, pp. 1–23.
- [51] S. Ramoji, P. Krishnan, and S. Ganapathy, "NPLDA: A deep neural PLDA model for speaker verification," in *Proc. Speaker Lang. Recognit. Workshop*, Nov. 2020, pp. 202–209.



SUNGHYUN YOON (Member, IEEE) received the B.S. and Ph.D. degrees in computer science from the University of Seoul, South Korea, in 2015 and 2020, respectively. Since 2021, he has been an Assistant Professor with the Department of Artificial Intelligence, Kongju National University. His research interests include speaker recognition, spoofed speech detection, and machine learning for time series analysis.