Interpolating Speaker Identities in Embedding Space for Data Expansion

Tianchi Liu*, Ruijie Tao*, Qiongqiong Wang†, Yidi Jiang*, Hardik B. Sailor†, Ke Zhang‡, Jingru Lin*, Haizhou Li‡*
* Department of Electrical and Computer Engineering, National University of Singapore, Singapore
† Institute for Infocomm Research (I²R), Agency for Science, Technology and Research (A*STAR), Singapore
‡ SRIBD, School of Artificial Intelligence, The Chinese University of Hong Kong, Shenzhen, China
E-mail: tianchi_liu@u.nus.edu

***Abstract*—The success of deep learning-based speaker verification systems is largely attributed to access to large-scale and diverse speaker identity data. However, collecting data from more identities is expensive, challenging, and often limited by privacy concerns. To address this limitation, we propose INSIDE (Interpolating Speaker Identities in Embedding Space), a novel data expansion method that synthesizes new speaker identities by interpolating between existing speaker embeddings. Specifically, we select pairs of nearby speaker embeddings from a pretrained speaker embedding space and compute intermediate embeddings using spherical linear interpolation. These interpolated embeddings are then fed to a text-to-speech system to generate corresponding speech waveforms. The resulting data is combined with the original dataset to train downstream models. Experiments show that models trained with INSIDE-expanded data outperform those trained only on real data, achieving 3.06% to 5.24% relative improvements. While INSIDE is primarily designed for speaker verification, we also validate its effectiveness on gender classification, where it yields a 13.44% relative improvement. Moreover, INSIDE is compatible with other augmentation techniques and can serve as a flexible, scalable addition to existing training pipelines.**

I. Introduction

Speaker verification (SV) is the task of determining whether a given speech segment belongs to a claimed speaker [1]. Recent progress in deep learning has significantly advanced this task, resulting in remarkable performance gains [2]–[6]. The availability of large-scale, diverse, and labeled datasets has played an important role in enabling models to learn robust and discriminative speaker representations [7]–[9]. However, collecting and annotating such datasets is often resource-intensive, time-consuming, and fraught with privacy concerns [10].

Previous studies have widely employed techniques such as additive noise, reverberation, and speed perturbation to improve robustness in speaker verification [11]. These techniques operate at the acoustic level and do not generate new speaker identities. An exception is speed perturbation [12], which has been argued by Yamamoto et al. [13] to create new speaker identities. However, this method does not fundamentally modify speaker-specific characteristics, thereby limiting the diversity of speaker identities in the training data.

To overcome these limitations, we propose a more flexible and effective approach that generates synthetic speaker identities directly in the speaker embedding space, rather than relying on superficial acoustic transformations. Speaker embeddings, which capture speaker-specific characteristics in a learned latent space, have become a core component in modern speaker recognition systems. Importantly, the geometric structure of this space allows for meaningful interpolation between different speaker embeddings. Inspired by this property, we introduce **INSIDE** (**In**terpolating **S**peaker **Id**entities in **E**mbedding Space), a data expansion method that creates new speaker identities by interpolating between embeddings of real speakers. These interpolated embeddings are then used to condition a text-to-speech (TTS) model, enabling the generation of synthetic speech samples that represent new speaker identities. The resulting data expands the diversity of training speakers, enhancing model robustness and generalization. The INSIDE framework offers several advantages:

- The proposed INSIDE is a scalable and privacy-friendly data expansion method that generates diverse speaker identities without requiring additional data collection.
- As the synthetic speakers are created through interpolation in the embedding space, the resulting samples preserve the semantic structure of speaker characteristics, which leads to more stable and effective model training.
- The method is controllable, allowing flexible adjustment of gender ratio, language, content, and the number of identities to match specific augmentation needs.
- While our primary focus is speaker verification, we also validate INSIDE on gender classification. The framework shows potential for broader speech-related tasks.

II. Related Work

A. Data Augmentation in Speaker Verification

Data augmentation plays a key role in improving the robustness of speaker verification systems. Most existing methods focus on acoustic-level transformations, such as adding noise and reverberation [11], [14], or using time-domain techniques like speed perturbation [12]. Speed perturbation alters the speaking rate of recordings, typically by applying speed factors. This method introduces variability and is sometimes used to simulate new speaker identities [13]. However, these approaches have clear limitations. Since speed perturbation and other acoustic-level methods only modify low-level signal

979-8-3315-7206-8/25/$31.00

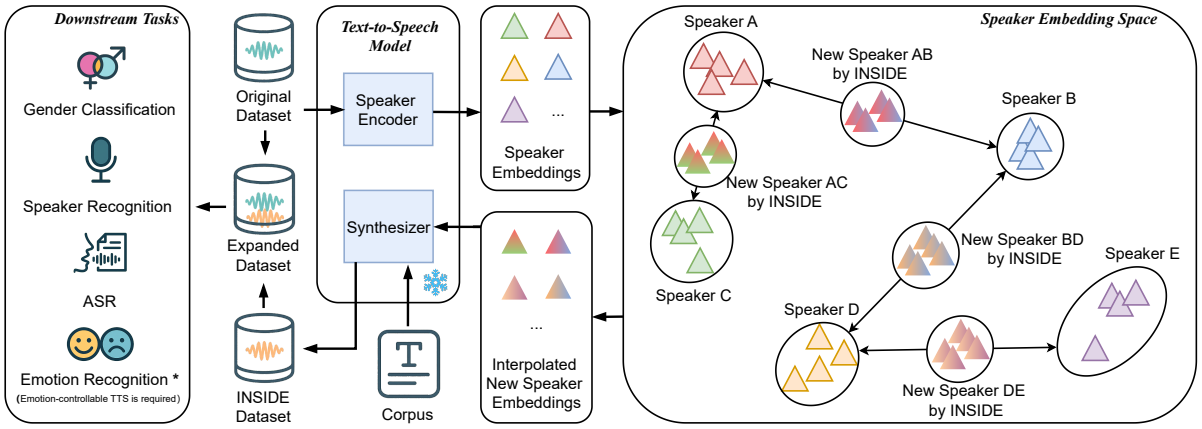

Fig. 1. Overview of the INSIDE data expansion pipeline. The snowflake icon denotes that the TTS model is pretrained and remains frozen.

characteristics, the generated utterances remain tightly coupled with the original speakers and phonetic content. They neither create truly new identities nor introduce sufficient variability in linguistic or phonetic expression. As a result, such methods struggle to capture intra-speaker variation and provide limited benefits when scaling to unseen speakers.

To address these limitations, we propose INSIDE, a framework that augments data directly in the embedding space. It generates new speaker identities by interpolating between real embeddings. This enables flexible, identity-level expansion with independent control over linguistic content, improving both speaker discriminability and intra-speaker variability.

B. Synthetic Speech Dataset

The advancement of speaker recognition systems relies on large-scale and diverse datasets. However, high collection costs and privacy concerns have led to the usage of synthetic datasets. These datasets aim to supplement or replace real-world data [10], offering benefits such as improved privacy, scalability, and the ability to simulate various conditions.

Synthetic audio has been widely used for data augmentation in many speech tasks [15]. For example, Libri2Vox [16] supports target speaker extraction, and SynthASR [17] is used in automatic speech recognition (ASR) where annotated data is limited or hard to obtain. In SV, synthetic data has also been explored. Tao et al. [18] use synthetic data to enhance defective datasets. SynVox2 [10] addresses privacy risks in datasets like VoxCeleb2 by generating anonymized yet useful synthetic alternatives. SynAug [19] augments text-dependent speaker verification by synthesizing fixed-transcript utterances conditioned on embeddings of real speakers. While effective, this method still relies on existing speaker identities and is therefore not scalable beyond the enrolled speaker set, limiting identity diversity and raising ongoing privacy concerns.

These constraints motivate the design of INSIDE, an identity-level data expansion framework that generates virtual speaker identities by interpolating between embeddings of real speakers in latent space. This enables scalable and diverse data generation without collecting new identity data, improving generalization while reducing privacy risks.

III. Method

A. Overview of the INSIDE

Diverse speaker identities are essential for speaker verification, while collecting large labeled datasets is expensive and raises privacy concerns [20], [21]. Since speaker embeddings typically lie in a structured latent space where real speakers are unevenly distributed and certain regions are sparsely populated [22], INSIDE leverages this property to generate intermediate embeddings between similar speakers, effectively filling underrepresented areas. This results in improved coverage of the embedding space and exposes the model to richer speaker variation, improving generalization. These new identities also smooth decision boundaries, helping reduce overfitting. As a result, INSIDE enhances both data diversity and distribution, leading to more reliable speaker verification.

B. Interpolating Speaker Identities in Embedding Space

The proposed INSIDE pipeline is illustrated in Fig. 1. We first use a pretrained speaker encoder from a controllable TTS model to extract averaged speaker embeddings $\mathbf{e}_i \in \mathbb{R}^N$ from a labeled dataset, where N is the embedding dimension.

To ensure natural interpolation, we group speakers by gender to avoid blending across major acoustic boundaries and preserve coherence in the generated identities. Let $\mathcal{E}$ be the set of embeddings with the same gender as $\mathbf{e}_i$, from which identity pairs $(\mathbf{e}_i, \mathbf{e}_j)$ are sampled for the interpolation process.

Given two source embeddings, several interpolation strategies can be considered, such as linear interpolation and spherical linear interpolation (SLERP) [23]. We adopt SLERP because it better fits the hyperspherical geometry of embedding spaces and preserves unit norm, ensuring that interpolated embeddings lie on the unit hypersphere, which aligns with the cosine similarity metric commonly used in speaker verification [24]. The SLERP operation is defined as:

$$\theta = \arccos\left(\frac{\mathbf{e}_i^\top \mathbf{e}_j}{|\mathbf{e}_i||\mathbf{e}_j|}\right), \tag{1}$$

$$\mathbf{e}_{ij} = \frac{\sin((1-\alpha)\theta)}{\sin(\theta)} \cdot \mathbf{e}_i + \frac{\sin(\alpha\theta)}{\sin(\theta)} \cdot \mathbf{e}_j, \quad \alpha \in [0,1]. \tag{2}$$

The coefficient α balances the two source speakers, enabling smooth transitions and avoiding artifacts from naive averaging in angular-sensitive spaces.

As the interpolated embeddings $\mathbf{e}_{ij}$ lie in the same embedding space as the original ones, they can be fed directly into the TTS synthesizer without introducing domain mismatch, as illustrated in Fig. 1. A set of texts with appropriate lengths is extracted from a large-scale corpus to serve as the linguistic content. For each synthetic speaker, a random subset of these texts is selected, and corresponding speech waveforms are synthesized. This process transforms abstract latent vectors into realistic speech data while preserving the identity attributes encoded in the embedding. The resulting utterances constitute the INSIDE dataset, which can be combined with real data to train downstream models.

C. Optimized Pair Selection via Nearest-Neighbor Traversal

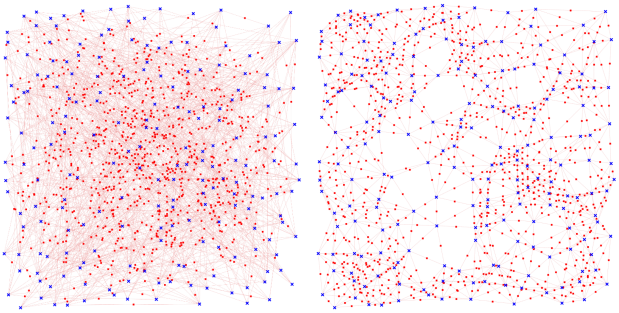

Fig. 2. Visualization of identity interpolation in the speaker embedding space. Blue points represent real speaker embeddings. Red lines connect selected pairs of real speakers. Red points denote synthetic speaker identities generated via interpolation between each pair.

As shown in Fig. 2 (a), while random pairing is effective, it results in uneven coverage of the embedding space. Interpolated embeddings $\mathbf{e}_{ij}$ tend to cluster in specific regions of the embedding space, leaving peripheral areas underrepresented.

To mitigate over-sampling in dense central regions and improve overall coverage, we introduce a layered nearest-neighbor interpolation strategy inspired by [18]. Specifically, for each embedding $\mathbf{e}_i$, we compute cosine distances to all other embeddings within the same gender group:

$$d_{ij} = 1 - \frac{\mathbf{e}_i^\top \mathbf{e}_j}{|\mathbf{e}_i||\mathbf{e}_j|}, \quad i \neq j. \tag{3}$$

We build identity pairs by gradually expanding the neighborhood. The n nearest neighbors of i are defined as:

$$\mathcal{N}_n(i) = j \in \mathcal{E} \mid \text{rank}_i(j) \leq n,\ j \neq i. \tag{4}$$

The candidate pair pool ($\mathcal{P}$) at neighborhood level n is defined as:

$$\mathcal{P}_n = (i, j) \mid j \in \mathcal{N}_n(i). \tag{5}$$

To avoid duplication due to symmetry (i.e., both (i, j) and (j, i)), we apply a uniqueness constraint. Let $\mathcal{S}$ be the set of selected pairs. We define:

$$\mathcal{S} = \bigcup_{n=1}^{n_{\max}} \text{UniquePairs}(\mathcal{P}_n \setminus \mathcal{S}) \quad \text{s.t.} \quad |\mathcal{S}| \geq T, \tag{6}$$

where $\setminus$ denotes the set difference. T is the desired number of synthetic identities. We begin with $n = 1$, collecting the closest neighbor for each identity. If the number of unique pairs is insufficient, n is incremented, and the next closest neighbors are included, repeating until $|\mathcal{S}| \geq T$. If the final level $n = k$ produces more candidates than required, we randomly sample the remainder from $\mathcal{P}_k \setminus \mathcal{S}$ to meet the target.

This method promotes a more uniform traversal of the embedding space. As shown in Fig. 2 (b), optimized pairing yields a distribution that more closely resembles that of real data compared to random sampling, while also filling some under-represented regions of the original dataset. This leads to smoother decision boundaries during training.

IV. Experimental Setup

A. Train Set

TABLE I
STATISTICS OF TRAIN SET.

Train Dataset	Set	# of male	# of female	# of speakers	# of samples
VoxCeleb2	Dev	3,682	2,312	5,994	1,092,009
INSIDE	Synthetic	3,682	2,312	5,994	1,092,009
INSIDE	Nearest-Neighbor	3,682	2,312	5,994	1,092,009
INSIDE	Identity-expanded	20,000	20,000	40,000	1,000,000

We follow the common speaker verification protocol by using the VoxCeleb2 [25] dataset as the training set for our baseline models and evaluating performance on the VoxCeleb1 [24], [26], [27]. In addition, we construct three synthetic datasets using our proposed INSIDE:

- **Synthetic (Syn)** comprises synthetic speech generated using the method described in Section III-B, where new speaker identities are created between random pairs.
- **Nearest-Neighbor (NN)** improves upon INSIDE-Syn dataset by optimizing the selection of speaker pairs, as introduced in Section III-C.
- **Identity-Expanded (ID-Exp)** follows the same method as NN but greatly increases the number of speaker identities, aiming to investigate whether further expanding identity quantity leads to additional performance gains.

B. Test Set

TABLE II
STATISTICS OF TEST SETS FOR SPEAKER VERIFICATION.

Test Dataset	# of identities	# of trial pairs
VoxCeleb1-Original (Vox1-O) [28]	40	37,611
VoxCeleb1-Extended (Vox1-E) [28]	1,251	579,818
VoxCeleb1-Hard (Vox1-H) [28]	1,251	550,894

For the speaker verification task, as shown in Table II, we evaluate our models on the widely used VoxCeleb1 [28] dataset, which includes three standard evaluation protocols: original, extended, and hard cases.

In addition to speaker verification, we investigate whether our proposed method can benefit other speech-related tasks. We use gender classification as a case study and evaluate on three public datasets—VoxCeleb1 [28], TIMIT [29], and Samrómur Children [30]—as well as a private dataset. Since gender classification typically achieves high accuracy, we

TABLE III

SPEAKER VERIFICATION RESULTS IN EER (%) AND MINDCF ON THE VOXCELEB1 DATASET. THE BEST RESULT IN EACH GROUP IS SHOWN IN **BOLD**, AND THE SECOND-BEST IS UNDERLINED. 'AVERAGED IMPROVEMENT' SHOWS THE AVERAGE RELATIVE GAIN OVER THE BASELINE ACROSS ALL TEST SUBSETS.

front-end	Backend	INSIDE Portion Added	Front-end Finetune	AS-Norm	Vox1-O		Vox1-E		Vox1-H		Averaged Improvement↑
					EER↓	minDCF↓	EER↓	minDCF↓	EER↓	minDCF↓	
WavLM -Large	ECAPA-TDNN (C=512)	N/A		✓	0.691	0.097	0.785	0.088	1.612	0.175	-
		Syn		✓	0.697	0.089	0.780	0.086	1.634	0.172	1.82%
		NN		✓	0.694	0.090	0.758	0.092	**1.532**	**0.165**	2.76%
		ID-Exp		✓	**0.649**	**0.089**	**0.747**	**0.084**	1.569	0.166	5.24%
	ECAPA-TDNN Global (C=512)	N/A			0.595	-	0.719	-	1.501	-	-
		ID-Exp			**0.545**	0.068	**0.699**	0.076	**1.488**	0.144	4.02%
		N/A		✓	0.548	-	0.656	-	1.355	-	-
		ID-Exp		✓	**0.518**	0.072	**0.635**	0.069	**1.334**	0.137	3.38%
		N/A	✓		0.542	-	0.635	-	1.355	-	-
		ID-Exp	✓		**0.489**	0.058	**0.619**	0.068	**1.337**	0.136	4.53%
		N/A	✓	✓	0.521	-	0.594	-	**1.237**	-	-
		ID-Exp	✓	✓	**0.479**	0.059	**0.583**	0.063	1.247	0.124	3.06%

TABLE IV

STATISTICS OF TEST SETS FOR GENDER CLASSIFICATION.

Test Dataset	Set	# of samples	Remarks
VoxCeleb1 [28]	-	153,516	In-domain
TIMIT [29]	Train	4,620	Cross-domain, English
	Test	1,680	
Private	Part 1 Train	2,264,026	Cross-domain, South-east Asia languages
	Part 1 Test	5,000	
	Part 2 Train	2,473,967	
	Part 2 Test	5,000	
Samrómur Children [30]	Train	129,446	Cross-domain, Icelandic children speech

include these diverse datasets to assess our method under varying acoustic conditions and speaker demographics, aiming to validate its generalizability and robustness.

C. Training Configuration

To construct the INSIDE datasets, we use $\alpha = 0.5$ to generate midpoint embeddings, representing a balanced blend of both identities. The speech synthesis model used is YourTTS [31], provided by Coqui[1]. Text samples are drawn from the LibriSpeech [32] and filtered to 100–250 words to match the average duration of VoxCeleb2 utterances [25].

All experiments are conducted using the WeSpeaker toolkit [24], and the same training strategies are applied to models trained with either the original dataset or its combination with INSIDE data. WavLM-Large [33] is used as the front-end; reverberation and additive noise are applied with a probability of 0.6; all models are optimized using SGD with momentum 0.9 and weight decay of 1×10^{-4}.

For speaker verification, the training settings differ by back-end architecture. Models with the ECAPA-TDNN [34] back-end are trained for 40 epochs using AAM-softmax with a scale of 32 and a margin that increases from 0 to 0.3 between epochs 5 and 10. The learning rate decays exponentially from 0.1 to 5×10^{-5} with a 2-epoch warm-up. Input utterances are cut to 300 frames, and speed perturbation is applied [12], [13]. For ECAPA-TDNN with global context attention (ECAPA-TDNN Global) [34], we follow the WeSpeaker configurations[2] and train for 150 epochs. The learning rate decays exponentially from 0.1 to 1×10^{-5} with a 6-epoch warm-up. The loss is AAM-softmax with sub-center, and inter top-k penalty [35] with a scale of 32 and a margin that increases exponentially from 0 to 0.2, starting at epoch 20 and fixed after epoch 40. Utterances are cut to 150 frames. Speed perturbation is also applied. When the front-end model is jointly finetuned with the back-end during training, the model is trained for 20 epochs with the learning rate decaying from 0.001 to 2.5×10^{-4} and a 1-epoch warm-up. Other settings remain the same as above.

For gender classification, we use a lightweight ECAPA-TDNN with channel size 128 and softmax loss. The front-end remains frozen. Training lasts for 3 epochs, with the learning rate decaying from 0.1 to 5×10^{-5} and no warm-up. Utterances are fixed to 200 frames.

[1]https://github.com/coqui-ai/TTS

[2]https://github.com/wenet-e2e/wespeaker

V. RESULT

A. Speaker Verification

We first evaluate the effectiveness of the proposed INSIDE on the SV task, using WavLM-Large as the front-end and ECAPA-TDNN as the back-end. Our baseline achieves strong performance on the VoxCeleb1 test sets, as shown in Table III. Augmenting the training data with synthetic identities generated by the INSIDE approach (Section III-B) yields an average relative improvement of 1.82%.

Further gains are achieved using the nearest-neighbor optimized pairing strategy (Section III-C), referred to as INSIDE-NN, which improves performance by 2.76%, indicating that the distribution of synthetic identities has a noticeable impact on the quality of the augmented dataset.

Taking advantage of INSIDE's scalability, we expand the number of synthetic identities to 40,000, creating the ID-Exp dataset. This leads to a substantial average relative improvement of 5.24%, demonstrating the importance of identity quantity in speaker-related tasks.

Additionally, we assess the robustness and compatibility of INSIDE within the WeSpeaker [24] framework under various settings, including whether front-end finetuning and AS-Norm are applied, and a different backend of ECAPA-TDNN Global [34]. In most of these settings, the proposed method consistently improves performance, yielding relative performance gains of 3.06% to 4.53%.

TABLE V

GENDER CLASSIFICATION RESULTS IN ERROR RATE (%). THE BEST RESULT IN EACH TEST TRACK IS SHOWN IN **BOLD**. 'AVERAGED IMP.' INDICATES THE AVERAGE RELATIVE IMPROVEMENT OVER THE BASELINE ACROSS ALL TEST SUBSETS.

front-end	Backend	INSIDE Portion Added	# of Total IDs	VoxCeleb1 [28]	TIMIT [29]		Private				Samrómur Children [30] ↓	Averaged Imp. ↑
					Train↓	Test↓	P1 Train↓	P1 Test↓	P2 Train↓	P2 Test↓		
WavLM -Large	ECAPA-TDNN (C=128)	N/A	5,994	1.24	0.50	0.89	2.20	2.42	2.55	2.36	37.01	-
		NN	11,988	**1.13**	0.25	**0.60**	2.16	2.46	2.55	2.42	34.17	12.09%
		ID-Exp	45,994	1.42	**0.17**	0.65	**2.10**	**2.26**	**2.49**	**2.30**	**32.18**	13.44%

B. Gender Classification

We hypothesize that since INSIDE selects same-gender speaker pairs for identity interpolation, the resulting synthetic identities are likely to preserve the gender of the original pair. This suggests that INSIDE may also be beneficial for gender classification tasks. To evaluate this hypothesis, we conducted experiments on several datasets, and the results are shown in Table V. We observe that the proposed INSIDE-NN consistently outperforms the baseline without INSIDE-based data expansion in most cases, achieving an average improvement of 12.09%. Furthermore, the ID-Exp variant, which introduces a larger number of synthetic identities, leads to even better performance with an average gain of 13.44%.

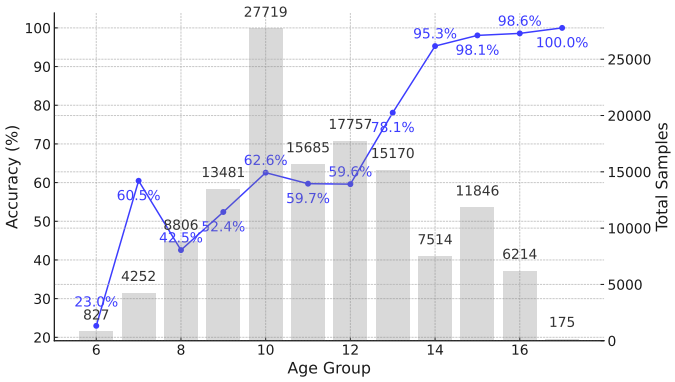

Fig. 3. Accuracy across different age groups (blue line, left Y-axis) and corresponding number of samples in each group (gray bars, right Y-axis).

A deeper analysis reveals that the Private and Samrómur Children datasets exhibit higher classification error rates. One possible reason for this is language mismatch [36], as both datasets differ from the English-based training data. Notably, Samrómur Children dataset shows high error rates. To further investigate this, we analyze the results by age group, as shown in Fig. 3. The findings indicate that gender classification is significantly more challenging for younger children (ages 6–13), with error rates increasing as age decreases. In contrast, for children aged 15 and above, the error rate drops below 2%.

VI. LIMITATIONS AND FUTURE WORK

Although the proposed INSIDE framework improves performance in both speaker verification and gender classification tasks, it still has several limitations.

First, the speaker encoder used in most current TTS systems is relatively lightweight and generally less accurate than those used in state-of-the-art speaker verification models based on speech foundation models [24]. This may limit the effectiveness of speaker embedding-based data expansion, particularly when applied to strong speaker verification baselines. For example, in our experiments with a highly competitive baseline, the relative performance gain is limited to 5.24%. Future work could explore TTS models with more powerful speaker encoders to improve the quality of identity interpolation.

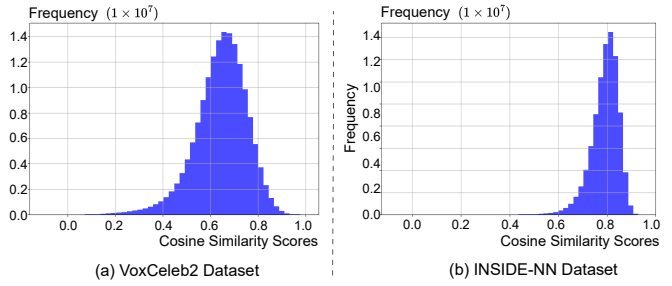

Fig. 4. Distribution of cosine similarity scores between different utterances of the same identity in the (a) VoxCeleb2 dataset, and (b) INSIDE-NN dataset.

Second, as shown in Fig. 4, we observe that cosine similarity scores between different utterances from the same synthetic identity are significantly higher than those of real identities. This indicates that synthetic identities exhibit lower intra-class uncertainty compared to real speakers. Such a discrepancy in data distribution may pose challenges for speaker verification models, particularly in learning to model speaker uncertainty, and could negatively impact robustness [37]. Future work may explore generating synthetic data that better mimics the statistical properties of real speaker distributions, including intra-class variability, to reduce this mismatch.

VII. CONCLUSIONS

In this work, we proposed INSIDE (Interpolating Speaker Identities in Embedding Space), a novel data expansion framework that generates new speaker identities via spherical interpolation between real speaker embeddings. By conditioning a TTS model on these interpolated embeddings, we synthesize realistic speech data, enhancing identity diversity without requiring additional data collection. We further introduce an optimized pair selection strategy and demonstrate INSIDE's compatibility with different training configurations. Experimental results show that INSIDE consistently improves performance on both speaker verification and gender classification tasks, achieving up to 5.24% and 13.44% average relative gains, respectively. Moreover, INSIDE is compatible with existing data augmentation methods and can further enhance performance when combined with them.

ACKNOWLEDGMENT

Ke Zhang and Haizhou Li are supported in part by the Shenzhen Science and Technology Program (Shenzhen Key Laboratory, Grant No. ZDSYS20230626091302006), the Shenzhen Science and Technology Research Fund (Fundamental Research Key Project, Grant No. JCYJ20220818103001002), and the Program for Guangdong Introducing Innovative and Entrepreneurial Teams (Grant No. 2023ZT10X044).

REFERENCES

[1] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Communication*, vol. 52, no. 1, pp. 12–40, 2010.

[2] I. Yakovlev, R. Makarov, A. Balykin, P. Malov, A. Okhotnikov, and N. Torgashov, "Reshape dimensions network for speaker recognition," *arXiv preprint arXiv:2407.18223*, 2024.

[3] T. Liu, K. A. Lee, Q. Wang, and H. Li, "Golden Gemini is all you need: Finding the sweet spots for speaker verification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 2324–2337, 2024.

[4] S. Peng, W. Guo, J. Zhang, *et al.*, "A study of multi-scale feature learning from pre-trained models on speaker verification," in *Proc. ICASSP*, 2025, pp. 1–5.

[5] J. Peng, L. Mošner, L. Zhang, *et al.*, "Ca-mhfa: A context-aware multi-head factorized attentive pooling for ssl-based speaker verification," in *Proc. ICASSP*, 2025, pp. 1–5.

[6] T. Liu, R. K. Das, K. A. Lee, and H. Li, "Neural acoustic-phonetic approach for speaker verification with phonetic attention mask," *IEEE Signal Processing Letters*, vol. 29, pp. 782–786, 2022.

[7] Y. Lin, M. Cheng, F. Zhang, Y. Gao, S. Zhang, and M. Li, "VoxBlink2: A 100k+ speaker recognition corpus and the open-set speaker-identification benchmark," in *Proc. Interspeech*, 2024, pp. 4263–4267.

[8] S. Zheng, L. Cheng, Y. Chen, H. Wang, and Q. Chen, "3d-speaker: A large-scale multi-device, multi-distance, and multi-dialect corpus for speech representation disentanglement," *arXiv preprint arXiv:2306.15354*, 2023.

[9] S. Wang, Z. Chen, K. A. Lee, Y. Qian, and H. Li, "Overview of speaker modeling and its applications: From the lens of deep speaker representation learning," *arXiv preprint arXiv:2407.15188*, 2024.

[10] X. Miao, X. Wang, E. Cooper, *et al.*, "Synvox2: Towards a privacy-friendly voxceleb2 dataset," in *Proc. ICASSP*, 2024, pp. 11 421–11 425.

[11] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *Proc. ICASSP*, 2018, pp. 5329–5333.

[12] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Proc. Interspeech*, 2015, pp. 3586–3589.

[13] H. Yamamoto, K. A. Lee, K. Okabe, and T. Koshinaka, "Speaker augmentation and bandwidth extension for deep speaker embedding," in *Proc. Interspeech*, 2019, pp. 406–410.

[14] S. Wang, Y. Yang, Z. Wu, Y. Qian, and K. Yu, "Data augmentation using deep generative models for embedding based speaker recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2598–2609, 2020.

[15] Z. Huang, J. Lin, M. Ge, *et al.*, "Augmenting short enrollment speech via synthesis for target speaker extraction," in *Proc. ICASSP*, 2025, pp. 1–5.

[16] Y. Liu, X. Liu, X. Miao, and J. Yamagishi, "Libri2Vox dataset: Target speaker extraction with diverse speaker conditions and synthetic data," *arXiv preprint arXiv:2412.12512*, 2024.

[17] A. Fazel, W. Yang, Y. Liu, *et al.*, "Synthasr: Unlocking synthetic data for speech recognition," in *Proc. Interspeech*, 2021, pp. 896–900.

[18] R. Tao, Z. Shi, Y. Jiang, T. Liu, and H. Li, "Voice conversion augmentation for speaker recognition on defective datasets," *arXiv preprint arXiv:2404.00863*, 2024.

[19] C. Du, B. Han, S. Wang, Y. Qian, and K. Yu, "Synaug: Synthesis-based data augmentation for text-dependent speaker verification," in *Proc. ICASSP*, 2021, pp. 5844–5848.

[20] 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)*, vol. 36, 2023, pp. 50 221–50 236.

[21] R. Tao, K. Aik Lee, R. Kumar Das, V. Hautamäki, and H. Li, "Self-supervised speaker recognition with loss-gated learning," in *Proc. ICASSP*, 2022, pp. 6142–6146.

[22] Y. Cai, L. Li, A. Abel, X. Zhu, and D. Wang, "Deep normalization for speaker vectors," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 733–744, 2021.

[23] K. Shoemake, "Animating rotation with quaternion curves," in *Proc. SIGGRAPH*, 1985, pp. 245–254.

[24] H. Wang, C. Liang, S. Wang, *et al.*, "Wespeaker: A research and production oriented speaker embedding learning toolkit," in *Proc. ICASSP*, 2023, pp. 1–5.

[25] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep speaker recognition," in *Proc. Interspeech*, 2018, pp. 1086–1090.

[26] Y. Chen, C. Deng, H. Wang, *et al.*, "Pushing the frontiers of self-distillation prototypes network with dimension regularization and score normalization," in *Proc. Interspeech*, 2025, pp. 3688–3692.

[27] T. Liu, R. K. Das, K. Aik Lee, and H. Li, "MFA: TDNN with multi-scale frequency-channel attention for text-independent speaker verification with short utterances," in *Proc. ICASSP*, 2022, pp. 7517–7521.

[28] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A large-scale speaker identification dataset," in *Proc. Interspeech*, 2017, pp. 2616–2620.

[29] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, and D. Pallett, "Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1," *NASA STI/Recon Technical Report N*, vol. 93, p. 27 403, 1993.

[30] C. D. Hernandez Mena, D. E. Mollberg, M. Borský, and J. Guðnason, "Samrómur children: An Icelandic speech corpus," in *Proc. LREC*, 2022, pp. 995–1002.

[31] E. Casanova, J. Weber, C. D. Shulby, A. C. Junior, E. Gölge, and M. A. Ponti, "YourTTS: Towards zero-shot multi-speaker TTS and zero-shot voice conversion for everyone," in *Proc. ICML*, 2022, pp. 2709–2720.

[32] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *Proc. ICASSP*, 2015, pp. 5206–5210.

[33] S. Chen, C. Wang, Z. Chen, *et al.*, "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[34] 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.

[35] M. Zhao, Y. Ma, M. Liu, and M. Xu, "The speakin system for voxceleb speaker recognition challange 2021," *arXiv preprint arXiv:2109.01989*, 2021.

[36] T. Liu, I. Kukanov, Z. Pan, Q. Wang, H. B. Sailor, and K. A. Lee, "Towards quantifying and reducing language mismatch effects in cross-lingual speech anti-spoofing," in *IEEE Spoken Language Technology Workshop (SLT)*, 2024, pp. 1185–1192.

[37] K. A. Lee, Q. Wang, and T. Koshinaka, "Xi-vector embedding for speaker recognition," *IEEE Signal Processing Letters*, vol. 28, pp. 1385–1389, 2021.