

Understanding Hyperspherical Geometry of ECAPA-TDNN Embedding and Its Impact on Zero-Shot Voice Conversion

Mathilde Abrassart
STMS Lab, IRCAM
CNRS, Sorbonne Université
Paris, France
mathilde.abrassart@ircam.fr

Nicolas Obin
STMS Lab, IRCAM
CNRS, Sorbonne Université
Paris, France
nicolas.obin@ircam.fr

Axel Roebel
STMS Lab, IRCAM
CNRS, Sorbonne Université
Paris, France
axel.roebel@ircam.fr

Abstract—Angular-margin speaker encoders are widely used in voice conversion, yet the geometry of their classifier prototypes remains poorly understood. We analyze ECAPA-TDNN classifier prototypes as points on the unit hypersphere and characterize their organization using rotation-invariant angular statistics together with global and local effective dimensionality measures. Our analysis shows that standard training can induce angular concentration and a substantial reduction in effective dimensionality. To address this, we investigate two geometric regularization strategies (hinged Riesz log-energy and effective-dimension maximization) applied to classifier prototypes to encourage more uniform hyperspherical coverage. The resulting prototype sets exhibit higher effective dimensionality and improved isotropy, with configuration-dependent effects on speaker-recognition performance. When the corresponding ECAPA-TDNN models are used as speaker encoders for Fast-VGAN, the regularized systems also exhibit improved robustness in zero-shot voice conversion, particularly for previously unseen speakers.

Index Terms—speaker embeddings, hyperspherical geometry, zero-shot voice conversion

I. INTRODUCTION

Speaker embeddings have become a fundamental component of modern speech processing systems, enabling robust speaker verification, diarization, and increasingly conditional generative modeling. Early approaches such as i-vectors [1] relied on generative factor analysis, while the introduction of discriminative neural embeddings such as x-vectors [2] marked a shift toward supervised representation learning. Subsequent architectural refinements, including ECAPA-TDNN [3], further improved discriminative power on large-scale benchmarks. Beyond speaker recognition, these embeddings are widely used as conditioning signals for multi-speaker speech synthesis and zero-shot voice conversion, where generalization to unseen speakers is critical.

In parallel, training objectives evolved from standard cross-entropy classification toward angular-margin-based losses rooted in hyperspherical metric learning, such as ArcFace [4]. These objectives normalize embeddings and classifier prototypes onto the unit hypersphere, making classification depend solely on angular separation. Both open-set speaker

verification and zero-shot voice conversion require generalization to unseen speakers. However, zero-shot voice conversion additionally relies on the speaker representation as a conditioning space, for which smooth and well-covered regions may facilitate the representation of speakers that were not observed during training. We therefore investigate whether improving the isotropy and coverage of this space benefits zero-shot voice conversion.

Recent advances in representation learning highlight the importance of geometric organization in such settings. The alignment-uniformity framework [5] shows that effective embeddings balance intra-class compactness and hyperspherical coverage, while Neural Collapse predicts convergence of class means toward simplex equiangular configurations in late-stage training [6]. Several studies have also related intrinsic dimensionality and spectral properties to representation quality [7]–[9]. Despite the widespread use of angular-margin objectives in speaker embedding systems, relatively little attention has been devoted to understanding the geometry of the resulting classifier prototypes or to evaluating how this geometry affects downstream applications such as zero-shot voice conversion. This paper makes the following contributions:

- 1) We present a systematic geometric analysis of the hyperspherical prototype spaces induced by angular-margin speaker encoders or as conditioning representations in generative models, characterizing their structure through global and local effective dimensionality and angular distance statistics.
- 2) Motivated by the observed angular concentration and reduced effective dimensionality, we introduce two geometric regularizers that encourage a more uniform hyperspherical distribution and higher effective dimensionality of classifier prototypes.
- 3) We provide a preliminary evaluation of the practical implications of these geometric properties in two settings: speaker classification and zero-shot voice conversion using Fast-VGAN [10], where ECAPA embeddings are used as conditioning representations for the generative model. The regularized systems exhibit

improved zero-shot voice conversion robustness together with substantially more isotropic and higher-dimensional prototype spaces, with configuration-dependent speaker-recognition results.

II. RELATED WORK

A. Speaker representation learning with angular margins

Speaker verification has progressively shifted from probabilistic latent-variable modeling to fully discriminative representation learning. The introduction of x-vectors [2] marked a transition to supervised neural embeddings, where a TDNN with statistical pooling learns utterance-level speaker representations. Recent systems incorporate stronger geometric constraints through ℓ_2 normalization and angular-margin objectives, projecting both embeddings $\mathbf{x}$ and classifier weights $\mathbf{w}_k$ onto the unit hypersphere, $\|\mathbf{x}\|_2 = \|\mathbf{w}_k\|_2 = 1$, so that classification depends only on the cosine similarity $\cos \theta_k = \mathbf{w}_k^\top \mathbf{x}$. This paradigm was first explored in face recognition through SphereFace [11], CosFace [12], and ArcFace [4], and was subsequently adopted in speaker verification [13], [14]. Modern ECAPA-TDNN systems commonly employ ArcFace-like objectives [3], [15]. ECAPA-TDNN further improved speaker embeddings through multi-scale temporal modeling, channel attention, and attentive pooling, achieving strong performance on VoxCeleb [3]. Numerous extensions have since improved robustness and discriminability [16]–[18], and ECAPA embeddings have been successfully transferred to tasks such as diarization and multi-speaker TTS [19], [20]. Despite strong downstream performance (e.g., EER, DER), the geometric structure of the embedding and classifier spaces—its isotropy, dimensionality, and angular organization—remains largely unexplored, motivating our study.

B. Geometric analysis of representation spaces

The geometry of learned representations has become a central topic in deep learning. In contrastive learning, the alignment-uniformity framework shows that effective embeddings balance intra-class compactness with uniform hyperspherical coverage [5]. In supervised settings, Neural Collapse predicts convergence of class means toward a simplex equiangular configuration during late training [6]. Several studies further analyze the intrinsic dimensionality of deep representations, reporting substantial dimensional compression and structured anisotropy across layers [7], [8]. Complementary spectral analyses relate eigenvalue structure to implicit regularization and effective rank in deep networks [9]. More recently, hyperspherical energy objectives have been proposed to promote uniformly distributed embeddings on the unit sphere [21], while participation-ratio-based measures have also been investigated both as representation quality indicators and optimization objectives [22]. While these geometric phenomena are well established in vision and theoretical deep learning, they remain largely unexplored in the speech domain, particularly for speaker verification and zero-shot voice conversion. We address this gap by explicitly characterizing the isotropy, intrinsic dimensionality, and angular organization

of ECAPA classifier prototypes trained with angular margins, and by studying how geometry-aware regularization affects downstream performance.

III. GEOMETRY OF THE ECAPA PROTOTYPE SPACE

In this section, we analyze the geometry of the ECAPA-TDNN speaker prototypes obtained by training an ECAPA-TDNN model on LibriTTS-R [23] derived from LibriTTS [24]. Our goal is therefore not to maximize isotropy for its own sake, but to test whether reducing prototype-space anisotropy improves hyperspherical coverage and ultimately benefits downstream zero-shot generation. We use the small ECAPA-TDNN variant introduced in [3]. Since our goal is not to achieve robust speaker identification, but to study the geometry of the resulting speaker prototypes and their impact on zero-shot voice conversion, we use Mel-spectrograms as input features, and do not use any of the data augmentation strategies proposed [3]. This choice aligns the speaker encoder with the representation used by the generative model.

Under the angular-margin classification objective, the columns of the speaker prototype matrix $W \in \mathbb{R}^{D \times C}$ are ℓ_2 -normalized and therefore lie on the unit hypersphere. Classifier prototypes and utterance-level embeddings are distinct and need not share the same distribution. However, the angular-margin objective directly couples them through their cosine similarities. We therefore analyze the prototype geometry explicitly constrained by the classifier, while downstream VC experiments evaluate the representations produced by the corresponding encoder. We characterize the global and local effective dimensionality of the prototype distribution using the participation ratio D_{PR} [25]

$$\Sigma = WW^T \quad (1)$$

$$D_{PR} = \frac{\text{Tr}(\Sigma)^2}{\text{Tr}(\Sigma^2)} = \frac{(\sum_i \lambda_i)^2}{\sum_i \lambda_i^2}, \quad (2)$$

where Σ is the prototype scatter matrix and λ_i its eigenvalues, and $\text{Tr}(\cdot)$ the trace operator. The trace formulation avoids explicit eigen-decomposition and is therefore efficient to compute as a batch-wise regularizer. Note that D_{PR} is computed globally (G_{PR}) to characterize the dimensionality of the full prototype set and locally (L_{PR}), where it is computed over the k nearest neighbors of each prototype. Because the number of speaker prototypes is limited, we use $k = 50$, which bounds the locally observable dimensionality by 50.

A. Regularization training strategy

Although the ArcFace loss promotes inter-speaker separability, its influence on the global organization of the classifier prototypes is mediated by the speakers represented in each training batch. We evaluated batch sizes of 32 and 160 with uniform speaker sampling, corresponding to 32 and 160 distinct speakers per batch. Increasing the batch size had only a limited effect on the resulting prototype geometry, as discussed in Section III-B. We therefore investigate geometry-aware regularization objectives that act directly on the classifier prototypes, independently of the composition of individual

training batches. The first promotes local angular separation between neighboring prototypes, while the second maximizes the effective dimensionality of the prototype distribution. Together, they improve the global coverage of the hypersphere for zero-shot speaker conditioning. The first regularizer is based on the Riesz log-energy potential [21], [26] and aims to prevent excessive angular concentration between neighboring prototypes. We employ the Riesz log energy according to [26]

$$R(\omega_N) = -\frac{\sum_{i \neq j} \log(1 - x_i \cdot x_j)}{2N(N-1)}. \quad (3)$$

Here $\omega_N = \{x_i\}_1^N$, with $\|x\| = 1$, $x_i \cdot x_j$ denotes cosine similarity, and N is the number of prototypes. Given (3) is a global potential and is calculated for all pairs of points with only very few approaching critical angular distances, the direct application of the regularization is rather costly and dilutes the effect of the regularizer due to the small impact of an individual pair. Instead, we use a hinged variant that penalizes only pairs exceeding a cosine similarity threshold c_t . Accordingly, this regularizer globally ensures local separability. The threshold depends on the number of points N and the representation dimension D . We determine c_t from the distribution of angular distances after weight initialization and use the 98th percentile of cosine similarities among nearest neighbors. The threshold used for our configuration is shown in Figure 1. Let $\mathcal{A}$ denote the set of prototype pairs whose cosine similarity exceeds the threshold c_t then the regularizer becomes

$$\mathcal{A} = \{(i, j) \mid i \neq j, x_i \cdot x_j > c_t\}. \quad (4)$$

$$\mathcal{L}_R = \frac{1}{|\mathcal{A}|} \sum_{(i,j) \in \mathcal{A}} -\frac{1}{2} \log \frac{1 - x_i \cdot x_j}{1 - c_t}. \quad (5)$$

Fig. 1 shows the threshold c_t and the rapid increase of the regularizer loss once the similarity exceeds the threshold.

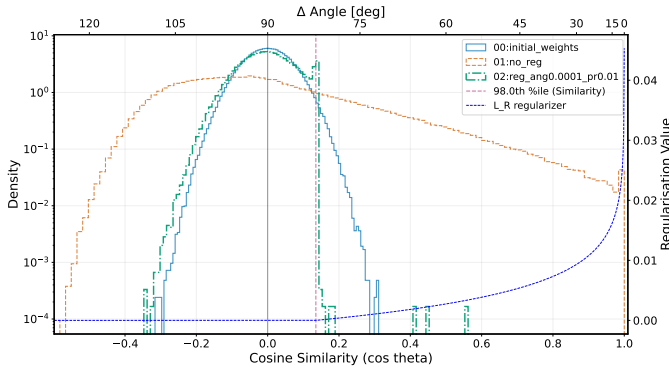


Fig. 1. Histogram of pairwise angles between ECAPA-TDNN speaker prototypes directly after initialization, and after training the model on LibriTTS-R with and without regularisation. The regularisation loss is shown in dashed blue.

The second regularizer maximizes the effective dimensionality of the prototype distribution through the participation

ratio. To exert sufficient pressure when the effective dimensionality decreases, we define

$$L_{PR} = -\log \left(\frac{D_{PR}}{D} + \epsilon \right). \quad (6)$$

Here, D_{PR} denotes the effective dimension from Eq. ((2)) and D the prototype dimension. Full regularization loss becomes

$$L_{reg} = \lambda_{PR} L_{PR} + \lambda_R L_R \quad (7)$$

Training the same ECAPA-TDNN model as before, using $\lambda_{PR} = 10^{-2}$, $\lambda_R = 10^{-4}$ substantially improves the prototype geometry. We use two additional angular metrics to characterize the relation between prototype and embedding geometry. RM° is the root-mean-square aggregation of the angular distance to the closest neighboring prototype, while $P2M^\circ$ is the 2nd percentile of nearest-neighbor inter-class angles, reflecting the weakest 2% of classes. For the jointly trained Fast-VGAN+ECAPA models analyzed in Section IV-A, we additionally report $Intra^\circ$, defined as the mean angular distance between utterance-level embeddings belonging to the same speaker, and NN° , defined as the mean angular distance from each classifier prototype to its nearest neighboring prototype. Comparing $Intra^\circ$ and NN° indicates whether the within-speaker angular spread of the embeddings remains smaller than the separation between neighboring classifier prototypes.

B. Emergence of hyperspherical structure

Model	RM $^\circ \uparrow$	P2M $^\circ \uparrow$	D _{PR} $\uparrow$	L _{PR} $\uparrow$	EER $\downarrow$	T5Acc $\uparrow$
Init	77.5	74.4	191.7	41.3	—	—
ECA S1	20.142	7.941	15.2	4.7	0.040	0.622
ECA S10	66.913	58.134	139.4	29.9	0.012	0.892
ECA S20	72.355	66.699	165.6	35.0	0.099	0.924
ECA S20 Reg	81.604	80.544	182.8	40.0	0.109	0.890
ECA S30	74.443	68.350	177.3	37.9	0.163	0.880

TABLE I
HYPERSPHERICAL GEOMETRY AND SPEAKER-RECOGNITION PERFORMANCE FOR ECAPA-TDNN. “S” DENOTES THE SCALE VALUE AND “REG” THE PROPOSED REGULARIZATION. THE MARGIN IS 0.2 AND THE BATCH SIZE IS 32.

As shown in Table I, the classifier scale strongly influences the geometry of the hyperspherical prototype space. With a small scale ($S = 1$), training produces a highly concentrated prototype distribution: both global and local participation ratios decrease drastically, while nearest-neighbor separation almost vanishes ($RM = 20.1^\circ$, $P2M = 7.9^\circ$). Increasing the scale progressively produces a better distributed prototype space. For $S = 20$, the geometric metrics are substantially improved relative to smaller scales, together with strong classification performance. Regularization further increases angular separation and participation ratios, yielding the most isotropic prototype distribution. However, it slightly degrades EER and Top-5 accuracy relative to $S = 20$. This illustrates that improved hyperspherical coverage does not necessarily translate into improved verification performance. Since EER primarily measures speaker discriminability rather than the suitability of the representation as a continuous conditioning space, we

treat it here as a secondary diagnostic. The downstream VC experiments provide the more relevant evaluation for the main question of this work. Interestingly, increasing the training batch size from 32 to 160 has only a marginal effect on the learned geometry. For the best-performing configuration ($S = 20$), the geometric metrics remain very similar (RM: $72.4^\circ \rightarrow 74.3^\circ$, P2M: $66.7^\circ \rightarrow 69.2^\circ$, G_{PR} : $165.6 \rightarrow 164.7$, L_{PR} : $35.0 \rightarrow 36.9$), while the validation performance changes only slightly (T5Acc: $0.924 \rightarrow 0.911$, EER: $0.099 \rightarrow 0.076$). This weak dependence on batch size is particularly encouraging for applications such as Fast-VGAN [10], where training is typically performed with a batch size of 32.

IV. IMPACT ON ZERO-SHOT VOICE-CONVERSION

This section investigates the use of ECAPA-based speaker representations for zero-shot voice conversion and examines the relation between hyperspherical geometry and conversion performance. Audio samples and supplementary material are available online.¹ We use Fast-VGAN [10] as the experimental backbone. We first vary the scale parameter of the ECAPA-TDNN angular-margin loss to analyze its effect on the geometry of the jointly trained speaker space. We then select a scale value and compare models trained with and without the proposed geometric regularization. Finally, we compare Fast-VGAN with four recent open-source zero-shot voice conversion baselines. The resulting systems are evaluated in terms of intelligibility, speaker similarity, perceptual quality, and generalization to unseen speakers. Fast-VGAN is a voice conversion model based on a 2D convolutional generator that predicts a target Mel-spectrogram from temporally aligned conditioning features such as F0, energy, phoneme representations, and linguistic content. Further architectural and implementation details are provided in [10]. The model is trained adversarially using a discriminator that distinguishes real from generated spectrograms. The resulting spectrograms are converted to waveforms using the neural vocoder MBExWN [27]. In the original any-to-many setting [10], speaker identity is encoded via a learned lookup embedding. In contrast, zero-shot voice conversion requires embeddings extracted from a speaker encoder. We therefore use ECAPA-TDNN embeddings and study how the geometry of the corresponding classifier prototypes relates to generalization to unseen speakers.

A. Hyperspherical structure

Table II reports the hyperspherical geometry metrics and speaker recognition performance for jointly trained Fast-VGAN and ECAPA-TDNN models under different scale values. For the best scale value, we added the result with regularization.

As observed with the standalone ECAPA-TDNN, modifying the scale parameter also affects the geometry of the learned speaker prototypes in the joint training setting. Increasing the scale from 15 to 30 progressively increases the geometric metrics, while the proposed regularization produces the highest

Model	Intra $^\circ$	NN $^\circ$	P2M $^\circ$	D $_{PR}$	L $_{PR}$	EER	T5Acc
FVG-Init	–	77.56	$\uparrow$ 74.6	$\uparrow$ 191.7	$\uparrow$ 39.0	–	–
FVG-S1	32.08	16.57	8.20	8.9	9.0	0.079	0.50
FVG-S15	56.01	63.73	59.1	50.0	24.5	0.031	0.87
FVG-S20	59.19	65.95	62.9	54.8	27.1	0.029	0.89
FVG-S30	61.37	68.19	64.6	63.8	30.3	0.029	0.90
FVG-S30 Reg	66.52	81.69	72.4	118.7	38.0	0.026	0.93
FVG-S40	62.84	68.92	61.35	61.75	28.7	0.029	0.89

TABLE II
HYPERSPHERICAL GEOMETRY AND SPEAKER-RECOGNITION PERFORMANCE FOR FAST-VGAN+ECAPA (FVG). Intra $^\circ$ AND NN $^\circ$ DENOTE THE MEAN INTRA-SPEAKER AND NEAREST-PROTOTYPE ANGLES, RESPECTIVELY; “S” IS THE SCALE VALUE AND “REG” THE PROPOSED REGULARIZATION.

participation ratios and angular separation among the evaluated configurations. In this joint Fast-VGAN setting, FVG-S30 Reg also yields the lowest EER (0.026) and highest Top-5 accuracy (0.93). This contrasts with the standalone ECAPA experiment, where regularization slightly degrades these recognition metrics, indicating that the effect on speaker-recognition performance depends on the training configuration.

Except for the strongly collapsed $S = 1$ configuration, the mean intra-speaker angular distance remains smaller than the mean nearest-neighbor prototype angle. This indicates that, once sufficient inter-prototype separation is established, utterance-level embeddings form speaker-specific clusters whose angular extent remains smaller than the separation between neighboring classifier prototypes. The regularized $S = 30$ model shows the largest margin between these quantities (66.5° vs. 81.7°), further supporting the correspondence between the geometry of the classifier prototypes and that of the utterance-level embeddings.

B. Neighborhood organization with respect to speaker gender

Model	Neighborhood size k					
	10	50	100	200	300	500
FVG S1	96.42	95.54	95.68	95.31	93.10	77.60
FVG S15	94.28	85.49	75.94	64.45	57.98	51.09
FVG S20	91.96	81.90	73.07	62.81	57.17	51.16
FVG S30	89.45	78.62	70.71	61.55	56.78	51.44
FVG S30 Reg	81.83	74.18	69.44	64.68	61.47	57.12
FVG S40	67.23	67.09	69.14	61.36	55.82	49.72

TABLE III
PERCENTAGE OF SAME-GENDER SPEAKERS AMONG THE k NEAREST NEIGHBORS OF EACH SPEAKER PROTOTYPE. HIGHER VALUES INDICATE A STRONGER GENDER-BASED ORGANIZATION OF THE PROTOTYPE SPACE.

As an additional analysis, Table III reports the gender composition of local neighborhoods in the learned hyperspherical space. For each speaker prototype, we measure the fraction of same-gender speakers among its k nearest neighbors. Although gender is not used during training, the prototypes exhibit a strong gender-based organization (Table III). This trend must be interpreted together with Table II: $S = 1$ collapses to $D_{PR} = 8.9$ with very small inter-prototype separation, whereas S30 Reg reaches $D_{PR} = 118.7$ and approaches the angular dispersion of random initialization. Thus, nearest neighbors do not represent comparable angular neighborhoods

¹<https://abrassartm.github.io/Reg-VGAN/>

across configurations. Nevertheless, gender organization remains strong after regularization, with 81.8% and 74.2% same-gender speakers for $k = 10$ and $k = 50$, respectively.

C. Experiments

Datasets & training: Fast-VGAN is trained on the LibriTTS-R train-clean subset [24] (~ 250 h, 1.1k speakers), with the augmentation recipe of [3] except reverberation. Zero-shot evaluation uses unseen VCTK speakers [28] at 22.5 kHz, with 20 gender-balanced conversion pairs (40 utterances) per condition. For EZ-VC [29], Seed-VC [30], GenVC [31], and kNN-VC [32], we evaluate the authors’ pretrained checkpoints on the same VCTK pairs using the same objective pipeline; VCTK is not among their reported training corpora.

D. Objective VC evaluation

We evaluate intelligibility using WER from a pretrained NVIDIA NeMo model², speaker similarity using cosine similarity between Resemblyzer embeddings³, and perceptual quality using UTMOSv2. Reported results include 95% bootstrap confidence intervals.

Model (Params)	WER (%) $\downarrow$	Sim. $\uparrow$	UTMOSv2 $\uparrow$
Ground truth	2.33 ± 2.43	0.77 ± 0.03	—
Fast-VGAN S30 (9M)	6.54 ± 3.26	0.65 ± 0.03	2.47 ± 0.17
Fast-VGAN S30 Reg (9M)	5.97 ± 3.96	0.69 ± 0.02	2.70 ± 0.19
EZ-VC [29] (910M)	16.93 ± 4.32	0.75 ± 0.03	3.28 ± 0.11
Seed-VC [30] (440M)	4.84 ± 1.87	0.69 ± 0.03	3.06 ± 0.11
Gen-VC [31] (620M)	10.17 ± 3.50	0.66 ± 0.03	2.38 ± 0.12
KNN-VC [32] (330M)	16.71 ± 3.88	0.73 ± 0.03	2.58 ± 0.11

TABLE IV

OBJECTIVE VC EVALUATION. LOWER WER IS BETTER, WHILE HIGHER SIMILARITY AND UTMOSv2 SCORES INDICATE BETTER PERFORMANCE.

Table IV shows that regularization improves Fast-VGAN in WER (6.54% to 5.97%), similarity (0.65 to 0.69), and UTMOSv2 (2.47 to 2.70). Despite using only 9M parameters and ~ 250 h of training data, Fast-VGAN remains competitive with much larger systems trained or pretrained on substantially larger corpora (12.8k h for EZ-VC, 101k h for Seed-VC, >46 k h for GenVC-Large, and 94k h for the WavLM encoder used by kNN-VC).

E. Subjective VC evaluation

We conducted a subjective evaluation to investigate whether the position of a target speaker in the speaker embedding space affects zero-shot voice conversion quality. A Mean Opinion Score (MOS) evaluation was conducted on the Prolific⁴ crowdsourcing platform with 50 participants. Listeners rated naturalness and speaker similarity on a 5-point Likert scale. Our hypothesis is that broader embedding-space coverage is associated with better generalization, particularly for speakers farther from the training distribution. VCTK speakers are ranked by their mean angular distance to the $k = 35$ nearest training speakers and split into three equal-sized groups: *Near*, *Medium*, and *Far*. Voice conversion samples were generated

with both the baseline ECAPA system and the proposed ECAPAREg system. Table V reports the results for the three distance groups, allowing us to assess whether embedding distance correlates with conversion difficulty and whether ECAPAREg provides larger improvements for speakers farther from the training distribution.

Model	Naturalness			Similarity		
	Near	Medium	Far	Near	Medium	Far
FVG S30	2.885	2.682	2.300	2.545	2.500	1.962
FVG S30 Reg	3.048	2.857	2.714	2.762	2.667	2.143
Ground truth	3.961 ± 1.231			4.351 ± 1.106		

TABLE V

SUBJECTIVE MOS EVALUATION ON VCTK SPEAKERS GROUPED ACCORDING TO THEIR AVERAGE ANGULAR DISTANCE TO THE k NEAREST TRAINING SPEAKERS. NEAR, MEDIUM, AND FAR DENOTE INCREASING DISTANCE FROM THE TRAINING SPEAKER DISTRIBUTION.

The results in Table V show that voice conversion quality decreases as the target speaker becomes more distant from the training speaker distribution. The proposed regularized model consistently improves both naturalness and speaker similarity across all three distance groups. Moreover, the largest improvement in naturalness is observed for the *Far* speakers, suggesting that the proposed regularization is particularly beneficial for speakers located farther from the training distribution. Overall, these results are consistent with the hypothesis that broader coverage of the speaker representation space is associated with better generalization to previously unseen speakers, although they do not establish a causal relation between isotropy and conversion quality.

F. Interpolation across low- and high-density regions

We next investigate regions of the speaker space that are not directly produced by the ECAPA-TDNN encoder for individual utterances, but are reached by interpolating between pairs of unseen VCTK speaker representations. For each speaker pair, we generate a SLERP trajectory between the two endpoint embeddings and evaluate the density of its interior with respect to the LibriTTS-R training-speaker prototypes. We define *Void* trajectories as those whose intermediate points remain far from the nearest training prototype, whereas *Dense* trajectories pass through regions closer to existing prototypes. The endpoints are excluded so that the density criterion characterizes the interpolated region itself rather than the observed VCTK speaker representations. To control for trajectory length, each Void trajectory is matched with a Dense trajectory having a similar angular separation between its two endpoint speakers.

As shown in Figure 2, Dense trajectories generally yield higher UTMOS and lower log-F0 RMSE and WER. In Void regions, regularization improves UTMOS and log-F0 RMSE and tends to reduce WER, while it does not degrade the Dense controls. These results are consistent with improved robustness in poorly covered regions rather than a trade-off with well-covered regions.

²<https://github.com/NVIDIA/NeMo/>

³<https://github.com/resemble-ai/Resemblyzer>

⁴<https://www.prolific.com/>

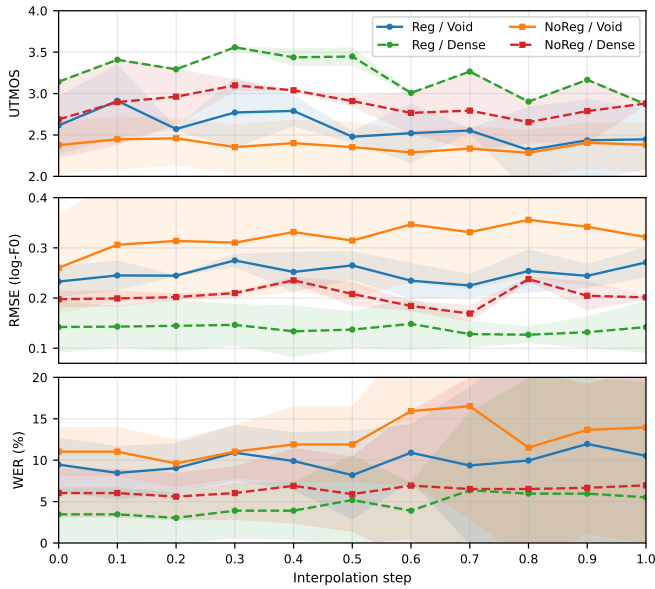


Fig. 2. Interpolation performance in low-density (*Void*) and high-density (*Dense*) regions, with trajectory pairs matched for the angular separation between their endpoint speakers.

G. Interpolation from one-shot speakers

We further investigate the robustness of the learned speaker space under limited enrollment using unseen speakers from VCTK. For each target speaker, the ECAPA-TDNN representation is extracted from a single short reference utterance of approximately one second. Interpolations are then generated between all pairs of these one-shot VCTK speaker representations using the same protocol as in the previous experiment.

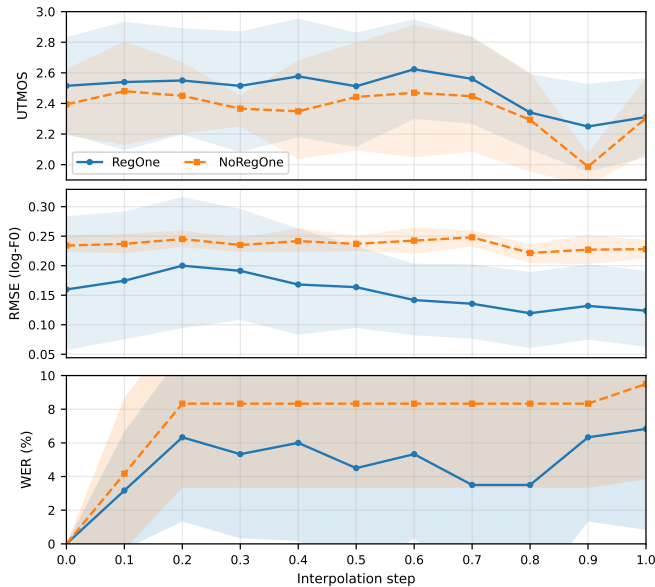


Fig. 3. Interpolation quality for speaker pairs represented by a single short enrollment utterance.

Figure 3 compares the baseline and the regularized models.

The regularized model consistently achieves higher UTMOS scores and substantially lower log-F0 RMSE across the interpolation trajectory. It also produces lower WER for nearly all interpolation coefficients, indicating that speech intelligibility is better preserved despite the limited speaker information. These results associate a more uniform speaker space with more robust interpolation under limited enrollment.

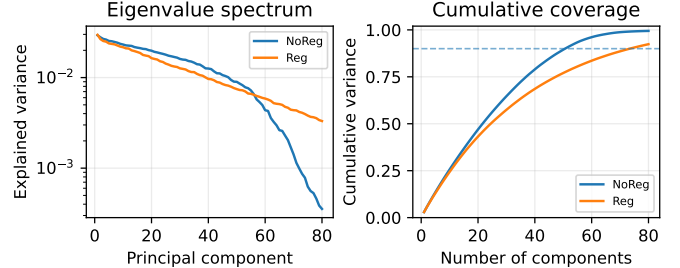


Fig. 4. PCA of the prototype space. Regularization flattens the eigenvalue spectrum and increases d_{90} from 45 to 65, indicating reduced anisotropy.

PCA further confirms this effect: regularization flattens the eigenvalue spectrum and increases the number of components required to explain 90% of the variance from 45 to 65.

V. DISCUSSION

Our experiments show that angular-margin training can produce strongly structured prototype spaces with substantial variations in angular separability, anisotropy, and effective dimensionality. The proposed geometric regularizers promote a more isotropic and higher-dimensional hyperspherical distribution by increasing angular separation and distributing variance across a larger number of dimensions. These geometric changes are associated with improved robustness in zero-shot voice conversion, particularly for unseen speakers, while speaker-recognition effects depend on the training configuration. Overall, geometry appears to provide a useful prior for zero-shot conditioning; isolating the effect of the two regularizers and training-seed variability remains future work.

VI. GENERATIVE AI USE DISCLOSURE

Generative AI has been used for improving, correcting, and compacting the English language, as well as for scripting related to data analysis and the generation of plots.

ACKNOWLEDGMENT

This work was partly performed using HPC resources from GENCI-IDRIS (Grant 2025-AD011011177R5), and partly funded by the ANR project BRUEL (ANR-22-CE39-0009).

REFERENCES

- [1] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2010.
- [2] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5329–5333, IEEE, 2018.

- [3] B. Desplanques, J. Thienpondt, and K. Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," *arXiv preprint arXiv:2005.07143*, 2020.
- [4] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4690–4699, 2019.
- [5] T. Wang and P. Isola, "Understanding contrastive representation learning through alignment and uniformity on the hypersphere," in *International conference on machine learning*, pp. 9929–9939, PMLR, 2020.
- [6] V. Pappas, X. Han, and D. L. Donoho, "Prevalence of neural collapse during the terminal phase of deep learning training," *Proceedings of the National Academy of Sciences*, vol. 117, no. 40, pp. 24652–24663, 2020.
- [7] A. Ansuini, A. Laio, J. H. Macke, and D. Zoccolan, "Intrinsic dimension of data representations in deep neural networks," *Advances in Neural Information Processing Systems*, vol. 32, 2019.
- [8] S. Recanatani, M. Farrell, M. Advani, T. Moore, G. Lajoie, and E. Shea-Brown, "Dimensionality compression and expansion in deep neural networks," *arXiv preprint arXiv:1906.00443*, 2019.
- [9] C. H. Martin and M. W. Mahoney, "Implicit self-regularization in deep neural networks: Evidence from random matrix theory and implications for learning," *Journal of Machine Learning Research*, vol. 22, no. 165, pp. 1–73, 2021.
- [10] M. Abrassart, N. Obin, and A. Roebel, "Fast-VGAN: Lightweight Voice Conversion with Explicit Control of F0 and Duration Parameters," in *Speech Synthesis Workshop*, 2025.
- [11] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 212–220, 2017.
- [12] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5265–5274, 2018.
- [13] Y. Li, F. Gao, Z. Ou, and J. Sun, "Angular softmax loss for end-to-end speaker verification," in *2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pp. 190–194, IEEE, 2018.
- [14] Y.-Q. Yu, L. Fan, and W.-J. Li, "Ensemble additive margin softmax for speaker verification," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6046–6050, IEEE, 2019.
- [15] J. Thienpondt, B. Desplanques, and K. Demuynck, "The idlab voxceleb speaker recognition challenge 2020 system description," *arXiv preprint arXiv:2010.12468*, 2020.
- [16] J. Yao, C. Liang, Z. Peng, B. Zhang, and X.-L. Zhang, "Branch-ECAPA-TDNN: A Parallel Branch Architecture to Capture Local and Global Features for Speaker Verification," in *Interspeech*, pp. 1943–1947, 2023.
- [17] B. Liu and Y. Qian, "ECAPA++: Fine-grained deep embedding learning for TDNN based speaker verification," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, pp. 3132–3136, 2023.
- [18] Z. Chen, S. Chen, Y. Wu, Y. Qian, C. Wang, S. Liu, Y. Qian, and M. Zeng, "Large-scale self-supervised speech representation learning for automatic speaker verification," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6147–6151, IEEE, 2022.
- [19] N. Dawalatabad, M. Ravanelli, F. Grondin, J. Thienpondt, B. Desplanques, and H. Na, "ECAPA-TDNN embeddings for speaker diarization," *arXiv preprint arXiv:2104.01466*, 2021.
- [20] J. Xue, Y. Deng, Y. Han, Y. Li, J. Sun, and J. Liang, "ECAPA-TDNN for Multi-speaker Text-to-speech Synthesis," in *2022 13th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pp. 230–234, IEEE, 2022.
- [21] W. Liu, R. Lin, Z. Liu, L. Liu, Z. Yu, B. Dai, and L. Song, "Learning towards minimum hyperspherical energy," *Advances in neural information processing systems*, vol. 31, 2018.
- [22] M. Kuznetsov, I. Butakov, M. Munkhoeva, A. Frolov, and I. Osledeets, "Information-Theoretic Unsupervised Embedding Quality Evaluation," in *ICLR 2026 Workshop on Representational Alignment (Re {textasciicircum} 4-Align)*, 2026.
- [23] Y. Koizumi, H. Zen, S. Karita, Y. Ding, K. Yatabe, N. Morioka, M. Bacchiani, Y. Zhang, W. Han, and A. Bapna, "Libritts-r: A restored multi-speaker text-to-speech corpus," *arXiv preprint arXiv:2305.18802*, 2023.
- [24] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, "LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech," *Interspeech 2019*, 2019.
- [25] S. Recanatani, S. Bradde, V. Balasubramanian, N. A. Steinmetz, and E. Shea-Brown, "A scale-dependent measure of system dimensionality," *Patterns*, vol. 3, p. 100555, Aug. 2022.
- [26] D. P. Hardin and E. Saff, "Discretizing Manifolds via Minimum Energy Points," *Notices of the AMS*, vol. 51, no. 10, pp. 1186–1194, 2004.
- [27] A. Roebel and F. Bous, "Neural Vocoding for Singing and Speaking Voices with the Multi-Band Excited WaveNet," *Information*, vol. 13, no. 3, p. 103, 2022.
- [28] C. Veaux, J. Yamagishi, K. MacDonald, *et al.*, "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit," *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*, vol. 6, no. 15, p. 17, 2017.
- [29] A. Joglekar, D. Singh, R. R. Bhatia, and S. Umesh, "EZ-VC: Easy Zero-shot Any-to-Any Voice Conversion," in *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 19768–19774, 2025.
- [30] S. Liu, "Zero-shot voice conversion with diffusion transformers," *arXiv preprint arXiv:2411.09943*, 2024.
- [31] Z. Cai, H. L. Xinyuan, A. Garg, L. P. García-Perera, K. Duh, S. Khudanpur, M. Wiesner, and N. Andrews, "Genvc: Self-supervised zero-shot voice conversion," *arXiv preprint arXiv:2502.04519*, 2025.
- [32] M. Baas, B. van Niekerk, and H. Kamper, "Voice conversion with just nearest neighbors," *arXiv preprint arXiv:2305.18975*, 2023.