BEYOND SPEAKER VERIFICATION: SPEAKER EMBEDDINGS FOR PERCEPTUAL SIMILARITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Speaker embedders are trained to verify whether two utterances come from the same speaker. However, the same embedders increasingly serve as evaluation metrics, reward models, and data filters for a different purpose: estimating how similar two voices sound to human listeners. We first show that verification-trained embedders do not reliably preserve perceptual similarity rankings, and that this limitation stems from the verification objective rather than from any particular model. For a speaker embedder widely used in text-to-speech (TTS) evaluation, scores on different-speaker pairs exhibit almost no correlation with human ratings. Our theoretical analysis examines the optimal embedding geometry under the verification objective, which collapses intra-speaker variation and spreads speakers uniformly. We prove that this geometry is invariant to inter-speaker perceptual similarity rankings and show empirically that existing speaker embedders closely approach it. Building on this analysis, we propose **Simpson**, a speaker embedder that measures **Sim**ilarity like a **p**er**son**. Simpson augments the verification objective with a ranking objective to learn perceptual similarity rankings without human similarity labels. The ordering targets are derived from a frozen self-supervised backbone and Simpson's own exponential moving average (EMA) teacher. An optional preference stage further refines the model using a small set of human similarity labels. Across natural, converted, and synthetic speech, Simpson achieves the highest correlation with human ratings among the publicly available embedders evaluated, reaching a Spearman rank correlation of 0.79 on different-speaker pairs where verification-trained embedders show near-zero correlation. We further demonstrate Simpson's practical effectiveness in downstream tasks, both as an evaluation metric for TTS and as a reward function for group relative policy optimization (GRPO). Audio samples are provided on an anonymous demo page[1].

1 INTRODUCTION

A speaker embedding is a fixed-length vector representation that encodes the characteristics of a speaker from an utterance. Speaker embeddings are conventionally trained for speaker verification, the task to decide whether two utterances come from the same speaker. With advances in objectives and architectures, speaker verification performance has become sufficient for commercial use. Beyond verification, the cosine similarity between embeddings (SIM) is now commonly used in place of human perceptual judgment to evaluate synthesized speech, to compute rewards for reinforcement learning, and to filter data. Such tasks require a graded measure of perceived similarity between two voices rather than a binary decision on speaker identity. Using verification-trained embedders for these tasks relies on the implicit assumption that an embedding accurate at separating speakers is also accurate at measuring how similar two voices sound.

However, embedding-based similarity has repeatedly been reported to diverge from human ratings of speaker similarity (Jia et al., 2025; Chen et al., 2025; Moon et al., 2026). For example, the embedder with the lower verification equal error rate (EER) showed a lower correlation between its similarity scores and human similarity ratings than the embedder with the higher EER (Ahn et al., 2024). To examine this discrepancy more closely, we separate human-labeled utterance pairs into

[1] https://simpson-demo.bitter-scene-cb40.workers.dev

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

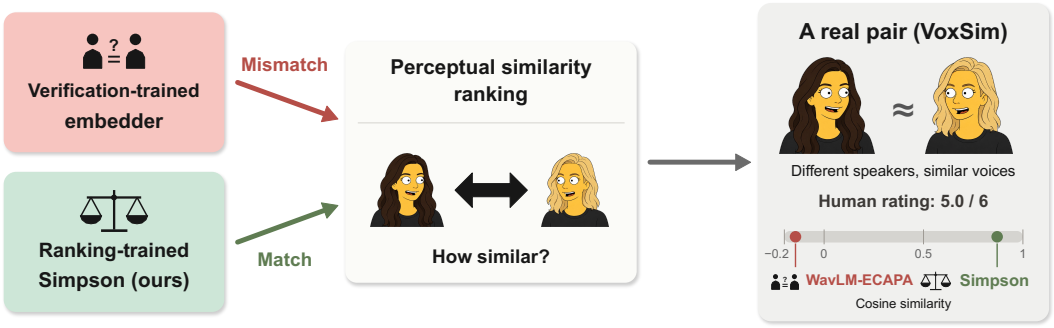


Figure 1: (Left): existing speaker embedders are trained for verification but used for perceptual similarity ranking, whereas Simpson is trained directly for ranking. (Right): a VoxSim pair of two different speakers rated 5 on a 6-point scale; WavLM(Large)-ECAPA (Existing) gives a cosine similarity of -0.143, whereas Simpson gives 0.871.

two types, same-speaker pairs and different-speaker pairs. Within different-speaker pairs, the correlation between embedding-based similarity and human ratings is close to zero (Figure 1). These observations suggest that learning to discriminate between speakers is distinct from learning the similarity between utterances. This inconsistency has been reported repeatedly, but its cause remains underexplored.

To identify the cause, we formalize the optimum of the verification objective for speaker embedders. At the optimum, the utterances of the same speaker collapse to one point, and the points of different speakers spread evenly on the hypersphere. We show that this geometry is invariant to the perceptual similarity ranking among speakers. We further demonstrate empirically that current embedders closely approach this geometry. Motivated by this analysis, we propose **Simpson**, a speaker embedder trained for perceptual similarity. Simpson learns to order utterance pairs by perceptual similarity with a loss that directly optimizes the pairwise ranking. Because human similarity labels are costly to collect and vary across annotators, we design Simpson to be trained without human labels. Instead, we use proxy similarities from two sources. One source is a pre-trained backbone, frozen at layers chosen by a layer-wise analysis; the other is an exponential-moving-average (EMA) teacher of Simpson itself, whose proxy similarities improve with the student. When a small set of human preference labels is available, a preference-optimization stage (Rafailov et al., 2023) further improves Simpson.

As a result, Simpson achieves the highest correlation with human ratings of perceptual similarity among public embedders across human-labeled test sets of natural speech, converted speech, and synthetic speech. On the different-speaker pairs of VoxSim Ahn et al. (2024), where the widely used WavLM(Large)-ECAPA reaches a correlation of 0.06, Simpson reaches 0.79 without any human similarity label. On the downstream tasks of speaker embedders, Simpson also produces results more consistent with human preference than those of existing embedders. As an evaluation metric, Simpson ranks TTS systems closer to listener ratings. As a reward for reinforcement learning to improve speaker similarity in TTS, Simpson produces a policy that synthesizes speech rated by listeners as more similar to the reference speaker.

Our contributions are as follows:

- We show that current speaker embedders fail to rank pairs of voices by similarity. Because of the verification objective, the learned geometry does not depend on the perceptual similarity ranking among speakers.
- We propose Simpson, a speaker embedder that encodes perceptual similarity in its embedding space. To our knowledge, Simpson is the first speaker embedder trained with a ranking objective, and the training requires no human similarity labels.
- We empirically demonstrate that Simpson ranks pairs of voices most consistently with human listeners among public embedders on natural, converted, and synthetic speech. On downstream tasks, Simpson further outperforms verification-trained embedders.

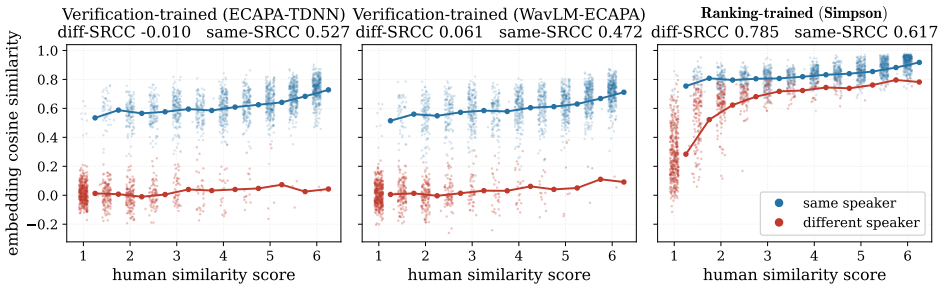


Figure 2: Human similarity rating (horizontal) versus embedding cosine (vertical); bold lines are bin means. For the verification-trained embedders (left, middle), cosines of different-speaker pairs (red) do not increase with the human rating; for Simpson (right) they increase for both pair types.

2 Motivation

2.1 The two tasks of a speaker embedder: verification vs. ranking

Let f_θ map utterances x and y to d-dimensional unit vectors $f_\theta(x)$ and $f_\theta(y)$, and let $s(x, y) = f_\theta(x)^\top f_\theta(y)$ denote their similarity. The first task of a speaker embedder is **verification**, deciding whether two utterances come from the same speaker by thresholding $s(x, y)$. Verification performance is measured by the EER. The second task is **ranking of perceived similarity**, ordering utterance pairs by how similar the two voices sound. Formally, given utterance pairs $\{(x_i, y_i), (x_j, y_j)\}$ with human similarity ratings $\{h_i, h_j\}$, the predicted similarities should follow the order of the ratings, that is, $s(x_i, y_i) > s(x_j, y_j)$ whenever $h_i > h_j$. Ranking performance is measured by the Spearman rank correlation between s and h.

Verification requires only that two speakers be separated, whether or not they sound alike. In contrast, ranking requires that pairs be ordered by how alike they sound. The verification metric itself does not measure ranking, so an embedder's EER should not be interpreted as a proxy for how the embedder ranks pairs by similarity.

2.2 Verification-trained embedders do not rank like humans

Figure 2 plots embedding cosines against the similarity ratings that listeners assigned to the same pairs. The two verification-trained embedders separate same-speaker pairs from different-speaker pairs but encode no further distinction within the different-speaker pairs. In other words, their cosines do not distinguish the pairs that listeners rate as similar from the pairs rated as dissimilar, and the correlation with the human ratings is near zero. Only our ranking-trained embedder provides cosines that increase with the human ratings.

2.3 The geometry of verification and ranking optima

Since similarity is generally measured by the cosine, every embedding can be assumed to lie on the unit sphere $\mathbb{S}^{d-1}$. Verification losses mainly fall into two families, classification losses (Desplanques et al., 2020; Yakovlev et al., 2024) and contrastive losses (Chung et al., 2020; Zhang & Yu, 2022), and the optimum of each family has been characterized. Cross-entropy and its margin variants drive each class to one point and spread the class means as evenly as possible on the sphere (Papyan et al., 2020; Jiang et al., 2024), and contrastive losses induce the same alignment and uniformity (Wang & Isola, 2020). For N speakers, where $N \gg d$, a classical bound gives a lower limit on the mean squared cosine between speakers.

Proposition 1 (Verification bound). *For any $N > d$ unit vectors $x_1, \ldots, x_N$ in $\mathbb{R}^d$, the mean squared cosine over pairs satisfies $\frac{1}{\binom{N}{2}} \sum_{i<j} (x_i^\top x_j)^2 \geq \frac{N-d}{d(N-1)}$, with equality if and only if the vectors are spread evenly over every direction, $\sum_i x_i x_i^\top = \frac{N}{d} I$ (Benedetto & Fickus, 2003; Welch, 1974).*

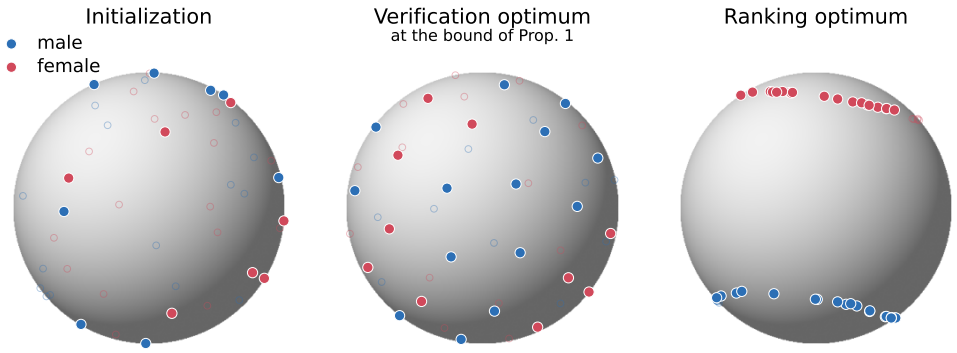


Figure 3: A toy example with $N > d$: 48 points on $\mathbb{S}^2$ with a two-level target similarity (male/female group, then an order within each group). Middle: a verification loss reaches the bound of Proposition 1, an even spread with no order. Right: from the same start, the ranking loss on the target similarity arranges the points into two arcs, separating the two groups by a distance that Corollary 1 bounds at the verification optimum.

Corollary 1. *When the bound is attained, for any partition of the speakers into groups of sizes N_A and N_B with group means μ_A and μ_B, the centroid gap satisfies $\|\mu_A - \mu_B\| \leq \frac{N}{\sqrt{N_A N_B}} \cdot \frac{1}{\sqrt{d}}$, which is $2/\sqrt{d}$ for a balanced partition.*

Proofs are in Appendix B.8 (the proposition is the frame-potential bound). Both statements are invariant to the ordering of the speakers, so the optimum is the same for every ordering among them. This invariance is why a verification-trained embedder does not encode a perceptual ranking. A ranking is expressed through differences between cosines, but at the bound the cosines have a root-mean-square of only about $1/\sqrt{d}$ (0.06 for $d = 256$), and their differences are typically of the same order. Figure 3 demonstrates this for the most salient perceptual distinction: gender. At the verification optimum, the male and female centroids lie within $2/\sqrt{d}$ of each other (middle), whereas the ranking optimum separates them (right).

3 SIMPSON: A SPEAKER EMBEDDER FOR PERCEPTUAL SIMILARITY

Figure 4 illustrates the overall architecture. The backbone is WavLM-Large (Chen et al., 2022) truncated to its three lowest transformer layers. We keep the lower layers because, in a pre-trained WavLM, the correlation between layer features and human ranking judgments increases up to the third layer and decreases with depth (Appendix Figure 6). Truncation reduces the backbone size from 317M to 51M parameters. The remaining layers stay frozen and are adapted only through LoRA (Hu et al., 2022b).

3.1 ARCHITECTURE

The readout follows the standard speaker-embedding head with one modification. The per-channel statistics that instance normalization (IN) (Ulyanov et al., 2016) removes before pooling also carry the overall spectral character of a voice. We compensate for their removal through a separate path,

$$\text{emb} \;=\; \text{ECAPA}\big(\text{IN}(g^{(\ell^*)})\big) \;+\; \sum_{\ell=0}^{\ell^*} W_\ell \,[\, m_\ell;\, v_\ell \,], \tag{1}$$

where $g^{(\ell)}$ is the hidden state of layer ℓ, m_ℓ and v_ℓ are its per-channel mean and standard deviation before normalization, and ℓ^* is the final layer. We maintain a teacher network with the same architecture alongside the student (Tarvainen & Valpola, 2017). The student parameters θ^{s} are updated by gradient descent, and the teacher parameters follow $\theta^{\text{t}} \leftarrow \lambda\theta^{\text{t}} + (1-\lambda)\theta^{\text{s}}$ with λ increasing from 0.9995 to 1 on a cosine schedule. The teacher receives a random 3-s segment of each utterance, and the student receives a random 1-s segment within the teacher's segment. The teacher supplies proxy similarities for the ranking loss (§3.3) and remains as the final embedder after training.

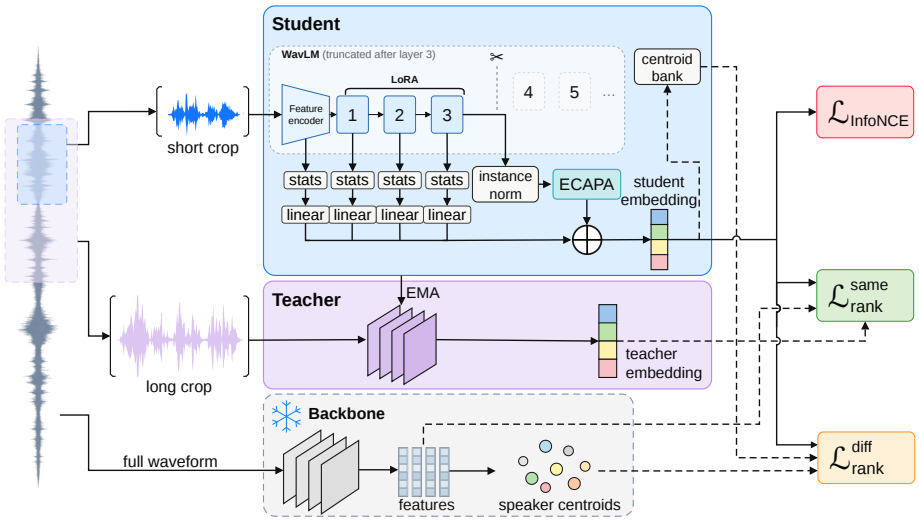


Figure 4: Overview of the proposed method. $\mathcal{L}^{\text{diff}}_{\text{rank}}$ takes its proxy similarities from speaker centroids of a frozen backbone mixed with a bank of student centroids, and $\mathcal{L}^{\text{same}}_{\text{rank}}$ from the utterance features of the frozen backbone and the EMA teacher (Eq. 3). Dashed arrows carry no gradient.

3.2 Contrastive Objective

The training objective consists of two types of losses, a contrastive loss and ranking losses. The first is InfoNCE (Oord et al., 2018) on speaker identity, with another utterance of the same speaker as the positive and the other speakers in the batch as the negatives, supervised only by speaker labels:

$$\mathcal{L}_{\text{InfoNCE}} = -\frac{1}{2B}\sum_{i=1}^{B}\left[\log\frac{e^{s_{ii}}}{\sum_{j=1}^{B} e^{s_{ij}}} + \log\frac{e^{s_{ii}}}{\sum_{j=1}^{B} e^{s_{ji}}}\right], \qquad s_{ij} = f^{\text{s}}_{\theta}(x_i)^{\top} f^{\text{s}}_{\theta}(x'_j)/\tau, \tag{2}$$

where x_i and x'_i are two utterances of speaker i in a batch of B speakers, f^{s}_{θ} is the student, and τ is a temperature. Trained with this term alone, the embedder would approach the optimum of §2.3, in which the utterances of each speaker collapse to one point and the points spread evenly.

3.3 Ranking Objective

Because Simpson is trained without human similarity labels, the ranking objective replaces the human ratings with proxy similarities, similarity-like measurements used only to determine the relative order of pairs. We compute the proxy similarities for two sets of pairs, different-speaker pairs S_{diff} and same-speaker pairs S_{same}, from the frozen backbone and the teacher. For a pair $p = (x, y)$, the proxy similarities are

$$\begin{aligned} t^{\text{diff}}_{p} &= (1-\alpha_{\text{d}})\, z\big(c_x^{\top} c_y\big) + \alpha_{\text{d}}\, z\big(\tilde{c}_x^{\top} \tilde{c}_y\big), \\ t^{\text{same}}_{p} &= (1-\alpha_{\text{s}})\, z\big(f^{\text{fr}}(x)^{\top} f^{\text{fr}}(y)\big) + \alpha_{\text{s}}\, z\big(f^{\text{t}}_{\theta}(x)^{\top} f^{\text{t}}_{\theta}(y)\big), \end{aligned} \tag{3}$$

where $f^{\text{fr}}(x)$ is the utterance feature of the frozen backbone, c_x and $\tilde{c}_x$ are the centroids of the frozen features and of the student embeddings for the speaker of x, looked up by speaker label (Eq. 7 in Appendix A.5), f^{t}_{θ} is the EMA teacher, and z is the standardization operator over the pairs in the batch. For a same-speaker pair, a centroid cosine is identical for every pair of that speaker and carries no order, so the proxy must come from the utterances. For a different-speaker pair, the order concerns the two voices rather than the two recordings, so the proxy comes from speaker centroids, which order pairs better than single utterances (0.72 against 0.42, Appendix Figure 6).

At the start of training, the frozen backbone is the more reliable source, whereas the student centroids and the teacher become reliable as training proceeds. The weights α_{d} and α_{s} therefore increase over training, which prevents the frozen features from limiting the proxy similarities.

Let $S = \{(x_p, y_p)\}$ be a set of utterance pairs with model scores $s_p = s(x_p, y_p)$ and proxy similarities $T = \{t_p\}$. The ranking loss takes the RankNet form (Burges et al., 2005) and penalizes misordered comparisons (p, q) sampled from the batch:

$$\mathcal{L}_{\text{rank}}(S;T) = \frac{1}{Z}\sum_{p,q} w_{pq}\,\text{softplus}\left(-\text{sign}(t_p - t_q)\frac{s_p - s_q}{\beta}\right), \quad w_{pq} = |t_p - t_q|, \;\; Z = \sum_{p,q} w_{pq}. \tag{4}$$

The sign uses only the order of the proxy similarities, and the factor $|t_p - t_q|$ weights each comparison by the difference of the standardized proxy similarities. The different-speaker term prevents the even spread of speakers and the same-speaker term prevents the collapse of a speaker's utterances, the two properties of the verification optimum in §2.3.

The full objective adds the contrastive loss of §3.2 to the two ranking losses:

$$\mathcal{L} = \mathcal{L}_{\text{InfoNCE}} + \lambda_d \underbrace{\mathcal{L}_{\text{rank}}(S_{\text{diff}}; t^{\text{diff}})}_{\mathcal{L}^{\text{diff}}_{\text{rank}}} + \lambda_s \underbrace{\mathcal{L}_{\text{rank}}(S_{\text{same}}; t^{\text{same}})}_{\mathcal{L}^{\text{same}}_{\text{rank}}}. \tag{5}$$

3.4 Preference-based refinement

Training so far has used no human similarity label. This section describes an optional refinement stage for the case in which human labels are available. Public similarity labels are few and vary considerably across raters, and fitting their values may transfer such noise into the model. Therefore, we adapt direct preference optimization (DPO) (Rafailov et al., 2023) to learn the human ratings as a ranking, using the labels as preferences rather than as ground truth. Each comparison is weighted by the difference of the human scores, $|h_p - h_q|$. In place of the likelihood ratio, the loss uses the margin of embedding scores, where s is the cosine of the model being refined and $\bar{s}$ that of its frozen copy:

$$\mathcal{L}_{\text{pref}} = \frac{1}{Z}\sum_{p,q} |h_p - h_q|\,\text{softplus}\left(-\text{sign}(h_p - h_q)\frac{(s_p - s_q) - (\bar{s}_p - \bar{s}_q)}{\beta}\right), \quad Z = \sum_{p,q} |h_p - h_q|, \tag{6}$$

where h is the human similarity score. The frozen reference regularizes the model so that a small, noisy label set cannot degrade the ordering learned without labels. This stage, like §3.3, supervises the ranking of the pairs; only the source of the ranking changes.

4 Experiments

4.1 Setup

We evaluate Simpson on three public test sets that carry human similarity labels, one for each origin of speech: natural speech (VoxSim (Ahn et al., 2024)), voice conversion (VCC2018 (Lorenzo-Trueba et al., 2018) and VCC2020 (Yi et al., 2020)), and speech synthesis (TTSDS2 (Minixhofer et al., 2026)). We follow the official protocol of each set and take the Spearman rank correlation coefficient (SRCC) as the primary metric. The linear correlation coefficient (LCC) is reported alongside, and the mean squared error (MSE) and accuracy (ACC) report absolute accuracy after a monotone mapping fitted on the official train split. The baselines are eight public speaker embedders trained without human similarity labels. We additionally take VoxSim-ft (Ahn et al., 2024), fine-tuned on such labels, as a reference upper bound rather than as a baseline. Appendices A.2–A.4 give the sizes of the data sets, the definitions of the metrics, and the checkpoint sources of the baselines. We train with AdamW (Loshchilov & Hutter, 2019), and the loss weights and schedules are listed in Appendix A.5. The choice of proxy similarities and the loss terms are ablated in Appendix C.1.

4.2 Main results

Table 1 compares the embedders on natural speech. Simpson surpasses every verification-trained embedder on all four metrics with no human similarity label (details in Appendix Table 20), and a small number of preference labels (DPO) narrows the remaining gap to VoxSim-ft on every axis. The gap is also visible in binary decisions. Among comparisons that listeners separate by at least

Table 1: Ranking on natural speech (VoxSim test). The first row is a hypothetical verifier that scores only speaker identity, so the within-class columns are undefined.

	different speaker (878)		same speaker (1,898)		full (2,776 pairs)			
Model	SRCC ↑	LCC ↑	SRCC ↑	LCC ↑	SRCC ↑	LCC ↑	MSE ↓	ACC ↑
Perfect verifier	*N/A*	*N/A*	*N/A*	*N/A*	*0.698*	*0.703*	*1.572*	*23.7*
CAM++	0.155	0.336	0.513	0.493	0.756	0.777	1.285	27.1
ERes2Net	0.102	0.326	0.491	0.473	0.745	0.772	1.308	26.2
Fast ResNet-34 (AP)	0.452	0.532	0.521	0.500	0.774	0.790	1.242	32.0
ResNet-34 (Softmax+AP)	0.333	0.403	0.522	0.504	0.769	0.783	1.273	30.8
WavLM(Base)-ECAPA	0.082	0.279	0.516	0.491	0.753	0.775	1.296	26.7
WavLM(Large)-ECAPA	0.061	0.245	0.472	0.449	0.736	0.763	1.357	25.6
ReDimNet-B3	0.049	0.248	0.491	0.469	0.741	0.768	1.323	26.8
ECAPA-TDNN	-0.010	0.231	0.527	0.502	0.751	0.776	1.253	27.4
Simpson	**0.785**	**0.753**	**0.617**	**0.593**	**0.822**	**0.827**	**0.998**	**46.6**
Simpson + DPO	**0.785**	**0.764**	**0.642**	**0.618**	**0.833**	**0.838**	**0.923**	**48.1**
VoxSim-ft (human labels)	*0.758*	*0.756*	*0.648*	*0.629*	*0.837*	*0.842*	*0.914*	*46.8*

Table 2: Ranking on voice conversion (VCC2018 and VCC2020), following the official SVSNet pair lists and aggregation. Per-subcondition numbers are in Appendix Tables 18 and 19.

	VCC2018 (5,600 groups)				VCC2020 (6,080 groups)			
Model	SRCC ↑	LCC ↑	MSE ↓	ACC ↑	SRCC ↑	LCC ↑	MSE ↓	ACC ↑
CAM++	0.468	0.490	0.936	30.8	0.694	0.703	0.280	62.7
ERes2Net	0.490	0.506	0.916	31.6	0.679	0.696	0.286	63.6
Fast ResNet-34 (AP)	0.510	0.522	0.896	31.7	0.621	0.647	0.331	58.2
ResNet-34 (Softmax+AP)	0.474	0.493	0.932	31.1	0.693	0.707	0.279	64.3
WavLM(Base)-ECAPA	0.428	0.457	0.974	30.6	0.666	0.674	0.308	61.6
WavLM(Large)-ECAPA	0.400	0.435	0.998	30.3	0.647	0.654	0.320	60.1
ReDimNet-B3	0.474	0.497	0.927	31.8	0.689	0.695	0.287	63.3
ECAPA-TDNN	0.484	0.502	0.921	30.8	0.683	0.699	0.284	63.3
Simpson	**0.547**	**0.555**	**0.852**	**41.4**	**0.739**	**0.752**	**0.329**	**63.7**
Simpson + DPO	**0.569**	**0.575**	**0.824**	**42.8**	**0.752**	**0.773**	**0.275**	**67.4**
VoxSim-ft (human labels)	*0.564*	*0.571*	*0.831*	*40.2*	*0.685*	*0.717*	*0.283*	*64.5*

one point, ECAPA-TDNN selects the pair rated less similar 50.2% of the time, at chance level, while Simpson reverses the order in only 8.1% and is the most accurate label-free embedder in every score band (Appendix Table 14).

In the full-correlation setup, the verification-trained embedders achieve an SRCC of 0.74–0.77. Most of this correlation comes from separating different-speaker pairs from same-speaker pairs. A hypothetical perfect verifier that assigns one constant similarity to every different-speaker pair and a higher constant to every same-speaker pair already reaches 0.698 (first row). Nevertheless, on different-speaker pairs alone the correlations fall to near zero for the widely used verification-trained embedders. WavLM(Large)-ECAPA, the embedder most widely used for TTS evaluation, reaches only 0.061 and ECAPA-TDNN falls below zero (Figure 2, Appendix Table 15, Appendix C.7, and Appendix Table 17).

Table 2 reports the same evaluation on voice conversion. In the Voice Conversion Challenges (VCC), listeners rated how closely each converted utterance resembles the target speaker, so the labels rank converted voices against real ones. Simpson again outperforms every verification-trained embedder on both VCC2018 and VCC2020. The lead on SRCC and LCC holds across every subcondition of reference-speaker role and language, including the cross-lingual condition, where the target speakers speak languages not included in the English-only training data (Appendix Tables 18 and 19).

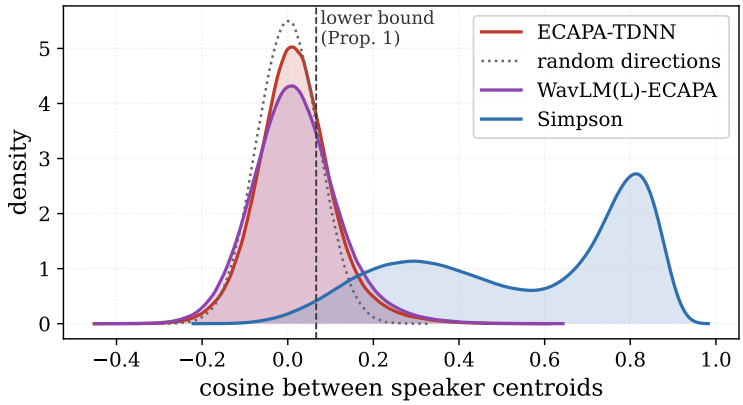


Figure 5: Distribution of speaker-centroid cosines over all speaker pairs of VoxCeleb1. Dashed: the bound of Proposition 1; dotted: randomly drawn directions. Exact numbers are in Appendix Table 9.

4.3 The geometry of the learned space

Figure 5 shows the speaker-centroid cosine distributions of trained embedders alongside the theoretical bound of Proposition 1. The cosines of the verification-trained embedders concentrate near the bound, at 1.2–1.9$\times$ its value and close to randomly drawn directions, whereas the cosines of Simpson lie about ten times above the bound (Appendix Table 9). The pull toward the bound is visible during training as well: a verification head trained on frozen backbone features lowers the different-speaker correlation with the human order from 0.417 to 0.102 while the EER improves (Appendix Figure 7).

To assess the perceptual relevance of this geometric difference, we also compare speaker-centroid cosines with human similarity ratings. Over the 795 speaker pairs that listeners rated, the rank correlation between centroid cosines and human ratings is 0.783 for Simpson against -0.007 for ECAPA-TDNN (Appendix B.5). These results suggeset that Simpson's speaker-centroid geometry reflects the perceptual similarity ranking among speakers.

5 Downstream tasks

5.1 As an evaluation metric for TTS

Speaker similarity is a standard metric for evaluating TTS systems, where the cosine between synthesized and reference speech (SIM) is reported in place of listener ratings. The choice of embedder for SIM may change the conclusions of an evaluation over the same set of systems. Here we compare Simpson against the two verification-trained embedders most widely used to compute SIM. TTS papers report SIM per system, so we compare the system ranking each embedder produces against the human similarity mean opinion score (SMOS) over the twenty TTS systems of TTSDS2 (Table 3). WavLM(Large)-ECAPA, the embedder most widely used for SIM, reaches a system-level rank correlation of 0.420, while Simpson with preference refinement reaches 0.720. Simpson also stays above the verification-trained embedders over the 91 individual utterances, at 0.415, and at 0.452 with preference refinement, against 0.320 for WavLM(Large)-ECAPA.

5.2 As a reward in reinforcement learning

Speaker similarity is also used as a reward when synthesis models are fine-tuned with reinforcement learning, such as GRPO (Shao et al., 2024), to improve similarity to the reference speaker. We fine-tune Llasa-1B (Ye et al., 2025) with GRPO, changing only the embedder used for the SIM term of the reward, and evaluate the two resulting policies in a listening test. Listeners prefer the policy trained with the ranking reward and rate it the closer match to the reference speaker (Table 5; Appendix A.6).

Table 3: Ranking on TTS synthetic speech (TTSDS2): correlation with human SMOS over twenty TTS systems on 91 clean-domain utterances (details in Appendix Table 16).

	Utterance level		System level	
Model	SRCC ↑	LCC ↑	SRCC ↑	LCC ↑
WavLM(Large)-ECAPA	0.320	0.270	0.420	0.368
ECAPA-TDNN	0.308	0.309	0.525	0.418
Simpson	0.415	0.398	0.583	0.577
Simpson + DPO	**0.452**	**0.465**	**0.720**	**0.618**
VoxSim-ft (human labels)	*0.406*	*0.377*	*0.612*	*0.563*

Table 4: Impersonation-risk screening on VCC2020: how often each embedder misses conversions that humans judged clearly similar, at matched false-alarm rates.

		Miss rate (%) ↓	
Model	AUC ↑	FA 10%	FA 5%
ECAPA-TDNN	0.9864	3.61	7.48
WavLM(Large)-ECAPA	0.9802	4.78	10.19
Simpson	0.9934	1.08	4.51
Simpson + DPO	**0.9949**	**0.99**	**2.89**

Table 5: GRPO with only the reward embedder varied. Preference: A/B choice between the two policies. SMOS: mean with 95% CI. Length gap: mean absolute duration difference from the real recording. Best in bold.

Reward embedder	Preference (%) ↑	SMOS ↑	SIM (VoxSim-ft) ↑	WER (%) ↓	Length gap (s) ↓
None (Llasa-1B)	N/A	3.88 $\pm$0.09	0.761	4.32	1.49
WavLM(Large)-ECAPA	42.2	3.94 $\pm$0.09	0.781	3.02	1.51
Simpson	**57.8**	**4.07** $\pm$0.08	**0.788**	**2.15**	**0.92**

The ranking reward increases SIM by a small margin (Table 5), yet listeners prefer the policy trained with it. It is known that similarity rewards are susceptible to reward hacking, in which the policy lengthens its outputs to increase SIM (Shin et al., 2026); the policy trained with the verification reward directly exhibits this behavior (length gap 1.51 s). In contrast, the policy trained with the ranking reward exhibits no such behavior without any length penalty or termination heuristic. Simpson's outputs stay close to the real recordings in duration (0.92 s), fewer of them are over-long or unterminated, and their word error rate is lower (paired difference -0.87 percentage points, 95% CI $[-1.56, -0.23]$). See Appendix B.3 and B.4 for further analyses.

5.3 As an impersonation-risk screening filter

Impersonation-risk screening requires the degree of similarity between two voices, which a binary verification decision does not report. Treating the converted speech of VCC2020 as impersonation attempts, we test how well each embedder separates the 1,109 groups that humans judged clearly similar from the 529 judged clearly dissimilar (Table 4), using the rank-based AUC and the miss rate at a matched false-alarm rate. At a ten-percent false-alarm rate, the operating point relevant to deployment, Simpson misses 1.08% of the attempts that listeners rated clearly similar, against 3.61% for ECAPA-TDNN and 4.78% for WavLM(Large)-ECAPA. Therefore, a low verification error alone does not guarantee an effective screen.

6 Conclusion

We presented Simpson, a speaker embedder that scores speaker similarity consistently with human listeners. We showed that the verification objective does not preserve perceptual similarity rankings between voices. The optimum of the verification objective is invariant to the order among speakers and among the utterances of one speaker, and current embedders lie close to that optimum. Simpson learns this order without human similarity labels and ranks closest to listeners among public embedders on natural, converted, and synthetic speech. A small set of preference labels improves the ranking further. Because Simpson ranks pairs of voices consistently with listeners, it outperforms verification-trained embedders on the downstream tasks that rely on a similarity score, namely TTS evaluation, reinforcement learning rewards, and impersonation-risk screening. We proposed that, where a similarity score substitutes for the judgment of a listener, the embedder should be trained for ranking rather than for verification.

AI USE STATEMENT

In this work, we used generative AI tools only for language editing to improve the readability of the manuscript. We did not use generative AI tools for any other task requiring disclosure, such as developing theory, designing or implementing methods, or analyzing results. We have reviewed and revised all AI-assisted text, and we take responsibility for the final content of this work.

ETHICS STATEMENT

The human evaluation in §5.2 involved only low-risk tasks, listening to short speech samples and rating them. Participants were 20 anonymous adult volunteers recruited from the authors' institution, who were informed of the purpose of the study and consented to participate; participation was unpaid and posed no physical or psychological risk, and no personal information was collected beyond the ratings. All other data used in this work are publicly available datasets. As with existing speaker embedders, a similarity score that follows listeners can strengthen impersonation screening but could also serve as a reward for voice cloning; we will release the model for evaluation and screening.

REPRODUCIBILITY STATEMENT

We will release the code and model checkpoints. We provide training settings in Appendix A.5, the datasets, metrics, and baseline checkpoints in Appendices A.2–A.4, the variation across training seeds in Appendix C.3, and the proofs in Appendix B.8.

REFERENCES

Junseok Ahn, Youkyum Kim, Yeunju Choi, Doyeop Kwak, Ji-Hoon Kim, Seongkyu Mun, and Joon Son Chung. VoxSim: A perceptual voice similarity dataset. In *Interspeech 2024*, pp. 2580–2584, 2024. doi: 10.21437/Interspeech.2024-646.

Philip Anastassiou, Jiawei Chen, Jitong Chen, Yuanzhe Chen, Zhuo Chen, Ziyi Chen, Jian Cong, Lelai Deng, Chuang Ding, Lu Gao, et al. Seed-tts: A family of high-quality versatile speech generation models. *arXiv preprint arXiv:2406.02430*, 2024.

Takanori Ashihara, Marc Delcroix, Takafumi Moriya, Kohei Matsuura, Taichi Asami, and Yusuke Ijima. What do self-supervised speech and speaker models learn? new findings from a cross model layer-wise analysis. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 10166–10170. IEEE, 2024.

John J Benedetto and Matthew Fickus. Finite normalized tight frames. *Advances in Computational Mathematics*, 18(2):357–385, 2003.

Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning*, pp. 89–96, 2005.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022.

Yafeng Chen, Siqi Zheng, Hui Wang, Luyao Cheng, Qian Chen, and Jiajun Qi. An Enhanced Res2Net with Local and Global Feature Fusion for Speaker Verification. In *Interspeech 2023*, pp. 2228–2232, 2023. doi: 10.21437/Interspeech.2023-1294.

Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, JianZhao JianZhao, Kai Yu, and Xie Chen. F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6255–6271, 2025.

Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. VoxCeleb2: Deep Speaker Recognition. In *Interspeech 2018*, pp. 1086–1090, 2018. doi: 10.21437/Interspeech.2018-1929.

Joon Son Chung, Jaesung Huh, Seongkyu Mun, Minjae Lee, Hee-Soo Heo, Soyeon Choe, Chiheon Ham, Sunghwan Jung, Bong-Jin Lee, and Icksang Han. In Defence of Metric Learning for Speaker Recognition. In *Interspeech 2020*, pp. 2977–2981, 2020. doi: 10.21437/Interspeech.2020-1064.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification. In *Interspeech 2020*, pp. 3830–3834, 2020. doi: 10.21437/Interspeech.2020-2650.

Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, et al. Cosyvoice 2: Scalable streaming speech synthesis with large language models. *arXiv preprint arXiv:2412.10117*, 2024.

Zhihao Du, Changfeng Gao, Yuxuan Wang, Fan Yu, Tianyu Zhao, Hao Wang, Xiang Lv, Hui Wang, Chongjia Ni, Xian Shi, et al. Cosyvoice 3: Towards in-the-wild speech generation via scaling-up and post-training. *arXiv preprint arXiv:2505.17589*, 2025.

Bing Han, Zhengyang Chen, and Yanmin Qian. Self-supervised learning with cluster-aware-dino for high-performance robust speaker verification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:529–541, 2023.

Cheng-Hung Hu, Yu-Huai Peng, Junichi Yamagishi, Yu Tsao, and Hsin-Min Wang. Svsnet: An end-to-end speaker voice similarity assessment model. *IEEE Signal Processing Letters*, 29:767–771, 2022a.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022b. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Dongya Jia, Zhuo Chen, Jiawei Chen, Chenpeng Du, Jian Wu, Jian Cong, Xiaobin Zhuang, Chumin Li, Zhen Wei, Yuping Wang, and Yuxuan Wang. DiTAR: Diffusion transformer autoregressive modeling for speech generation. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=8tRtweTTwv.

Jiachen Jiang, Jinxin Zhou, Peng Wang, Qing Qu, Dustin G. Mixon, Chong You, and Zhihui Zhu. Generalized neural collapse for a large number of classes. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=D4B7kkB89m.

Minoru Kishi, Hayato Yagi, Shinnosuke Takamichi, and Yuki Saito. Do speech foundation models perceive speaker similarity as humans do? *arXiv preprint arXiv:2606.05739*, 2026.

Suyuan Liu, Molly Babel, and Jian Zhu. A comparison of voice similarity through acoustics, human perception and deep neural network (DNN) speaker verification systems. In *Interspeech 2024*, pp. 3674–3678, 2024. doi: 10.21437/Interspeech.2024-2331.

Jaime Lorenzo-Trueba, Junichi Yamagishi, Tomoki Toda, Daisuke Saito, Fernando Villavicencio, Tomi Kinnunen, and Zhenhua Ling. The Voice Conversion Challenge 2018: Promoting Development of Parallel and Nonparallel Methods . In *The Speaker and Language Recognition Workshop (Odyssey 2018)*, pp. 195–202, 2018. doi: 10.21437/Odyssey.2018-28.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Christoph Minixhofer, Ondrej Klejch, and Peter Bell. Ttsds2: resources and benchmark for evaluating human-quality text to speech systems. In *International Conference on Learning Representations*, volume 2026, pp. 29228–29247, 2026.

Junwon Moon, Yejin Lee, Seungbeom Kim, Hoseong Ahn, Sewoong Park, Heeseung Kim, and Kyuhong Shim. Delta-tts: Adapting autoregressive model into diffusion language model for text-to-speech. *arXiv preprint arXiv:2607.04140*, 2026.

Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. VoxCeleb: A Large-Scale Speaker Identification Dataset. In *Interspeech 2017*, pp. 2616–2620, 2017. doi: 10.21437/Interspeech.2017-950.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Vardan Papyan, Xiao Y Han, and David L Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40): 24652–24663, 2020.

Ankita Pasad, Ju-Chieh Chou, and Karen Livescu. Layer-wise analysis of a self-supervised speech representation model. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 914–921. IEEE, 2021.

Ankita Pasad, Bowen Shi, and Karen Livescu. Comparative layer-wise analysis of self-supervised speech models. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Seungyoun Shin, Dongha Ahn, Jiwoo Kim, and Sungwook Jeon. No verifiable reward for prosody: Toward preference-guided prosody learning in tts. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2256–2260. IEEE, 2026.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017.

Ismail Rasim Ulgen, John HL Hansen, Carlos Busso, and Berrak Sisman. Rethinking speaker embeddings for speech generation: Sub-center modeling for capturing intra-speaker diversity. *arXiv preprint arXiv:2407.04291*, 2024.

Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016.

Haoxu Wang, Biao Tian, Weiqin Li, Xiang Lv, Han Zhao, and Xiangang Li. Flowtts-grpo: Online reinforcement learning with multi-objective reward optimization for flow-matching based text-to-speech. *arXiv preprint arXiv:2606.23190*, 2026.

Hui Wang, Siqi Zheng, Yafeng Chen, Luyao Cheng, and Qian Chen. CAM++: A Fast and Efficient Network for Speaker Verification Using Context-Aware Masking. In *Interspeech 2023*, pp. 5301–5305, 2023. doi: 10.21437/Interspeech.2023-1513.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International conference on machine learning*, pp. 9929–9939. PMLR, 2020.

Lloyd Welch. Lower bounds on the maximum cross correlation of signals (corresp.). *IEEE Transactions on Information theory*, 20(3):397–399, 1974.

Ivan Yakovlev, Rostislav Makarov, Andrei Balykin, Pavel Malov, Anton Okhotnikov, and Nikita Torgashov. Reshape Dimensions Network for Speaker Recognition. In *Interspeech 2024*, pp. 3235–3239, 2024. doi: 10.21437/Interspeech.2024-2116.

Zhen Ye, Xinfa Zhu, Chi-Min Chan, Xinsheng Wang, Xu Tan, Jiahe Lei, Yi Peng, Haohe Liu, Yizhu Jin, Zheqi Dai, et al. Llasa: Scaling train-time and inference-time compute for llama-based speech synthesis. *arXiv preprint arXiv:2502.04128*, 2025.

Zhao Yi, Wen-Chin Huang, Xiaohai Tian, Junichi Yamagishi, Rohan Kumar Das, Tomi Kinnunen, Zhen-Hua Ling, and Tomoki Toda. Voice Conversion Challenge 2020 — Intra-lingual semi-parallel and cross-lingual voice conversion —. In *Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*, pp. 80–98, 2020. doi: 10.21437/VCCBC.2020-14.

Chunlei Zhang and Dong Yu. C3-dino: Joint contrastive and non-contrastive self-supervised learning for speaker verification. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1273–1283, 2022.

A Related work and experimental details

A.1 Related work

Speaker embedders are now deployed well beyond verification, in tasks where they substitute for a listener's judgment of a voice rather than decide whether two recordings share a speaker. The cosine between synthesized and reference speech (SIM) is now the standard metric for comparing speaker similarity across synthesis systems (Anastassiou et al., 2024; Chen et al., 2025; Du et al., 2024), and in reinforcement fine-tuning it enters the reward alongside word error rate (Wang et al., 2026). In both roles the order of the scores determines the outcome, so these uses require a graded ranking rather than a binary decision.

In several system reports, however, the ranking by SIM and the ranking by listeners' SMOS disagree (Moon et al., 2026; Jia et al., 2025; Chen et al., 2025). Both the SIM value and the system ranking it produces change with the choice of embedder (Du et al., 2025). A study comparing acoustics, human perception, and DNN speaker verification systems finds that the systems diverge from listeners at the fine phonetic level (Liu et al., 2024), and work on speech generation finds that a single embedding per speaker discards the within-speaker variation that synthesis needs (Ulgen et al., 2024). Later work measures embedders directly against human perception with public labels (Ahn et al., 2024; Hu et al., 2022a). Yet human similarity labels are scarce and vary across raters, and whether a model fitted to them generalizes beyond that data has not been established. These studies established the gap and the means to measure it, but their remedy is limited to fine-tuning an embedder on human labels, without addressing the cause. How the training objective organizes the representation, what it erases, and whether the result suits the tasks it is deployed for remain unexplored.

Layer-wise probes of self-supervised speech models show that lower layers carry acoustic features while phonetic and lexical content emerges in higher layers, at a depth that depends on the pretraining objective (Pasad et al., 2021; 2023). Studies that probe speaker information specifically find it concentrated in the lower layers (Ashihara et al., 2024). Alignment with human similarity judgments likewise decreases in the deeper layers of most models (Kishi et al., 2026). The information relevant to perceived similarity is thus concentrated below the layers from which a verification head usually takes its input, and our backbone is truncated at that depth. Speaker representations are also learned on these models without labels, by self-distillation that matches a student to an exponential-moving-average teacher (Zhang & Yu, 2022; Han et al., 2023).

A.2 Training and evaluation datasets

Training uses only VoxCeleb2 dev (5,994 speakers) (Chung et al., 2018), and the only label taken from it is speaker identity (§3.2). VoxCeleb1, on which VoxSim is built, shares no speakers with VoxCeleb2, so every evaluation speaker is unseen in training. Only the post-training stage (DPO, §3.4) uses the human scores of the VoxSim train split. Evaluation covers three public test sets in which human similarity judgments exist as labels. VoxSim (Ahn et al., 2024) rates pairs of natural-speech utterances from VoxCeleb1 (Nagrani et al., 2017) on a 1–6 scale; its test split contains 1,898 same-speaker and 878 different-speaker pairs. VCC2018 and VCC2020 (Lorenzo-Trueba et al., 2018; Yi et al., 2020), the Voice Conversion Challenges, rate the outputs of voice conversion against reference utterances, with 5,600 and 6,080 utterance-pair groups, respectively. TTSDS2 (Minixhofer et al., 2026) provides human SMOS for the synthetic speech of twenty TTS systems; we use the 91 utterances of the clean domain, the only domain with public reference recordings.

A.3 Evaluation metrics

Following the official protocol of each data set, we compare the model score of an utterance pair against the mean human score of that pair. The Spearman rank correlation (SRCC) and the linear correlation (LCC) measure agreement independently of scale, and since training optimizes an ordering, SRCC serves as the primary metric. The mean squared error (MSE) and the accuracy (ACC, the fraction of predictions within 0.5 of the human score) measure absolute accuracy after a monotone mapping of the cosines onto the human rating scale, fitted on each data set's official train split and applied identically to every model, which leaves SRCC unchanged (VCC2020 reuses the VCC2018 mapping, following SVSNet; TTSDS2, whose scale differs, reports correlations only).

A.4 BASELINES

We compare against eight public speaker embedders trained without human similarity labels. They differ in architecture and objective but share verification training: ECAPA-TDNN (Desplanques et al., 2020), Fast ResNet-34 (AP) and ResNet-34 (Softmax+AP) (Chung et al., 2020), CAM++ (Wang et al., 2023), ERes2Net (Chen et al., 2023), ReDimNet-B3 (Yakovlev et al., 2024), and WavLM(Base)-ECAPA and WavLM(Large)-ECAPA (Chen et al., 2022), which place a verification head on a self-supervised backbone. For CAM++ and ERes2Net we use the VoxCeleb checkpoints of 3D-Speaker[2], and for the rest, the official release of the respective authors. As a reference upper bound rather than a baseline, we also list VoxSim-ft, the released WavLM(Large)-ECAPA fine-tuned on VoxSim human labels (Ahn et al., 2024)[3].

A.5 IMPLEMENTATION DETAILS

The two centroids of Eq. 3 are computed as

$$c_k = \frac{1}{|U_k|} \sum_{x \in U_k} f^{\mathrm{fr}}(x), \qquad \tilde{c}_k \leftarrow m\,\tilde{c}_k + (1-m)\,\bar{e}_k, \tag{7}$$

where U_k is the training utterances of speaker k, $\bar{e}_k$ is the mean student embedding of speaker k in the batch, and $m = 0.99$. c_k is computed once before training; $\tilde{c}_k$ is updated every batch, starting from $\bar{e}_k$. Both are L2-normalized and centred over speakers before use.

Table 6: Training settings. The upper block is Simpson and its preference refinement, the lower block the GRPO experiment of §5.2.

Component	**Setting**
Data	VoxCeleb2 dev, 5,994 speakers, speaker labels only; batch 256 $\times$ 2 utterances per speaker; 3-s crops
Backbone	WavLM-Large, CNN front end + lowest 3 transformer layers, frozen; LoRA $r = 32$, $\alpha = 64$ on attention and feed-forward projections
Readout	instance norm $\rightarrow$ ECAPA head (SE-Res2 $\times$3, attentive statistics pooling) $\rightarrow$ 256-d; per-layer mean/std side path (layers 0–3)
Loss weights	$\lambda_d = 3.0$, $\lambda_s = 1.5$; temperatures $\tau = 0.05$ for InfoNCE and $\beta = 0.1$ for ranking
Proxy similarities	different-speaker: frozen speaker centroids + student centroid bank ($m = 0.99$); same-speaker: frozen utterance features + EMA teacher (Eq. 3)
Proxy weights	α_{d} and α_{s}: 0 for the first 8k steps, then linear to 0.5 and 0.8, respectively, over 16k steps
Teacher	EMA of the student, λ 0.9995 $\rightarrow$ 1.0 (cosine); teacher 3-s / student random 1-s segments; released model: the teacher
Optimizer	AdamW, lr 10^{-3}, weight decay 0.01, step decay 0.95 per epoch, 10 epochs
Refinement (§3.4)	batch 64, lr 5×10^{-5}, 2 epochs, reference = the label-free model
GRPO policy	Llasa-1B, LoRA $r = 32$, $\alpha = 64$ on the attention projections
GRPO data	zero-shot prompts from LibriTTS, each a reference utterance with a target sentence from a different speaker; one epoch; 8 rollouts per prompt
GRPO reward	$z(1 - \mathrm{WER}) + z(\mathrm{SIM})$, each term standardized within the group; WER from Whisper-large-v3-turbo
GRPO optimizer	KL coefficient 0.04, lr 10^{-6}
GRPO evaluation	Seed-TTS test-en; WER by the official protocol (Whisper-large-v3), similarity by VoxSim-ft

Table 6 lists the settings omitted from §3 and §5.2; the two GRPO runs differ only in the embedder used for the SIM term, and the reward includes no length penalty and no termination term.

[2] https://github.com/modelscope/3D-Speaker

[3] https://github.com/kaistmm/voxsim_trainer

A.6 Subjective evaluation details

In the listening test reported in Table 5, 20 adult volunteers, recruited as described in the ethics statement, each evaluate 100 utterances per system. Speaker similarity is rated against the reference utterance on a five-point scale (SMOS), reported as a mean with a 95% confidence interval and compared between systems with a Wilcoxon signed-rank test (Simpson against WavLM(Large)-ECAPA, $p < 0.001$; against the base policy, $p < 0.001$). Preference is collected on a separate set of 100 utterances as an A/B choice between the two systems.

B Analysis

B.1 Layer-wise ranking signal

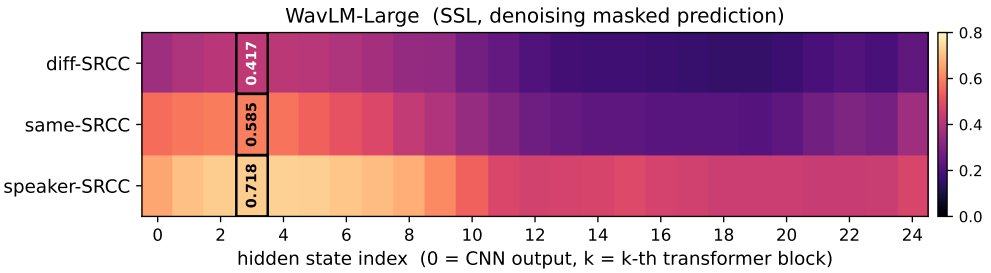


Figure 6: Correlation between human judgment and the mean-pooled cosine at each layer of a frozen WavLM-Large. The top two rows use different-speaker and same-speaker utterance pairs; the bottom row aggregates the same quantity at the speaker level, the signal our proxy similarities use. Boxes mark layer 3, where the backbone is truncated.

Figure 6 provides the evidence for the layer choice of §3.1: the ordering signal for different-speaker pairs increases up to the third layer and decreases with depth, so the backbone is truncated after the third layer.

B.2 Verification training over epochs

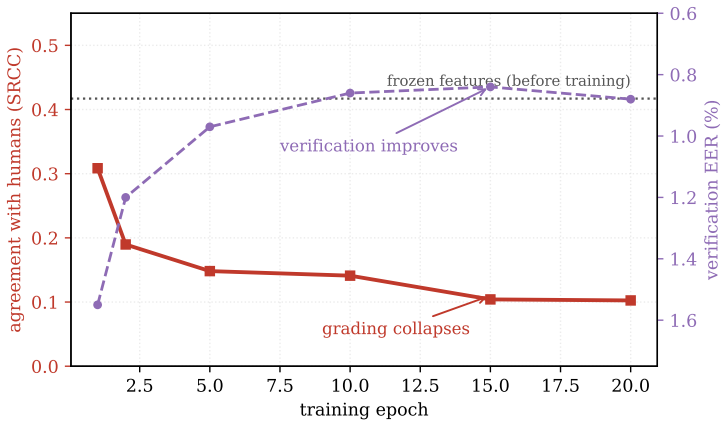


Figure 7: During one training run, the verification EER and the different-speaker ranking correlation move in opposite directions. We train a verification head (AAM) on frozen backbone features and measure both metrics at every epoch. The dotted line marks the ranking correlation of the frozen features before training.

Figure 7 shows the two metrics over the epochs of the verification-head training cited in §4.3: as the head trains, the ranking correlation on different-speaker pairs falls below its frozen value.

B.3 Duration and speaking rate

Table 7: Normalized similarity on LibriTTS test-clean under two controlled changes, shortening the audio and changing the speaking rate. Similarity is scaled so that 0 is an embedder's mean cosine to other speakers and 1 its mean cosine to another utterance of the same speaker. Neither direction is uniformly better, since a drop under time stretching is the intended behavior.

	WavLM(Large)-ECAPA	**Simpson**	*VoxSim-ft*
Duration (crop to N seconds)			
1 s	0.63	**0.85**	*0.66*
2 s	0.83	**0.94**	*0.83*
4 s	0.95	**0.98**	*0.95*
full	1.00	**1.00**	*1.00*
Delivery rate (time-stretch, pitch preserved)			
$0.8\times$	0.93	**0.76**	*0.75*
$0.9\times$	0.97	**0.77**	*0.76*
$1.1\times$	0.96	**0.83**	*0.79*
$1.25\times$	0.93	**0.79**	*0.75*

The probe cited in §5.2 alters one factor at a time on LibriTTS test-clean and measures each embedder's similarity. Raw cosines are not comparable across embedders, so the similarity is normalized: 0 is an embedder's mean cosine to other speakers and 1 its mean cosine to another utterance of the same speaker. *Crop* truncates an utterance to N seconds and compares it against another utterance of the same speaker; *Stretch* resamples it in time with the pitch preserved and compares it against the original.

In Table 7, from one second to the full utterance WavLM(Large)-ECAPA gains 0.37 while Simpson gains 0.15, and under rate changes between $0.8\times$ and $1.25\times$ its similarity is nearly unchanged while that of Simpson decreases. VoxSim-ft differs across the two axes: on duration it behaves like WavLM(Large)-ECAPA, the backbone it inherits, and on rate like Simpson. Its sensitivity to rate therefore appears to come from the human labels it was fine-tuned on rather than from the backbone.

B.4 Length and error rate of the GRPO outputs

Table 8: Outputs of the three policies on Seed-TTS test-en against the real recording of each sentence and the reference prompt.

Reward embedder	Off by >2 s (%) ↓	Unterminated (%) ↓	WER, terminated (%) ↓	Rate deviation, terminated ↓
Llasa-1B	7.0	1.84	3.32	0.195
WavLM(Large)-ECAPA	13.3	1.29	2.52	0.244
Simpson	**6.1**	**0.37**	**1.97**	**0.194**

Table 8 gives the numbers behind §5.2. Each policy synthesizes the 1,088 sentences of Seed-TTS test-en, and its outputs are compared with the real recording of the same sentence. *Off by >2 s* is the fraction of outputs more than two seconds longer or shorter than the recording, *unterminated* the fraction that reach the maximum length without an end token, *WER* the word error rate over the outputs that terminate, and *rate deviation* the mean of $|\log(r_{\text{out}}/r_{\text{ref}})|$ over the same outputs, where r is characters per second of the output and of the reference prompt of the same speaker. The verification reward raises the length deviation and the rate deviation above the base policy (paired Wilcoxon on the rate deviation, $p < 10^{-28}$), whereas the ranking reward lowers the length deviation and leaves the rate deviation unchanged; the gap in error rate remains once unterminated outputs are excluded.

These outcomes follow from what each reward scores. Verification-trained embedders assign higher similarity to longer audio and are insensitive to speaking rate (Appendix B.3), so a policy optimized against such a reward learns to lengthen its outputs, for example by slowing down, regardless of how the speech sounds to a listener. The similarity of Simpson gains little from duration and falls when

the speaking rate departs from the reference, so the policy trained with it keeps the rate of the base policy and gains nothing from longer outputs.

B.5 Exact numbers behind the speaker geometry

Table 9: Speaker-centroid cosine RMS against the bound $\sqrt{(N-d)/(d(N-1))}$ of Proposition 1. The random-directions row replaces the centroids with N vectors drawn uniformly on the sphere.

Speakers	Model	N	d	RMS	bound	ratio
VoxCeleb1 (all 1,251)	ECAPA-TDNN	1251	192	0.093	0.066	1.41
	WavLM(Large)-ECAPA	1251	256	0.104	0.056	1.87
	random directions	1251	256	0.063	0.056	1.12
	Simpson	1251	256	**0.629**	0.056	**11.28**
VoxCeleb2 dev (5,994)	ECAPA-TDNN	5994	192	0.083	0.071	1.16
	WavLM(Large)-ECAPA	5994	256	0.088	0.061	1.43
	Simpson	5994	256	**0.598**	0.061	**9.78**

Table 9 gives the numbers behind Figure 5. The centroid-cosine RMS of the verification-trained embedders is 1.16 to $1.87\times$ the bound of Proposition 1, a randomly drawn arrangement of directions reaches $1.12\times$, and Simpson reaches 9.8 to $11.3\times$. The random-direction row is the reference point for that comparison. With one vector fixed, the inner product of two vectors drawn uniformly on $\mathbb{S}^{d-1}$ has the distribution of a single coordinate of the other, so its mean is 0 and its variance is $1/d$. Its RMS of $1/\sqrt{d}$ therefore lies just above the bound.

B.6 The gender gap of trained embedders

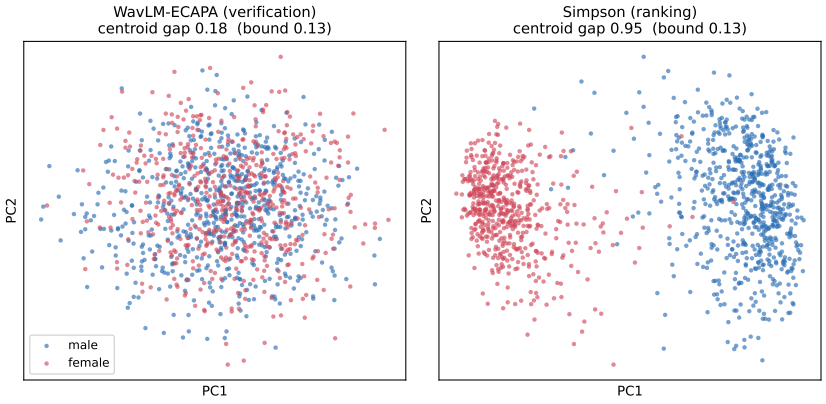


Figure 8: Speaker centroids of VoxCeleb1 on their first two principal components. In the verification-trained embedder (left) male and female speakers are mixed and the centroid gap (0.18; 0.15 for ECAPA-TDNN) lies at the Corollary 1 bound for the 561/690 gender partition of VoxCeleb1 (0.13 for $d = 256$, 0.15 for $d = 192$); in Simpson (right) gender is the first principal axis and the gap is 0.95. VoxSim-ft, the human-label embedder, keeps a comparable gap (0.90, over the 109 speakers of the VoxSim test split).

Figure 8 shows the speaker centroids behind the gap numbers of §4.3: in the verification-trained embedder male and female speakers are mixed, whereas in Simpson gender is the first axis along which they separate.

B.7 Realizing the human order when $N > d$

Proposition 2 shows that the sphere can encode the human order. Speaker points fitted to the human scores of the VoxSim training split reach, for $d \geq 16$, the maximum Spearman correlation permitted

by the ties in the scores. What verification-trained embedders lack is therefore not capacity but an objective that requires the order.

Table 10: Spearman correlation between the cosines of the constructed arrangement and the human scores, $N = 1{,}142$ speakers.

d	4	8	16	32	64	256
SRCC (human order)	0.878	0.941	0.948	0.948	0.948	0.948

Proposition 2 (Ranking is realizable)**.** *Given human similarity scores h over speaker pairs, the Spearman correlation between h and the cosines of any spherical arrangement with distinct cosines is at most $\rho^*(h) \leq 1$, a ceiling set by the ties in h (defined in Appendix B.8). For $d \geq N$ the ceiling is attained; for $N > d$ we construct arrangements that attain it (below).*

We take the $N = 1{,}142$ speakers of the VoxSim training split with human scores on $15{,}764$ speaker pairs. For each d we place N unit vectors on the sphere by minimizing the pairwise order loss of §3.3 against the human scores, then measure the Spearman correlation between the resulting cosines and h (Table 10). The tie ceiling $\rho^*(h)$ is 0.948, and the construction reaches it for $d \geq 16$. The same procedure on shuffled scores, which destroys the structure but keeps the ties, reaches 0.378 at $d = 16$.

This arrangement is an existence proof, not a model: it uses the human scores it is measured on.

B.8 Proofs

Proposition 1. For unit vectors $x_1, \ldots, x_N \in \mathbb{R}^d$ write $G = \sum_i x_i x_i^\top$, so that $\mathrm{tr}(G) = N$ and, by Cauchy–Schwarz, $\mathrm{tr}(G^2) \geq \mathrm{tr}(G)^2/d = N^2/d$. Since $\mathrm{tr}(G^2) = \sum_{i,j}\langle x_i, x_j\rangle^2 = N + \sum_{i\neq j}\langle x_i, x_j\rangle^2$, this gives $\frac{1}{N(N-1)}\sum_{i\neq j}\langle x_i, x_j\rangle^2 \geq \frac{N-d}{d(N-1)}$, equivalently the same bound for the mean over unordered pairs. Equality in the Cauchy–Schwarz step holds if and only if all eigenvalues of G are equal, that is, $G = (N/d)I$, which is the tight-frame condition; for $N \gg d$ the bound approaches $1/d$. □

Corollary 1. For a partition of the speakers into A and B with $N = N_A + N_B$, write $\Delta = \|\mu_A - \mu_B\|$ and $g = (\mu_A - \mu_B)/\Delta$. For a set attaining the bound, $\sum_i (x_i \cdot g)^2 = N/d$ for every unit vector g. Cauchy–Schwarz within each group gives $\sum_{i\in A}(x_i \cdot g)^2 \geq N_A a_A^2$ and $\sum_{i\in B}(x_i \cdot g)^2 \geq N_B a_B^2$, where $a_A = \mu_A \cdot g$, $a_B = \mu_B \cdot g$, and $a_A - a_B = \Delta$. The minimum of $N_A a_A^2 + N_B a_B^2$ under that constraint is $\frac{N_A N_B}{N}\Delta^2$, so $\frac{N_A N_B}{N}\Delta^2 \leq \frac{N}{d}$, that is, $\Delta \leq \frac{N}{\sqrt{N_A N_B}} \cdot \frac{1}{\sqrt{d}}$. □

Proposition 2. Spearman correlation is a correlation of ranks, so for cosines that are all distinct the rearrangement inequality shows that it is maximized when they are placed in the order of h, with the pairs tied in h placed in any order; every such placement gives the same value, which defines $\rho^*(h)$, and $\rho^*(h) = 1$ only when h has no ties. For $d \geq N$, take any symmetric S with unit diagonal whose off-diagonal entries are distinct and ordered as h: for small enough ϵ the matrix $(1-\epsilon)I + \epsilon S$ is positive semidefinite with unit diagonal, hence the Gram matrix of unit vectors, and it realizes the ceiling. For $N > d$ we fit a rank-d Gram matrix to h directly; the construction and its shuffled control are in §B.7. □

C Additional results

C.1 Target and loss ablation

Table 11 fixes the backbone, the data, and the batch, and varies the loss terms and, in the last row, the different-speaker proxy. Each ranking term raises the axis it targets and changes the other little. Relative to the same-speaker term alone, the full objective loses 0.005 on the same-speaker axis and gains 0.29 on the different-speaker axis, so both terms are kept. The last row replaces the speaker centroids of the different-speaker proxy with utterance-level features, the construction of

Table 11: Target and loss ablation. All rows share the backbone, data, and batch, and are reported at the same epoch. The last row keeps the full objective but replaces the speaker centroids in the different-speaker proxy with utterance-level features, the same construction as the same-speaker proxy.

	full		different speaker		same speaker	
Loss	SRCC ↑	LCC ↑	SRCC ↑	LCC ↑	SRCC ↑	LCC ↑
$\mathcal{L}_{\text{InfoNCE}}$	0.790	0.780	0.462	0.513	0.564	0.529
$\mathcal{L}_{\text{InfoNCE}} + \mathcal{L}_{\text{rank}}^{\text{diff}}$	0.796	0.764	0.773	0.713	0.538	0.499
$\mathcal{L}_{\text{InfoNCE}} + \mathcal{L}_{\text{rank}}^{\text{same}}$	0.813	0.804	0.496	0.544	**0.622**	0.584
$\mathcal{L}_{\text{InfoNCE}} + \mathcal{L}_{\text{rank}}^{\text{diff}} + \mathcal{L}_{\text{rank}}^{\text{same}}$	**0.822**	**0.827**	**0.785**	**0.753**	0.617	**0.593**
Full, utterance-level t^{diff}	0.694	0.691	0.380	0.484	0.405	0.354

t^{same} applied to different-speaker pairs. The different-speaker correlation falls from 0.785 to 0.380, below the InfoNCE-only model, and the same-speaker axis degrades as well, which confirms that different-speaker pairs require speaker-level proxies (§3.3).

C.2 Verification error

Table 12: Speaker verification EER (%) on VoxCeleb1-O with cosine scoring, for the checkpoints of Table 1.

Model	EER ↓
CAM++	0.73
ERes2Net	0.80
Fast ResNet-34 (AP)	2.28
ResNet-34 (Softmax+AP)	0.99
WavLM(Base)-ECAPA	1.05
WavLM(Large)-ECAPA	0.56
ReDimNet-B3	0.50
ECAPA-TDNN	0.90
Simpson	8.76
Simpson + DPO	9.22

Table 12 lists the verification EER of every embedder for reference; Simpson is not trained for verification.

C.3 Training variance across seeds

Table 13: SRCC of Simpson on VoxSim test as reported in Table 1, with the standard deviation over five training seeds.

	full	different speaker	same speaker
Simpson	0.822 ± 0.009	0.785 ± 0.005	0.617 ± 0.003

Simpson is trained five times with different random seeds to reduce the effect of random initialization, and Table 13 reports the standard deviation over these runs.

C.4 Decision accuracy

Table 14 gives the full results for the decision accuracy summarized in §4.2.

Table 14: Decision accuracy. Reversal is the rate of choosing the pair listeners rated less similar, over comparisons of two different-speaker pairs that listeners separated by at least one point. The two-alternative forced choice (2AFC) columns give the rate of choosing correctly, binned by the difference between the listeners' ratings of the two pairs, starting at half a point.

Model	Reversal (%) ↓	2AFC 0.5–1.0 ↑	1.0–1.5 ↑	1.5–2.0 ↑	>2.0 ↑
CAM++	40.0	49.7	55.0	56.8	67.6
ERes2Net	43.5	49.8	52.0	54.4	62.7
Fast ResNet-34 (AP)	24.1	58.7	68.8	74.5	84.7
ResNet-34 (Softmax+AP)	31.4	56.8	62.4	67.4	76.2
WavLM(Base)-ECAPA	44.9	50.5	51.5	51.9	61.1
WavLM(Large)-ECAPA	46.7	51.6	48.9	53.1	58.2
ReDimNet-B3	46.7	49.5	50.1	50.5	58.6
ECAPA-TDNN	50.2	48.6	44.5	49.6	55.9
Simpson	**8.1**	**75.9**	**88.3**	**92.2**	**95.7**
Simpson + DPO	**7.8**	**75.5**	**88.3**	**92.4**	**96.3**
VoxSim-ft (human labels)	*9.1*	*73.6*	*85.8*	*91.1*	*96.4*

C.5 Confidence intervals

Table 15: Bootstrap 95% confidence intervals for SRCC on VoxSim test. The utterances are resampled 5,000 times with replacement, the same draws shared by every model.

Model	full ↑	different speaker ↑	same speaker ↑
CAM++	$+0.756$ [$+0.739$, $+0.773$]	$+0.155$ [$+0.083$, $+0.223$]	$+0.513$ [$+0.477$, $+0.547$]
ERes2Net	$+0.745$ [$+0.727$, $+0.762$]	$+0.102$ [$+0.033$, $+0.170$]	$+0.491$ [$+0.454$, $+0.526$]
Fast ResNet-34 (AP)	$+0.774$ [$+0.757$, $+0.791$]	$+0.452$ [$+0.393$, $+0.506$]	$+0.521$ [$+0.487$, $+0.554$]
ResNet-34 (Softmax+AP)	$+0.769$ [$+0.750$, $+0.785$]	$+0.333$ [$+0.268$, $+0.397$]	$+0.522$ [$+0.488$, $+0.554$]
WavLM(Base)-ECAPA	$+0.753$ [$+0.735$, $+0.770$]	$+0.082$ [$+0.009$, $+0.152$]	$+0.516$ [$+0.482$, $+0.549$]
WavLM(Large)-ECAPA	$+0.736$ [$+0.718$, $+0.754$]	$+0.061$ [-0.012, $+0.128$]	$+0.472$ [$+0.437$, $+0.506$]
ReDimNet-B3	$+0.741$ [$+0.724$, $+0.759$]	$+0.049$ [-0.022, $+0.116$]	$+0.491$ [$+0.454$, $+0.525$]
ECAPA-TDNN	$+0.751$ [$+0.734$, $+0.768$]	-0.010 [-0.083, $+0.060$]	$+0.527$ [$+0.493$, $+0.558$]
Simpson	$+0.822$ [$+0.806$, $+0.837$]	$+0.785$ [$+0.756$, $+0.808$]	$+0.617$ [$+0.587$, $+0.646$]
Simpson + DPO	$+0.833$ [$+0.818$, $+0.847$]	$+0.785$ [$+0.755$, $+0.809$]	$+0.642$ [$+0.614$, $+0.670$]
VoxSim-ft (human labels)	$+0.837$ [$+0.822$, $+0.851$]	$+0.758$ [$+0.724$, $+0.785$]	$+0.648$ [$+0.618$, $+0.675$]

Table 16: Bootstrap 95% confidence intervals for the TTSDS2 correlations of Table 3. Utterance-level intervals resample the 91 utterances and system-level intervals the twenty systems.

Model	Utterance level SRCC ↑	System level SRCC ↑
WavLM(Large)-ECAPA	0.320 [$+0.10$, $+0.51$]	0.420 [-0.11, $+0.80$]
ECAPA-TDNN	0.308 [$+0.09$, $+0.49$]	0.525 [$+0.03$, $+0.81$]
Simpson	**0.415** [$+0.23$, $+0.57$]	**0.583** [$+0.14$, $+0.84$]
Simpson + DPO	**0.452** [$+0.28$, $+0.59$]	**0.720** [$+0.36$, $+0.87$]
VoxSim-ft (human labels)	*0.406* [$+0.22$, $+0.57$]	*0.612* [$+0.12$, $+0.85$]

Table 15 gives bootstrap 95% confidence intervals for the SRCC values of Table 1. The utterances are resampled, keeping a pair when both of its utterances are drawn, with the same draws shared by every model. Table 16 gives the same intervals for TTSDS2. The twenty systems are resampled with replacement for the system-level intervals and the 91 utterances for the utterance-level ones, 5,000 draws each, with the same draws shared by every model.

Table 17: Different-speaker SRCC on VoxSim test: all pairs and same-gender pairs only.

Model	all (878)	same gender (431)
WavLM(Large)-ECAPA	0.061	0.115
ECAPA-TDNN	−0.010	0.114
Simpson	**0.785**	**0.491**
Simpson + DPO	**0.785**	**0.525**
VoxSim-ft (human labels)	*0.758*	*0.501*

C.6 Different-speaker ranking split by gender

To check how much of the different-speaker correlation comes from separating genders, Table 17 restricts the 878 different-speaker pairs of VoxSim test to the 431 same-gender pairs and recomputes the SRCC. On same-gender pairs gender is uninformative, and Simpson and the human-label embedder retain four times the correlation of the verification-trained embedders.

C.7 Position of listener-rated similar pairs in each embedder's score distribution

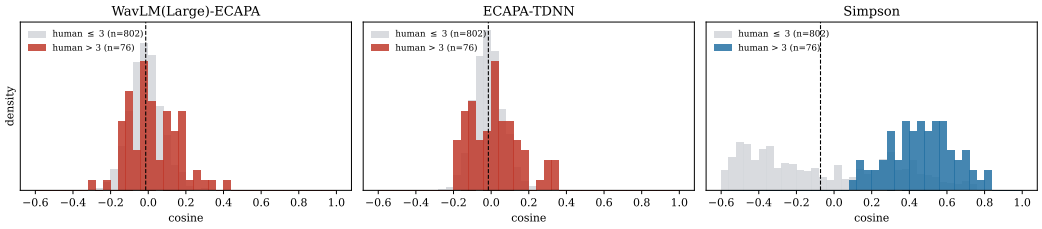


Figure 9: VoxSim test, 878 different-speaker pairs: cosine distribution of the pairs listeners rated above 3 (colored) and at most 3 (grey), per embedder. The dashed line is the embedder's median over all 878 pairs.

To show that the pair of Figure 1 is not an exception, Figure 9 splits the 878 different-speaker pairs of VoxSim test into those listeners rated above 3 and at most 3 and plots the cosine distribution of each group for each embedder. For the verification-trained embedders the two distributions overlap, and 38–42% of the pairs listeners found similar fall below the embedder's own median; for Simpson the two distributions separate and none of those pairs falls below the median.

C.8 Voice conversion by subset

Table 18: VCC2018 by subset. *toS* compares a conversion with its source speaker, *toT* with its target speaker, and *nat* is a reference pair of natural utterances.

	toS (2,685)				toT (2,716)				nat (199)			
Model	SRCC ↑	LCC ↑	MSE ↓	ACC ↑	SRCC ↑	LCC ↑	MSE ↓	ACC ↑	SRCC ↑	LCC ↑	MSE ↓	ACC ↑
CAM++	0.315	0.409	1.003	24.6	0.339	0.350	0.909	34.5	0.791	0.909	0.395	63.3
ERes2Net	0.354	0.427	0.979	24.6	0.352	0.361	0.892	36.3	0.790	0.909	0.391	63.3
Fast ResNet-34 (AP)	0.375	0.449	0.952	24.9	0.342	0.361	0.880	36.1	0.792	0.918	0.353	62.8
ResNet-34 (Softmax+AP)	0.353	0.429	0.987	24.9	0.335	0.345	0.919	34.8	0.792	0.911	0.364	62.8
WavLM(Base)-ECAPA	0.349	0.429	1.029	25.4	0.332	0.336	0.962	33.2	0.773	0.906	0.398	63.3
WavLM(Large)-ECAPA	0.311	0.406	1.049	24.5	0.297	0.298	0.990	33.7	0.775	0.908	0.422	62.8
ReDimNet-B3	0.313	0.411	1.000	24.9	0.355	0.366	0.892	36.3	0.781	0.900	0.420	63.3
ECAPA-TDNN	0.354	0.428	0.979	24.0	0.338	0.346	0.903	35.1	0.751	0.904	0.384	62.8
Simpson	**0.467**	**0.507**	**0.890**	**43.0**	**0.365**	**0.393**	**0.851**	**36.9**	**0.831**	**0.922**	**0.362**	**81.4**
Simpson + DPO	**0.477**	**0.533**	**0.864**	**45.3**	**0.403**	**0.431**	**0.821**	**37.3**	**0.805**	**0.932**	**0.305**	**83.4**
VoxSim-ft (human labels)	*0.467*	*0.525*	*0.866*	*40.9*	*0.385*	*0.412*	*0.834*	*36.7*	*0.819*	*0.935*	*0.311*	*78.4*

Table 19: VCC2020 by subset. *intra* is English-to-English conversion and *cross* is cross-lingual conversion, where the target speaker is Finnish, German, or Mandarin.

	intra-lingual				cross-lingual			
Model	SRCC ↑	LCC ↑	MSE ↓	ACC ↑	SRCC ↑	LCC ↑	MSE ↓	ACC ↑
CAM++	0.738	0.767	0.252	65.1	0.548	0.574	0.303	60.6
ERes2Net	0.716	0.759	0.245	68.7	0.511	0.556	0.319	59.7
Fast ResNet-34 (AP)	0.674	0.736	0.325	57.0	0.452	0.503	0.335	59.3
ResNet-34 (Softmax+AP)	0.701	0.750	0.267	65.6	0.566	0.594	0.290	63.2
WavLM(Base)-ECAPA	0.727	0.749	0.274	63.9	0.487	0.511	0.335	59.7
WavLM(Large)-ECAPA	0.741	0.752	0.275	63.1	0.487	0.499	0.356	57.6
ReDimNet-B3	0.720	0.765	0.239	67.6	0.522	0.545	0.325	59.8
ECAPA-TDNN	0.710	0.765	0.254	65.0	0.523	0.567	0.307	61.9
Simpson	**0.771**	**0.841**	**0.189**	**77.9**	**0.582**	**0.630**	**0.435**	**52.9**
Simpson + DPO	**0.758**	**0.837**	**0.187**	**78.0**	**0.614**	**0.661**	**0.341**	**59.4**
VoxSim-ft (human labels)	*0.711*	*0.804*	*0.214*	*70.2*	*0.527*	*0.577*	*0.335*	*60.3*

The main text reports the official aggregate of each data set. Broken down by the role of the reference speaker and by the language condition (Tables 18 and 19), the conclusion on rank correlation is the same in every subset.

C.9 Absolute prediction on VoxSim

Table 20: Absolute prediction on VoxSim, split by axis.

	different speaker		same speaker	
Model	MSE ↓	ACC ↑	MSE ↓	ACC ↑
CAM++	0.816	31.8	1.502	24.9
ERes2Net	0.822	31.1	1.534	24.0
Fast ResNet-34 (AP)	0.666	52.5	1.509	22.4
ResNet-34 (Softmax+AP)	0.771	46.5	1.505	23.5
WavLM(Base)-ECAPA	0.848	31.2	1.504	24.7
WavLM(Large)-ECAPA	0.866	31.3	1.584	23.0
ReDimNet-B3	0.863	31.4	1.536	24.7
ECAPA-TDNN	0.869	31.0	1.431	25.7
Simpson	**0.398**	**72.0**	**1.276**	**34.8**
Simpson + DPO	**0.383**	**70.7**	**1.172**	**37.6**
VoxSim-ft (human labels)	*0.394*	*69.6*	*1.154*	*36.2*

Table 20 splits the absolute-prediction columns of Table 1 by axis.

C.10 Scores on the outputs of one TTS system

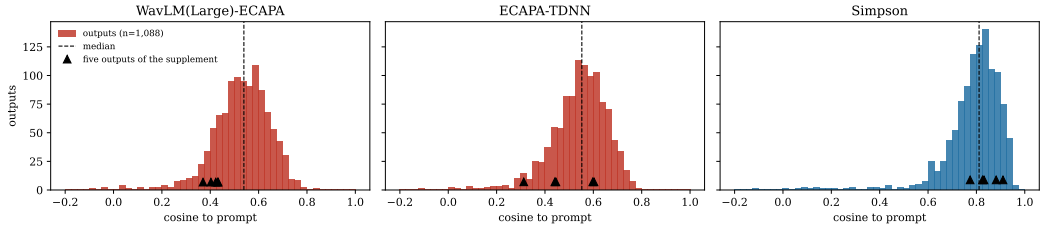


Figure 10: Seed-TTS test-en, 1,088 outputs of Llasa-1B: cosine between each output and its prompt, per embedder. Triangles mark the five outputs of the samples page; the dashed line is the embedder's median.

Figure 10 shows, for the 1,088 Seed-TTS test-en outputs of Llasa-1B, the distribution of the cosine each embedder assigns between an output and its prompt, and marks the five outputs of the samples page, which are transcribed without error and match the prompt speaker. They lie in the lower tail of WavLM(Large)-ECAPA and around the median of Simpson.

D FUTURE WORK

Standard benchmarks for speaker representation consist of discriminative tasks such as verification, diarization, and identification, and none of them measures the graded judgment that embeddings are deployed to make. This explains why the gap reported here has gone unnoticed, and a benchmark for speaker representation that includes ranking is needed. Ranking losses compare pairs within a batch, so the batch size and the diversity of voices in a corpus limit the ordering a model can learn. How far the remaining gap to listeners closes with larger batches and with open corpora beyond VoxCeleb2 is left for future work.