

Cosine Scoring with Uncertainty for Neural Speaker Embedding

Qiongqiong Wang, *Member, IEEE*, Kong Aik Lee, *Senior Member, IEEE*

Abstract—Uncertainty modeling in speaker representation aims to learn the variability present in speech utterances. While the conventional cosine-scoring is computationally efficient and prevalent in speaker recognition, it lacks the capability to handle uncertainty. To address this challenge, this paper proposes an approach for estimating uncertainty at the speaker embedding front-end and propagating it to the cosine scoring back-end. Experiments conducted on the VoxCeleb and SITW datasets confirmed the efficacy of the proposed method in handling uncertainty arising from embedding estimation. It achieved improvement with 8.5% and 9.8% average reductions in EER and minDCF compared to the conventional cosine similarity. It is also computationally efficient in practice.

Index Terms—Cosine similarity, scoring, speaker embeddings, speaker recognition, uncertainty propagation

I. INTRODUCTION

SPEAKER recognition is the automated process of recognizing who is speaking by analyzing the speech and extracting speaker-specific information [1]. Modern speaker recognition systems consist of a front-end that produces speaker embeddings followed by a scoring back-end [2]–[6]. Speaker embeddings are fixed-length continuous-value representations of input sequences of acoustic feature vectors [7]. Representing speech utterances as embedding simplifies speaker recognition tasks because simple geometric operations are easily applicable in the embedding space.

A speaker-embedding neural network consists of three main modules [8]: (1) a frame-level feature encoder, (2) a temporal aggregation pooling layer that converts frame-level features into a fixed-dimensional condensed vector, and (3) a decoder that produces the embedding representations. During training, a softmax function is used at the output layer, and a cross-entropy loss is used for network optimization. A substantial amount of work has been reported on neural network architectures that produce embedding vectors with improved speaker representations [8]–[14]. Commonly used scoring back-ends include the *probabilistic linear discriminant analysis* (PLDA) [15]–[17] and *cosine similarity* measure.

Speech utterances exhibit intrinsic variability owing to the physiological nature of the vocal apparatus and psychological states such as emotion and mental health, as well as extrinsic variability from background noises, channel distortions, and duration [8]. Despite being speaker-independent, the extrinsic

variabilities do affect how representative the extracted embedding vectors are of the speaker’s voice [18]. For example, it has been noted that speech utterances of shorter length generate less dependable embeddings, leading to poor performance in speaker verification [19]. Embeddings of the same utterance extracted using different networks may also differ in terms of their representative power. It depends on the conditions and amount of training data, the network capacity, and so on. Therefore, measuring the precision, or equivalently, the uncertainty of the embeddings is essential.

Considering uncertainty in both front-end embedding extraction and back-end scoring has been found effective. For the front-end, an auxiliary neural network has been employed to predict the uncertainty by utilizing the output from the statistics pooling layer of the original x-vector [20]. A posterior inference pooling has been proposed in a speaker-embedding neural network, referred to as the xi-vector network [8], [21]. It predicts the uncertainty of each frame in the input and subsequently propagates it to the embedding estimation. In [22], an uncertainty-propagated PLDA (UP-PLDA) introduces an increase in the within-speaker covariance by the posterior covariance corresponding to the embedding uncertainty. It improves the robustness against local and random perturbation, an inherent property of speech utterances.

In addition to the speaker-embedding network architectures, advances in loss functions have brought significant progress in speaker representation learning [23]–[26]. To increase the compactness within the same speaker class, various techniques incorporate margin penalties to the conventional softmax activation function following a cross-entropy loss [27]–[29]. The use of large-margin embeddings amplifies speaker-discriminative capabilities, leading to a shift from the PLDA back-end to a more straightforward cosine similarity measure. While cosine scoring is computationally efficient and does not necessitate training, it lacks the capability to manage uncertainty [24], [25], [30], [31].

This paper explores methods for uncertainty propagation from the embeddings to the back-end of speaker recognition tasks, specifically, cosine scoring. To the best of our knowledge, this is the first work taking uncertainty into consideration for cosine scoring.

II. SPEAKER EMBEDDINGS AND COSINE SCORING

The estimation of the speaker embedding is influenced by the variability in utterances and the property of embedding extraction networks. Hence, it is essential to measure the uncertainty, or precision, of the embeddings.

Qiongqiong Wang is with the Institute for Infocomm Research (I²R), Agency for Science, Technology and Research (A*STAR), Singapore 138632 (e-mail: wang_qiongqiong@i2r.a-star.edu.sg).

Kong Aik Lee is with the Dept. of Electrical and Electronic Engineering, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong (email: kong-aik.lee@polyu.edu.hk)

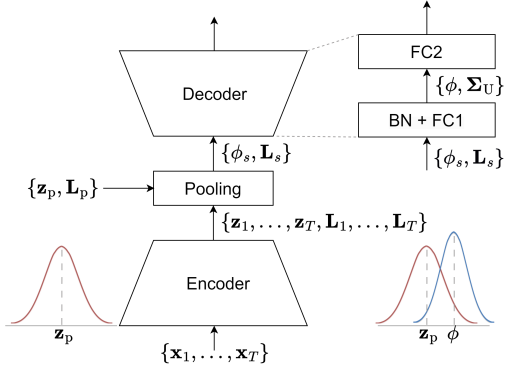


Fig. 1: A speaker-embedding neural network with uncertainty propagation. “BN” and “FC” are shorthand for batch normalization and a fully-connected layer, respectively. The Gaussian distributions are to illustrate that the posterior distribution of an embedding differs from the prior distribution.

A. Uncertainty of neural speaker embedding

As shown in Fig. 1, we assume that an encoder maps an input sequence $\{x_1, \dots, x_T\}$ of T frames to another sequence $\{z_1, \dots, z_T\}$ with precision $\{L_1^{-1}, \dots, L_T^{-1}\}$. The encoder is commonly known as the backbone neural network. Popular architectures are ResNet34 [32], ECAPA-TDNN [30], TDNN [9], and so on. By defining the temporal pooling as a Gaussian inference process [8], a point estimate ϕ_s and its precision L_s are obtained as

$$\phi_s = L_s^{-1} \left(\sum_{t=1}^T L_t z_t + L_p z_p \right) \quad (1)$$

$$L_s = \sum_{t=1}^T L_t + L_p \quad (2)$$

where μ_p and L_p represent the prior mean and precision, respectively. The posterior precision L_s measures the uncertainty associated with the point estimate in (1). Referring to the black arrows in Fig. 1, the processes are summarized as follows:

$$\{z_1, \dots, z_T, L_1^{-1}, \dots, L_T^{-1}\} \leftarrow f_{\text{enc}}(x_1, \dots, x_T) \quad (3)$$

$$\{\phi_s, L_s\} \leftarrow g_{\text{pool}}(z_1, \dots, z_T, L_1^{-1}, \dots, L_T^{-1}) \quad (4)$$

While the posterior precision L_s is not directly used in the network training [8], it can be propagated to the speaker-embedding space through the feed-forward layers in a similar manner as the posterior mean (see Fig. 1), as follows:

$$\Sigma_U \leftarrow W \Sigma_{s, \text{BN}} W^T \quad (5)$$

$$\phi \leftarrow W \phi_{s, \text{BN}} + b \quad (6)$$

where W denotes the weight, and b represents the bias of the first FC layer. The results of batch normalization of $\{\phi_s, L_s^{-1}\}$ are represented as $\{\phi_{s, \text{BN}}, \Sigma_{s, \text{BN}}\}$. More details can be found in [22]. The network outputs the speaker embeddings represented by a distribution with the mean ϕ and a variance Σ_U that is referred to as the uncertainty of an embedding.

To understand the scale of the embedding uncertainty Σ_U , we plot the diagonal values of the average uncertainty in

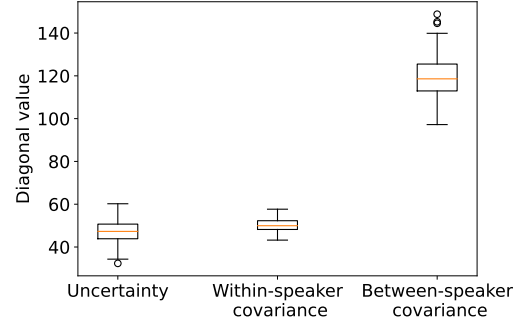


Fig. 2: Boxplots of the diagonal values of the average uncertainty estimation, and the within- and between-speaker covariance matrices estimated on the VoxCeleb2 development set. The off-diagonal values are disregarded.

Fig. 2 alongside that of between- and within-speaker covariance matrices computed from the embedding means. The off-diagonal values are discarded. The embeddings are extracted from the VoxCeleb2 development dataset [11] using the xi-vector network [8]. Notably, the between-speaker covariance exhibits significantly higher values than the within-speaker covariance, indicating discriminativeness across speakers and compactness within the same speaker. This aligns with our findings in [31]. The notable overlap between uncertainty and within-speaker covariance values suggests considering uncertainty in computing within-speaker covariance [22]. This highlights the need to propagate uncertainty for improving accuracy in back-end scoring.

B. Cosine scoring

The objective of a speaker verification task is to verify whether the testing speech is from the enrolled speaker. To achieve this, a similarity is computed between the embeddings of the enrollment speech (ϕ_e) and the test speech (ϕ_t). One widely used approach is the cosine similarity scoring, known for its computational efficiency. It computes the cosine of the angle between the two embedding vectors, functioning as an indicator of their similarity [24], [25], [30], [31], as follows

$$s(\phi_e, \phi_t) = \frac{\langle \phi_e, \phi_t \rangle}{\|\phi_e\| \|\phi_t\|} \quad (7)$$

where $\langle \cdot \rangle$ represents the inner product between two vectors, and $\|\cdot\|$ signifies l_2 normalization. Note that cosine scoring in (7) has no ability to handle uncertainty.

III. UNCERTAINTY-AWARE COSINE SCORING

Cosine scoring (7) measures the similarity between embeddings on a unit circle. It calculates the Euclidean lengths of the enrollment and test embeddings in the denominator, treating all dimensions equally, as follows:

$$s(\phi_e, \phi_t) = \frac{\phi_e^T \phi_t}{\sqrt{\phi_e^T \mathbf{I} \phi_e} \sqrt{\phi_t^T \mathbf{I} \phi_t}} \quad (8)$$

With the consideration of uncertainty in the embedding representations, it's evident that each dimension of the embeddings is no more independent or identically distributed. Therefore, they should be treated differently with the reference

of covariance Σ_U in (5). For example, the same distance in the dimensions with larger variances should be counted as less contribution in the “length calculation” process, and the opposite for that in the dimensions with smaller variances.

As uncertainty is additive, the uncertainty posterior covariance of the embedding can be added to the identity matrix with a factor of $\frac{1}{d}$, as follows:

$$s_{UP} = \frac{\phi_e^T \phi_t}{\sqrt{\phi_e^T (\mathbf{I} + \frac{1}{d} \Sigma_{Ue})^{-1} \phi_e} \sqrt{\phi_t^T (\mathbf{I} + \frac{1}{d} \Sigma_{Ut})^{-1} \phi_t}} \quad (9)$$

Dividing the uncertainty covariance $\{\Sigma_{Ue}, \Sigma_{Ut}\}$ by the dimensionality of the embedding vectors d keeps the scale of the denominator the same. The uncertainty-aware cosine scoring propagates the uncertainty of speaker embedding extraction to the back-end scoring.

The uncertainty-propagated cosine (UP-Cos) scoring in (9) can be interpreted as an inner production between the enrollment and test embeddings normalized by the Mahalanobis distance lengths of the two embeddings. Mahalanobis distance [33] is a multivariate statistical distance measure that scales the contribution of individual variables to the distance value according to their variance. We define the Mahalanobis distance length $\|\phi\|_{ML}$ of an embedding as the Mahalanobis distance between the coordinate origin and the endpoint of the embedding vector, as follows:

$$\|\phi\|_{ML} = \sqrt{(\phi - \mathbf{0})^T \Sigma^{-1} (\phi - \mathbf{0})} \quad (10)$$

In the uncertainty-aware cosine scoring (9), the variance Σ for Mahalanobis distance is

$$\Sigma_e = \mathbf{I} + \frac{1}{d} \Sigma_{Ue} \quad \Sigma_t = \mathbf{I} + \frac{1}{d} \Sigma_{Ut} \quad (11)$$

respectively. The scoring can be re-written as

$$s_{UP} = \frac{\langle \phi_e, \phi_t \rangle}{\sqrt{\phi_e^T \Sigma_e^{-1} \phi_e} \sqrt{\phi_t^T \Sigma_t^{-1} \phi_t}} \quad (12)$$

In the case where the uncertainty is not considered, i.e., $\Sigma_U = \mathbf{0}$ and $\Sigma_e = \Sigma_t = \mathbf{I}$, the Mahalanobis length (10) reduces to the Euclidean length. In addition, the uncertainty-aware cosine similarity score s_{UP} in (9) becomes equivalent to the original cosine similarity score as in (7).

Total covariance captures the global covariance of embeddings and can be an indicator for the contribution of each dimension to the Mahalanobis distance. When considering embedding uncertainty, the total covariance attributed to an individual embedding is computed as a sum of the uncertainty matrix and the total covariance Σ_{tot} derived from embedding means (see UP-Cos 2 in Table I). UP-Cos 3 and UP-Cos 4 are the variations of the previous UP-Cos scorings considering the possible correlation between the enroll and test embeddings.

The relation between the uncertainty-aware cosine similarity and the original cosine similarity is shown as follows:

$$s_{UP}(\phi_e, \phi_t) = \alpha_e \alpha_t \cos(\phi_e, \phi_t) \quad (13)$$

where α_e and α_t are

$$\alpha_e = \left(\frac{\phi_e^T \phi_e}{\phi_e^T \Sigma_e^{-1} \phi_e} \right)^{\frac{1}{2}} \quad \alpha_t = \left(\frac{\phi_t^T \phi_t}{\phi_t^T \Sigma_t^{-1} \phi_t} \right)^{\frac{1}{2}} \quad (14)$$

TABLE I: Four UP-Cos methods

Scoring	Enrol Σ	Test Σ
UP-Cos 1	$d^{-1} \Sigma_{Ue} + \mathbf{I}$	$d^{-1} \Sigma_{Ut} + \mathbf{I}$
UP-Cos 2	$d^{-1} (\Sigma_{Ue} + \Sigma_{tot})$	$d^{-1} (\Sigma_{Ut} + \Sigma_{tot})$
UP-Cos 3	$d^{-1} (\Sigma_{Ue} + \Sigma_{Ut}) + \mathbf{I}$	
UP-Cos 4	$d^{-1} (\Sigma_{Ue} + \Sigma_{Ut} + \Sigma_{tot})$	

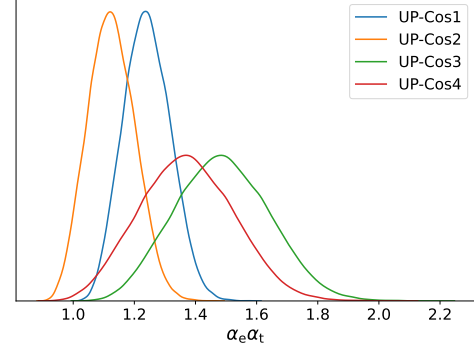


Fig. 3: Distribution of the product $\alpha_e \alpha_t$ when using the four methods of UP-Cos shown in Table I. The investigation is done on the VoxCeleb1-O test dataset.

Therefore, the uncertainty of enrollment and test speaker embedding representations is reflected in the effect of scaling in the scores. The scaling factors α_e and α_t are scalars calculated from a function of the embedding vector ϕ and uncertainty covariance Σ_U . It is worth noting that the uncertainty influences the range of UP-Cos scores s_{UP} . Due to the factor $\alpha_e \alpha_t$ shown in Fig. 3, it no longer falls within the range of $[-1, 1]$ as with cosine similarity scores (7).

IV. EXPERIMENTS

A. Experimental settings

We evaluate the proposed methods on both the VoxCeleb1 [34] and the Speaker in the Wild (SITW) core-core eval sets [35]. For the former, we test two trial lists: the original test set Vox1-O and the hard test set Vox1-H. We use the development set of the VoxCeleb2 dataset [11] as the training data. There are no speakers common to both our training and evaluation sets. Data augmentation is utilized to increase training data diversity and quantity. Techniques involve speed perturbation [36], random drops of audio chunks and frequency bands [37], environmental corruptions using room impulse responses, and adding noise [38].

For the front-end, we extract 80-dimensional filterbank features from the utterance audio and input them to speaker-embedding networks. We employ the ECAPA-TDNN backbone with AAM-Softmax cross-entropy loss [30]. Attentive statistics pooling and posterior inference pooling [8] are used in the networks to produce the two types of embedding vectors differing in whether to take uncertainty under consideration in embedding extraction, respectively, referred to as Emb and UP-Emb in the experiments. For UP-Emb, frame-wise precision $\mathbf{L}_t$ is assumed to be diagonal [8], resulting in embeddings' posterior covariance Σ_U having significantly larger diagonal values than off-diagonal ones. We utilize them as

TABLE II: Performance of the proposed uncertainty-propagated scoring and baselines on Vox1-O, Vox1-H, and SITW datasets. Emb and UP-Emb represent the embeddings extracted from the ECAPA-TDNN network with attentive statistics pooling and posterior inference pooling, respectively. Four UP-Cos scoring methods are evaluated with UP-Emb. Results are shown as EER/minDCF.

System	Vox1-O	Vox1-H	SITW
Emb+Cos	1.22/0.176	2.46/0.235	1.79/0.170
Emb+PLDA	1.23/0.173	2.46/0.236	1.69/0.172
UP-Emb+Cos	1.23/0.145	2.42/0.234	1.75/0.168
UP-Emb+PLDA	1.23/0.141	2.43/0.235	1.77/0.168
UP-Emb+UP-Cos 1	1.05/ 0.115	2.20/ 0.219	1.72/ 0.164
UP-Emb+UP-Cos 2	1.04/0.117	2.20/0.221	1.72/ 0.164
UP-Emb+UP-Cos 3	1.01/0.127	2.17/0.233	1.67/ 0.164
UP-Emb+UP-Cos 4	0.99 /0.131	2.17 /0.240	1.64/0.172
UP-Emb+UP-PLDA	1.01/0.124	2.18/0.224	1.59 /0.166

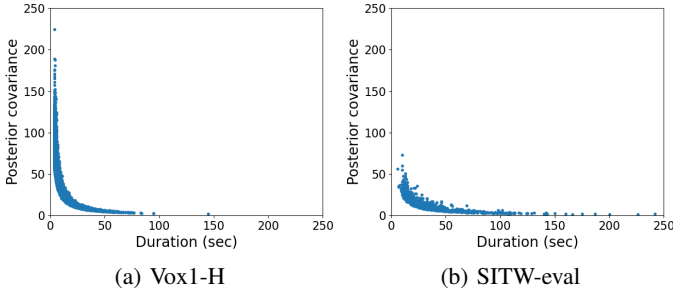


Fig. 4: Scatter plot of average posterior covariance versus utterance duration for Vox1-H and SITW-eval datasets.

diagonal matrices, discarding off-diagonal values. We use the SpeechBrain open-source toolkit [39] for the implementation of the front-end and extraction of embeddings.

The two baselines are the cosine scoring (Cos) on both types of embeddings. For the proposed scoring with uncertainty (UP-Cos), we use the UP-Emb and propagate the frame-wise uncertainty to the embedding space following II-A. We evaluate four variations of UP-Cos in Table I. For UP-Cos 3 and 4, the total covariance is pre-computed using the same training data and subsequently diagonalized for convenience and alignment with uncertainty covariance matrices Σ_U . Additionally, we employ PLDA and uncertainty-propagated PLDA (UP-PLDA) [22] for comparison. Length scaling (LS) and uncertainty-propagated LS (UP-LS) pre-processing techniques are applied to PLDA and UP-PLDA, respectively, as recommended in [22]. The training is conducted on the same training data with concatenations of segments belonging to the same session.

B. Results and analysis

We first compare the four systems of two back-ends (Cos and PLDA) with the conventional embeddings (Emb) and uncertainty-aware embeddings (UP-Emb), as shown in the first four rows in Table II. We observe that with the Emb, the use of Cos and PLDA give similar performance. It is expected since a large margin softmax is used in the network training [31]. In both Cos and PLDA systems, UP-Emb consistently shows superior or comparable performance in comparison to Emb. It is confirmed that considering uncertainty in deriving speaker embedding extraction helps extract better speaker embeddings for speaker recognition tasks, while the effect is limited.

We next propagate the uncertainty, represented as posterior covariance, to the speaker embedding space and evaluate the proposed cosine scoring with uncertainty. As shown in Table II, all of the UP-Cos methods outperform the Cos baseline in terms of EER over all three evaluation sets, and UP-Cos 1, 2, and 3 reduce minDCF in addition. It indicates that further propagating the uncertainty to the back-end is helpful for speaker recognition tasks, and the proposed cosine scoring with uncertainty has the capability to handle the uncertainty from embedding estimation in cosine similarity. According to Table I, Σ for UP-Cos 1 and 2 are dependent on the embedding only, while for UP-Cos 3 and 4, Σ depends on both the enrollment and test embeddings in a trial. Therefore, UP-Cos 1 and 2 are more computationally efficient in the testing phase. In addition, the total covariance in UP-Cos 2 and 4 is calculated, in advance, from a dataset other than the evaluation set. Considering the computational cost, UP-Cos 1 is the most recommended for practical use. It achieved consistent improvement with average 8.5% and 9.8% reductions in EER and minDCF, respectively compared with the baseline.

We further compare the use of uncertainty in cosine scoring and PLDA scoring in Table II. We observe a consistency with [22] that UP-PLDA outperforms PLDA when using UP-Emb. Compared with UP-Cos 1, UP-PLDA achieves better EERs but worse minDCF in Vox1 sets while the opposite in the SITW set. Compared to UP-PLDA, UP-Cos takes much less computational cost.

Finally, we investigate the correlation between uncertainty with speech duration. By averaging the diagonal values of each posterior covariance, across the 192 dimensions, we generate scalars. They are then plotted against the utterance duration in Fig. 4. A negative correlation is observed between the two variables. It is consistent with our expectation that embeddings extracted from a shorter utterance are less reliable, i.e., larger uncertainty. Therefore, it is confirmed that the posterior covariance measures the uncertainty, and is related to its nature. The posterior covariance of SITW-eval exhibits higher values compared to that of the Vox1-H for the same speech duration, possibly indicating a domain mismatch from the training data. Consequently, the extracted embeddings show increased uncertainties.

V. CONCLUSION

We have shown the importance of taking uncertainty into consideration in speaker recognition because of speaker-independent variabilities exhibited in speech utterances and the extraction networks. On the basis of neural speaker embeddings, we show how to propagate the posterior covariance (i.e., uncertainty) to the back-end. We also propose uncertainty-aware scoring. To the best of our knowledge, it is the first work of non-parametric scoring with uncertainty for speaker recognition. Experiments confirmed the efficacy of the proposed scoring and showed consistent improvement in performance on the VoxCeleb and the SITW datasets. It achieved consistent improvement with 9.5%/15.0% average reductions in EER/minDCF compared to the setting in which no uncertainty is considered and is 8.5%/9.8% compared to the cosine similarity following certainty-aware embeddings.

REFERENCES

- [1] Z. Bai and X.-L. Zhang, "Speaker recognition based on deep learning: An overview," in *Neural Network*, vol. 140, 2021, pp. 65–99.
- [2] P. Matejka, O. Plchot, O. Glembek, L. Burget, J. Rohdin *et al.*, "13 years of speaker recognition research at BUT, with longitudinal analysis of NIST SRE," in *Computer Speech & Language*, vol. 63, 2020, p. 101035.
- [3] J. Villalba, N. Chen, D. Snyder, D. Garcia-Romero, A. McCree *et al.*, "State-of-the-art speaker recognition with neural network embeddings in NIST SRE18 and Speakers in the Wild evaluations," in *Computer Speech & Language*, vol. 60, 2020, p. 101026.
- [4] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "Voxceleb: Largescale speaker verification in the wild," in *Computer Speech & Language*, vol. 60, 2020, p. 101027.
- [5] K. A. Lee, H. Yamamoto, K. Okabe, Q. Wang, L. Guo *et al.*, "NEC-TT system for mixed-bandwidth and multi-domain speaker recognition," in *Computer Speech & Language*, vol. 61, 2020, p. 101033.
- [6] L. Ferrer, M. McLaren, and N. Brummer, "A speaker verification back-end with robust performance across conditions," in *Computer Speech & Language*, vol. 71, 2022, p. 101258.
- [7] N. Brummer, A. Silnova, L. Burget, and T. Stafylakis, "Gaussian meta-embeddings for efficient scoring of a heavy-tailed PLDA model," in *Proc. Odyssey*, 2018, pp. 349–356.
- [8] K. A. Lee, Q. Wang, and T. Koshinaka, "Xi-vector embedding for speaker recognition," in *IEEE Signal Processing Letters*, 2021, pp. 1385–1389.
- [9] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. IEEE ICASSP*, 2018, pp. 5329–5333.
- [10] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive statistics pooling for deep speaker embedding," in *Proc. Interspeech*, 2018, pp. 2252–2256.
- [11] J. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep speaker recognition," in *Proc. Interspeech*, 2018, pp. 1086–1090.
- [12] J. Rohdin, A. Silnova, M. Diez, O. Plchot, P. Matejka, L. Burget, and O. Glembek, "End-to-end DNN based text-independent speaker recognition for long and short utterances," in *Computer Speech & Language*, 2020, pp. 22–35.
- [13] T. Liu, R. K. Das, K. A. Lee, and H. Li, "MFA: TDNN with multi-scale frequency-channel attention for text-independent speaker verification with short utterances," in *Proc. IEEE ICASSP*, 2022, pp. 7517–7521.
- [14] T. Liu, K. A. Lee, Q. Wang, and H. Li, "Golden gemini is all you need: Finding the sweet spots for speaker verification," *arXiv preprint arXiv:2312.03620*, 2023.
- [15] S. Ioffe, "Probabilistic linear discriminant analysis," in *Proc. ECCV*, 2006, pp. 531–542.
- [16] S. Prince and J. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Proc. ICCV*, 2007, pp. 1–8.
- [17] Q. Wang, K. Okabe, K. A. Lee, and T. Koshinaka, "Generalized domain adaptation framework for parametric back-end in speaker recognition," in *IEEE Transactions on Information Forensics and Security*, vol. 18, 2023, pp. 3936–3947.
- [18] K. Takahashi and T. Murakami, "A measure of information gained through biometric systems," in *Image and Vision Computing*, vol. 32, 2014, pp. 1194–1203.
- [19] H. Zeinali, K. A. Lee, J. Alam, and L. Burget, "SdSV challenge 2020: Large-scale evaluation of short-duration speaker verification," in *Proc. Interspeech*, 2020, pp. 731–735.
- [20] A. Silnova, N. Brummer, J. Rohdin, T. Stafylakis, and L. Burget, "Probabilistic embeddings for speaker diarization," in *Proc. Odyssey*, 2020, pp. 24–31.
- [21] T. Liu, K. A. Lee, Q. Wang, and H. Li, "Disentangling voice and content with self-supervision for speaker recognition," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.
- [22] Q. Wang, K. A. Lee, and T. Liu, "Incorporating uncertainty from speaker embedding estimation to speaker verification," in *Proc. IEEE ICASSP*, 2023.
- [23] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," in *Proc. IEEE ICASSP*, 2018, pp. 4879–4883.
- [24] Y. Liu, L. He, and J. Liu, "Large margin softmax loss for speaker verification," in *Proc. Interspeech*, 2019, pp. 2873–2877.
- [25] D. Zhou, L. Wang, K. A. Lee, Y. Wu, M. Liu, J. Dang, and J. Wei, "Dynamic margin softmax loss for speaker verification," in *Proc. Interspeech*, 2020, pp. 3800–3804.
- [26] Z. Li and M.-W. Mak, "Speaker representation learning via contrastive loss with maximal speaker separability," in *Proc. APSIPA*, 2022.
- [27] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "Cosface: Large margin cosine loss for deep face recognition," in *IEEE CVPR*, 2018, pp. 5265–5274.
- [28] F. Wang, J. Cheng, W. Liu, and H. Liu, "Additive margin softmax for face verification," in *IEEE Signal Processing Letters*, vol. 25, 2018, pp. 926–930.
- [29] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *IEEE CVPR*, 2019, pp. 4690–4699.
- [30] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in *Proc. Interspeech*, 2020, pp. 3830–3834.
- [31] Q. Wang, K. A. Lee, and T. Liu, "Scoring of large-margin embeddings for speaker verification: Cosine or PLDA?" in *Proc. Interspeech*, 2022, pp. 600–604.
- [32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE CVPR*, 2016.
- [33] P. C. Mahalanobis, "On the generalized distance in statistics," in *Proceedings of the National Institute of Sciences (Calcutta)*, vol. 2, 1936, pp. 49–55.
- [34] A. Nagrani, J. Chung, and A. Zisserman, "VoxCeleb: a large-scale speaker identification dataset," in *Proc. Interspeech*, 2017, pp. 2616–2620.
- [35] M. McLaren, L. Ferrer, D. Castan, and A. Lawson, "The speakers in the wild (SITW) speaker recognition database," in *Proc. Interspeech*, 2016, pp. 818–822.
- [36] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Proc. Interspeech*, 2015, pp. 3586–3589.
- [37] D. Park, W. Chan, Y. Zhang, C. Chiu, B. Zoph, E. Cubuk, and Q. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Proc. Interspeech*, 2019, pp. 2613–2617.
- [38] T. Ko, V. Peddinti, D. Povey, M. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *Proc. IEEE ICASSP*, 2017, pp. 5220–5224.
- [39] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell *et al.*, "SpeechBrain: A general-purpose speech toolkit," in *arXiv:2106.04624*, 2021.