TIDE: TOPOLOGICAL IDENTITY DYNAMICS AND EVIDENCE LEARNING FOR SHORT-UTTERANCE SPEAKER VERIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Short-utterance speaker verification is challenging because only limited speaker information is observed. Short speech has poor phonetic coverage and misses many identity related transitions. Its representation is therefore sparse, fragmented, and sensitive to nuisance factors. Most existing systems still compress the available frames into a fixed dimensional embedding, without explicitly modeling missing identity structure or evidence reliability. We propose TIDE, which views a short utterance as an incomplete observation of a speaker trajectory. Frame-level features and their local temporal differences form position velocity state clouds. Multi-scale persistent homology captures stable identity structures across temporal scales. During training, a momentum long-view branch provides a richer topological target. A duration-gated residual module estimates the recoverable structural correction from the short view without reconstructing speech. The completed topology then modulates the vector field of a neural controlled differential equation, enabling continuous identity evolution along the observed acoustic path. A thinning-consistency loss improves robustness to missing observations. For verification, dynamic states contribute time-weighted directional evidence to a von Mises Fisher posterior. Its mean represents speaker identity and its concentration reflects evidence coherence. Experiments on VoxCeleb show consistent gains over strong speaker-verification baselines, with the largest benefits under limited speech.

1 INTRODUCTION

Speaker identity is not sampled uniformly over time. A one-second segment may capture an informative vocal-tract transition, but it may also be dominated by a narrow phonetic event, noise, or channel coloration. Standard speaker-verification systems hide this imbalance by compressing an utterance into one fixed-dimensional embedding. Strong encoders have improved this paradigm, including x-vector (Snyder et al., 2018), ECAPA-TDNN (Desplanques et al., 2020), RawNet3 (Jung et al., 2022), MFA-Conformer (Zhang et al., 2022), ECAPA2 (Thienpondt & Demuynck, 2023), CAM++ (Wang, Zheng, et al., 2023), and ERes2Net (Chen et al., 2023). Self-supervised learning adds transferable representations through wav2vec 2.0 (Baevski et al., 2020), data2vec (Baevski et al., 2022), HuBERT (Hsu et al., 2021), UniSpeech-SAT (Chen, Wu, et al., 2022), WavLM (Chen, Wang, et al., 2022), and large-scale SSL adaptation for speaker verification (Chen, Qian, et al., 2022). Short-utterance verification nevertheless remains difficult because limited speech reduces phonetic coverage, truncates local transitions, and magnifies nuisance factors. ERes2NetV2 (Chen et al., 2024), MR-RawNet (Kim, Lim, et al., 2024), adaptive large-margin fine-tuning (Zhang et al., 2023), and DAME (Jung et al., 2026) reduce this mismatch. They still focus mainly on estimating a better vector from fewer frames.

We take a different view. Short speech is not simply a lower-quality version of long speech. It is a partial observation of a structured speaker trajectory. Missing observations remove both speaker states and transitions between them. Test-time speech extension can add evidence (Zhao et al., 2025), but synthesized duration is not equivalent to naturally observed context. Uncertainty-aware methods provide another correction. Xi-vector propagates frame-level precision (Lee et al., 2021), uncertainty can enter scoring directly (Wang, Lee, et al., 2023), empirical studies separate duration- and noise-related uncertainty (Jing et al., 2025), and Xi+ supervises uncertainty explicitly (Li et al.,

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.

2026). These approaches quantify confidence, but do not identify which structural relations were undersampled or how the remaining trajectory fragments should be integrated.

TIDE couples structural support, temporal evolution, and verification evidence. As summarized in Figure 1, speaker-oriented frame representations first form position–velocity state clouds. Multi-scale persistent homology retains structures that survive changes in neighborhood scale. A momentum long-view teacher then supervises a duration-gated structural completion module. The completed topology conditions the vector field of a neural controlled differential equation (CDE) (Kidger et al., 2020); topology therefore changes how the observed acoustic path is integrated rather than acting as a post-hoc descriptor. Finally, the dynamic states generate directional evidence that is accumulated in physical time into a von Mises–Fisher (vMF) posterior. The mean direction acts as the speaker representation, while concentration grows only when the accumulated evidence is directionally coherent. We use this concentration as a task-coupled measure of coherent speaker evidence rather than a universal Bayesian epistemic uncertainty, consistent with recent cautions on generic evidential outputs (Shen et al., 2024).

Contributions. We make three contributions. First, we formulate short-utterance verification as identity inference from a partially observed latent trajectory and introduce multi-scale persistent identity topology with duration-gated long-view completion. Second, we place the completed topology inside a neural CDE vector field and regularize the resulting trajectory under missing observations. Structural context therefore governs continuous identity dynamics instead of being fused after temporal aggregation. Third, we derive physical-time spherical evidence accumulation and a vMF verification geometry that couple speaker direction, evidence concentration, and uncertainty-aware scoring in the same representation.

A broader discussion of speaker representation learning, topological sequence modeling, and uncertainty-aware verification is provided in Appendix A.

2 METHOD

2.1 PROBLEM FORMULATION AND OVERVIEW

Let $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_T)$ be an utterance and let y be its training speaker label. A speaker encoder F_θ produces frame-level states

$$\mathbf{H} = F_\theta(\mathbf{x}) = [\mathbf{h}_1, \ldots, \mathbf{h}_T]^\top \in \mathbb{R}^{T \times d_h}. \tag{1}$$

Conventional systems pool $\mathbf{H}$ into one vector. This assumes that the observed segment adequately samples speaker variation. For short speech, speaker-informative states may be absent, local transitions may be truncated, and nuisance factors may dominate the remaining frames. We instead treat the utterance as a partial observation of an underlying identity trajectory.

During training, we sample a long view $\mathbf{x}^\ell$ and a nested short view $\mathbf{x}^s \subset \mathbf{x}^\ell$, with durations $D_\ell > D_s$. The nesting constraint ensures that the long view contains the complete short observation and only adds surrounding acoustic context. The long branch is used for training supervision only. At inference, the model requires only the observed enrollment and test utterances.

Figure 1 shows the full inference chain: persistent identity structure is first extracted from the short view; a momentum target supervises structural completion; the completed topology conditions continuous identity dynamics; and the resulting states are integrated as spherical speaker evidence.

2.2 MULTI-SCALE PERSISTENT IDENTITY TOPOLOGY

Dynamics-aware state construction. Persistent homology should operate on a speaker-oriented geometry rather than raw acoustic frames. For each encoder state, we first compute

$$\mathbf{r}_i = \frac{\mathbf{W}_r \mathbf{h}_i}{\|\mathbf{W}_r \mathbf{h}_i\|_2 + \epsilon}, \tag{2}$$

where $\mathbf{W}_r$ is learned jointly with the encoder. Let τ_i be the physical timestamp of frame i. With finite-difference stride Δ, local motion is

$$\mathbf{v}_i = \frac{\mathbf{r}_{i+\Delta} - \mathbf{r}_{i-\Delta}}{\tau_{i+\Delta} - \tau_{i-\Delta}}, \tag{3}$$

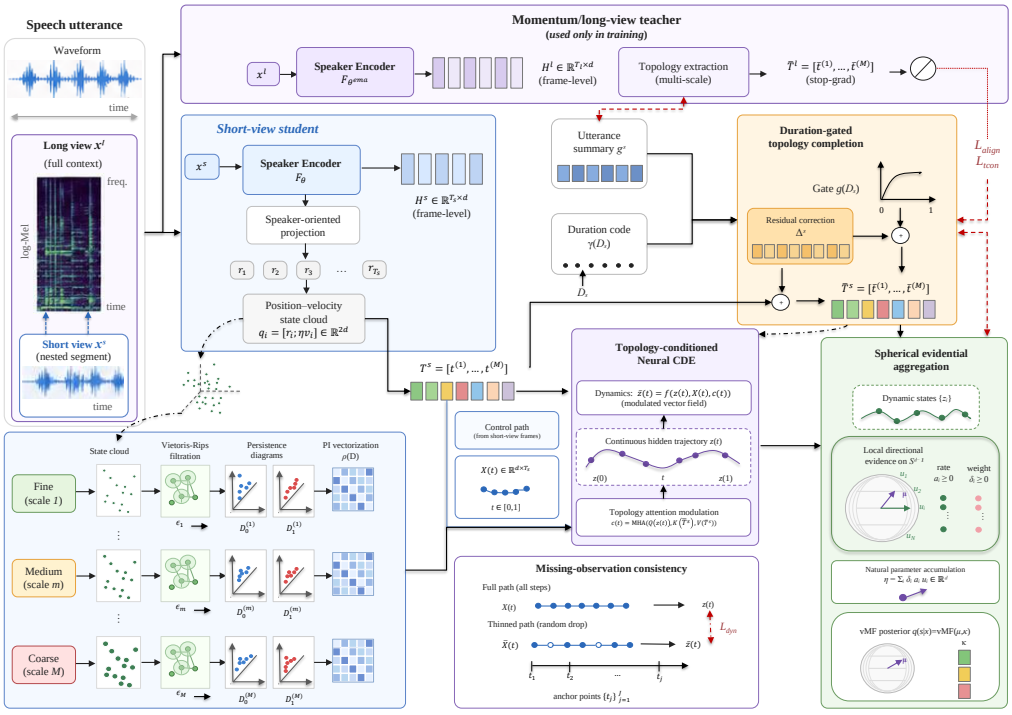

Figure 1: Overview of TIDE for speaker verification from partial acoustic observations. The short-view branch constructs multi-scale persistent identity topology from position velocity state clouds. A training-only momentum long-view teacher provides structural supervision for duration-gated topology completion. The completed topology conditions the neural CDE vector field to model continuous identity dynamics along the observed acoustic path. Missing-observation consistency regularizes the resulting trajectory, and spherical evidential aggregation converts the dynamic states into a vMF posterior for verification. The long-view branch is removed at inference.

with one-sided differences at sequence boundaries. We normalize its magnitude within the utterance,

$$\widetilde{\mathbf{v}}_i = \frac{\mathbf{v}_i}{\text{median}_j \, \|\mathbf{v}_j\|_2 + \epsilon}, \qquad \mathbf{q}_i = [\mathbf{r}_i; \alpha_v \widetilde{\mathbf{v}}_i], \tag{4}$$

where α_v controls the contribution of local dynamics.

A single temporal resolution is unstable when only a few acoustic events are observed. We therefore form M state clouds with adaptive temporal pooling at progressively larger supports. Each cloud contains at most N_τ landmarks. At scale m, pairwise distances are robustly normalized by

$$\widetilde{d}_{ij}^{(m)} = \frac{\|\mathbf{q}_i^{(m)} - \mathbf{q}_j^{(m)}\|_2}{\text{median}_{a<b} \, \|\mathbf{q}_a^{(m)} - \mathbf{q}_b^{(m)}\|_2 + \epsilon}. \tag{5}$$

This removes global feature-scale changes while retaining the relative neighborhood geometry used by the filtration.

Persistent topology tokenization. For each scale, the distances in Equation equation 5 define a Vietoris–Rips filtration. We retain the zero- and one-dimensional persistence diagrams $\mathcal{D}_0^{(m)}$ and $\mathcal{D}_1^{(m)}$. For a persistence pair $p = (b, d)$, let $\ell = d - b$. A learnable persistence-image component is

$$\text{PI}_j(\mathcal{D}) = \sum_{p \in \mathcal{D}} w(\ell_p) \exp\left(-\frac{\|[b_p, \ell_p]^\top - \mathbf{c}_j\|_2^2}{2\sigma_j^2} \right), \tag{6}$$

where $\mathbf{c}_j$ and σ_j are learnable basis parameters and $w(\ell)$ is the smooth lifetime weighting used to suppress near-diagonal pairs. The two homology orders are fused as

$$\mathbf{t}^{(m)} = \mathbf{W}_t[\mathrm{PI}(\mathcal{D}_0^{(m)}); \mathrm{PI}(\mathcal{D}_1^{(m)})] + \mathbf{b}_t, \qquad \mathbf{T}^s = [\mathbf{t}^{(1)}, \ldots, \mathbf{t}^{(M)}]. \tag{7}$$

We preserve one token per temporal scale. This allows the evolving identity state to query fine and coarse structure separately.

The persistence operator is implemented with a differentiable topology layer. Between filtration-order changes, birth and death values are differentiable with respect to the selected pairwise distances. Gradients pass through these values and the persistence-image mapping, while the current simplex pairing is treated as locally fixed (Brüel-Gabrielsson et al., 2020; Carrière et al., 2021).

2.3 Duration-gated momentum topology completion

Topology extracted from a very short segment is itself incomplete. Directly conditioning the dynamics on $\mathbf{T}^s$ would preserve this fragmentation. We therefore learn a residual short-to-long structural completion map.

Let ϑ denote the trainable parameters of the topology branch and $\bar{\vartheta}$ their momentum copy. The teacher update is

$$\bar{\vartheta} \leftarrow \zeta\bar{\vartheta} + (1-\zeta)\vartheta, \qquad 0 \le \zeta < 1. \tag{8}$$

The stopped long-view target is

$$\bar{\mathbf{T}}^\ell = \mathrm{stopgrad}\left(\Phi_{\bar{\vartheta}}(\mathbf{x}^\ell)\right), \qquad \mathbf{T}^s = \Phi_\vartheta(\mathbf{x}^s), \tag{9}$$

where Φ denotes the complete state-cloud and persistent-topology pipeline.

From the short-view features we also obtain an utterance summary $\mathbf{g}^s$ and a continuous duration code $\gamma(D_s/D_{\mathrm{ref}})$, where D_{ref} fixes the duration scale. A cross-scale completion network predicts

$$\Delta\mathbf{T}^s = \mathcal{C}_\psi\left(\mathbf{T}^s, \mathbf{g}^s, \gamma(D_s/D_{\mathrm{ref}})\right). \tag{10}$$

We use a monotone gate $g_D(D_s) \in (0,1]$, constrained such that $\partial g_D/\partial D_s \le 0$, and form

$$\widehat{\mathbf{T}}^s = \mathbf{T}^s + g_D(D_s)\Delta\mathbf{T}^s. \tag{11}$$

The residual form retains reliable observed structure. The decreasing gate limits correction as more acoustic context becomes available.

Completion is supervised at two levels. Scale-wise structural alignment is

$$\mathcal{L}_{\mathrm{align}} = \frac{1}{M}\sum_{m=1}^{M}\left\|\widehat{\mathbf{t}}_s^{(m)} - \bar{\mathbf{t}}_\ell^{(m)}\right\|_2^2. \tag{12}$$

We also use cross-view supervised contrastive learning. Let

$$\mathbf{e}_i^s = \mathrm{norm}\left(\mathbf{P}\,\mathrm{vec}(\widehat{\mathbf{T}}_i^s)\right), \qquad \mathbf{e}_i^\ell = \mathrm{norm}\left(\mathbf{P}\,\mathrm{vec}(\bar{\mathbf{T}}_i^\ell)\right),$$

and let $\mathcal{P}(i)$ contain the long-view samples in the mini-batch sharing speaker label y_i. Then

$$\mathcal{L}_{\mathrm{tcon}} = -\frac{1}{B}\sum_{i=1}^{B}\frac{1}{|\mathcal{P}(i)|}\sum_{p\in\mathcal{P}(i)}\log\frac{\exp(\mathbf{e}_i^{s\top}\mathbf{e}_p^\ell/\tau_c)}{\sum_{a=1}^{B}\exp(\mathbf{e}_i^{s\top}\mathbf{e}_a^\ell/\tau_c)}. \tag{13}$$

The topology objective is

$$\mathcal{L}_{\mathrm{top}} = \mathcal{L}_{\mathrm{align}} + \lambda_{\mathrm{tcon}}\mathcal{L}_{\mathrm{tcon}}. \tag{14}$$

Alignment recovers same-instance structure; the contrastive term keeps that structure speaker-discriminative.

2.4 TOPOLOGY-CONDITIONED CONTINUOUS IDENTITY DYNAMICS

Persistent topology provides structural context but does not determine how each observed acoustic increment should update the identity estimate. We therefore model identity formation as a continuous hidden trajectory driven by the short speech itself. Neural CDEs provide this formulation because hidden states evolve with increments of an observed path rather than an autonomous clock (Kidger et al., 2020; Morrill et al., 2021).

Let $\widetilde{\tau}_i = (\tau_i - \tau_1)/(\tau_{T_s} - \tau_1) \in [0, 1]$. The short-view control knots are

$$\mathbf{X}(\widetilde{\tau}_i) = \mathbf{W}_x \mathbf{h}_i^s + \mathbf{b}_x, \tag{15}$$

and $\mathbf{X}(t)$ is their piecewise-linear interpolation. Absolute duration remains available through $\gamma(D_s/D_{\mathrm{ref}})$.

We initialize from the first observed control value only,

$$\mathbf{z}(0) = \phi_0(\mathbf{X}(0)). \tag{16}$$

This prevents a globally pooled utterance vector from bypassing path integration. At time t, the hidden state queries the completed topology,

$$\mathbf{c}(t) = \mathrm{MHA}\left(\mathbf{W}_q \mathbf{z}(t), \mathbf{W}_k \widehat{\mathbf{T}}^s, \mathbf{W}_v \widehat{\mathbf{T}}^s\right). \tag{17}$$

The trajectory evolves as

$$\mathbf{z}(t) = \mathbf{z}(0) + \int_0^t f_\phi\left(\mathbf{z}(u), \mathbf{c}(u), \gamma(D_s/D_{\mathrm{ref}})\right) d\mathbf{X}(u). \tag{18}$$

Topology therefore acts inside the vector field. Because $\mathbf{c}(t)$ depends on the current hidden state, different temporal scales can contribute at different points of the trajectory. The long-view teacher never supplies acoustic values to $\mathbf{X}(t)$.

Missing-observation consistency. For each short view, we randomly remove a subset of non-boundary control knots while retaining their timestamps. Recomputing topology completion and the CDE on this thinned path gives $\widetilde{\mathbf{z}}(t)$. The original short-view trajectory is stopped and used as the reference. At anchor times $\{a_j\}_{j=1}^J$,

$$\mathcal{L}_{\mathrm{dyn}} = \frac{1}{J} \sum_{j=1}^{J} \|\widetilde{\mathbf{z}}(a_j) - \mathrm{stopgrad}(\mathbf{z}(a_j))\|_2^2. \tag{19}$$

The denser short path is thus the target and the sparser path is required to preserve its identity trajectory.

2.5 SPHERICAL EVIDENTIAL IDENTITY AGGREGATION

The CDE produces a sequence of identity states rather than one deterministic embedding. TIDE represents the utterance by a vMF posterior on the same unit hypersphere used for speaker comparison.

At observed time τ_i, a dynamic state produces a unit evidence direction

$$\mathbf{u}_i = \frac{\mathbf{W}_u \mathbf{z}_i}{\|\mathbf{W}_u \mathbf{z}_i\|_2 + \epsilon} \in \mathbb{S}^{d-1}, \tag{20}$$

and a bounded non-negative evidence rate

$$a_i = a_{\max} \sigma\left(\mathbf{w}_a^\top \mathbf{z}_i + b_a\right). \tag{21}$$

The bound prevents arbitrary concentration inflation by rescaling the evidence head.

Evidence is accumulated in physical time. Let δ_i be trapezoidal quadrature weights computed from the physical timestamps. Under a uniform prior over latent speaker direction $\mathbf{s} \in \mathbb{S}^{d-1}$, each local contribution has the exponential-family form

$$\varphi_i(\mathbf{s}) \propto \exp\left(\delta_i a_i \mathbf{u}_i^\top \mathbf{s}\right). \tag{22}$$

The accumulated natural parameter is

$$\boldsymbol{\eta} = \sum_i \delta_i a_i \mathbf{u}_i, \qquad \kappa = \|\boldsymbol{\eta}\|_2, \qquad \boldsymbol{\mu} = \frac{\boldsymbol{\eta}}{\kappa + \epsilon}. \tag{23}$$

The utterance posterior is

$$q(\mathbf{s} \mid \mathbf{x}) = C_d(\kappa) \exp\left(\kappa \boldsymbol{\mu}^\top \mathbf{s}\right), \tag{24}$$

where $C_d(\kappa)$ is the d-dimensional vMF normalizer.

We distinguish available evidence from its directional coherence:

$$E = \sum_i \delta_i a_i, \qquad r_{\text{coh}} = \frac{\kappa}{E + \epsilon}. \tag{25}$$

By the triangle inequality, $0 \le r_{\text{coh}} \le 1$. More observed speech can increase E, but κ rises only when new evidence agrees with the accumulated speaker direction.

Uncertainty-aware verification. Let enrollment and test utterances produce natural parameters $\boldsymbol{\eta}_e, \boldsymbol{\eta}_t$ and concentrations κ_e, κ_t. We compare the complete vMF posteriors with the normalized mutual-likelihood score

$$s_{\text{TIDE}}(e,t) = \frac{1}{2}\log C_d(2\kappa_e) + \frac{1}{2}\log C_d(2\kappa_t) - \log C_d\left(\|\boldsymbol{\eta}_e + \boldsymbol{\eta}_t\|_2\right). \tag{26}$$

The score is the logarithm of the L_2-normalized posterior overlap. It therefore depends on both direction and evidence concentration. All normalizers are evaluated in log space using stable modified-Bessel routines.

2.6 End-to-end verification

We train direction, concentration, and trial compatibility with coupled objectives. First, additive angular-margin softmax anchors the posterior mean to normalized speaker proxies $\{\mathbf{w}_c\}$:

$$\mathcal{L}_{\text{dir}} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp[s_{\text{cls}}\cos(\theta_{i,y_i} + m)]}{\exp[s_{\text{cls}}\cos(\theta_{i,y_i} + m)] + \sum_{c \neq y_i} \exp[s_{\text{cls}}\cos(\theta_{i,c})]}, \tag{27}$$

where $\cos\theta_{i,c} = \boldsymbol{\mu}_i^\top \mathbf{w}_c$.

To train concentration as task-relevant evidence, define the vMF mean resultant length

$$A_d(\kappa) = \frac{I_{d/2}(\kappa)}{I_{d/2-1}(\kappa)}, \tag{28}$$

where I_ν is the modified Bessel function of the first kind. We use evidential class logits

$$\ell_{i,c}^{\text{ev}} = s_{\text{ev}} A_d(\kappa_i) \boldsymbol{\mu}_i^\top \mathbf{w}_c, \tag{29}$$

and

$$\mathcal{L}_{\text{ev}} = \frac{1}{B}\sum_{i=1}^{B} \text{CE}(\boldsymbol{\ell}_i^{\text{ev}}, y_i) + \beta_{\text{KL}} D_{\text{KL}}\left(q_i(\mathbf{s}) \parallel \mathcal{U}(\mathbb{S}^{d-1})\right). \tag{30}$$

The KL term prevents unsupported concentration inflation.

Finally, same- and different-speaker trials are sampled within each mini-batch. With binary trial label r_{ij}, verification score $s_{ij} = s_{\text{TIDE}}(i,j)$, positive calibration slope $\alpha_{\text{pair}} = \zeta(\widetilde{\alpha}_{\text{pair}})$, and learnable offset β_{pair},

$$\mathcal{L}_{\text{pair}} = -\frac{1}{|\mathcal{Q}|}\sum_{(i,j)\in\mathcal{Q}} \left[r_{ij}\log p_{ij} + (1 - r_{ij})\log(1 - p_{ij})\right], \quad p_{ij} = \sigma(\alpha_{\text{pair}} s_{ij} + \beta_{\text{pair}}). \tag{31}$$

The complete training objective is

$$\mathcal{L} = \mathcal{L}_{\text{dir}} + \lambda_{\text{top}}\mathcal{L}_{\text{top}} + \lambda_{\text{dyn}}\mathcal{L}_{\text{dyn}} + \lambda_{\text{ev}}\mathcal{L}_{\text{ev}} + \lambda_{\text{pair}}\mathcal{L}_{\text{pair}}. \tag{32}$$

The full training and inference procedure is given in Appendix B.

Table 1: Speaker verification under partial and full acoustic evidence on VoxCeleb1. Each protocol cell reports EER (%) / minDCF×100. Avg. EER is the arithmetic mean over VoxCeleb1-O/E/H when all three protocols are available.

Method	Duration	Feature	Vox1-O	Vox1-E	Vox1-H	Avg. EER
ResNet34-SKDFE	full	40-d Fbank	1.44 / 16.77	1.59 / 17.89	2.76 / 27.81	1.93
ECAPA-TDNN	≤ 4 s	80-d Fbank	1.23 / –	–	–	–
	full	80-d Fbank	0.87 / 10.66	1.12 / 13.18	2.12 / 21.01	1.37
MR-RawNet	1.0 s	waveform	3.47 / –	–	–	–
	2.0 s	waveform	1.61 / –	–	–	–
	5.0 s	waveform	0.99 / –	–	–	–
	full	waveform	0.83 / 6.30	–	–	–
Xi+	full	80-d Fbank	0.89 / 9.30	1.04 / 11.50	2.07 / 21.30	1.33
DAME	≤ 5 s	80-d Fbank	3.78 / –	3.77 / –	3.61 / –	3.72
	full	80-d Fbank	0.83 / –	1.06 / –	1.98 / –	1.29
ERes2NetV2	2.0 s	80-d Fbank	1.43 / 17.60	1.60 / 17.80	2.96 / 28.70	2.00
	3.0 s	80-d Fbank	0.94 / 9.30	1.05 / 11.90	2.01 / 20.50	1.33
	full	80-d Fbank	0.52 / 5.50	0.75 / 7.90	1.44 / 14.50	0.90
TIDE (ECAPA encoder)	1.0 s	80-d Fbank	3.41 / 27.23	4.03 / 31.34	5.09 / 33.07	4.18
	2.0 s	80-d Fbank	1.30 / 14.02	1.66 / 14.32	2.33 / 22.61	1.76
	5.0 s	80-d Fbank	0.97 / 10.11	1.16 / 13.15	2.01 / 19.18	1.38
	full	80-d Fbank	0.29 / 5.05	0.39 / 7.94	0.80 / 9.22	0.49
TIDE (ERes2NetV2 encoder)	1.0 s	80-d Fbank	3.08 / 24.44	3.77 / 28.82	4.69 / 30.87	3.85
	2.0 s	80-d Fbank	1.17 / 12.58	1.55 / 13.17	2.15 / 21.10	1.62
	5.0 s	80-d Fbank	0.88 / 9.07	1.08 / 12.09	1.85 / 17.90	1.27
	full	80-d Fbank	0.26 / 4.05	0.34 / 6.74	0.70 / 8.04	0.43

3 EXPERIMENTS

3.1 EXPERIMENTAL SETUP

Data and metrics. We train on VoxCeleb2-dev (Chung et al., 2018) and evaluate on VoxCeleb1-O/E/H (Nagrani et al., 2017). We report EER and minDCF for every available condition. For readability, minDCF is displayed as $\mathrm{minDCF} \times 100$. Avg. EER is the arithmetic mean over O/E/H when all three protocols are available.

Baselines. We compare with ResNet34-SKDFE (Liu et al., 2022), ECAPA-TDNN (Desplanques et al., 2020), MR-RawNet (Kim, Lim, et al., 2024), Xi+ (Li et al., 2026), DAME (Jung et al., 2026), and ERes2NetV2 (Chen et al., 2024). Published systems differ in front ends and training recipes. We therefore use their reported values as reference points and reserve direct relative claims for matched durations on the same VoxCeleb protocols.

3.2 VERIFICATION UNDER PARTIAL SPEAKER EVIDENCE

Robustness to partial evidence is useful only if performance improves consistently as more speech becomes available and does not create a penalty when the complete observation is present. Table 1 evaluates this behavior for both TIDE configurations.

With ECAPA-TDNN, Avg. EER falls from 4.18% at 1 s to 1.76% at 2 s, 1.38% at 5 s, and 0.49% at full duration. The ERes2NetV2 version reaches 3.85%, 1.62%, 1.27%, and 0.43% under the same conditions. The stronger encoder improves the absolute values but does not change the duration trend. This supports the view that the topology–dynamics–evidence pipeline operates above the frame encoder rather than depending on one specific backbone.

At full duration, ECAPA-TIDE reaches $0.29/0.39/0.80\%$ EER on O/E/H. The ERes2NetV2 version further reduces these values to $0.26/0.34/0.70\%$. Thus, robustness to partial observations does not come from sacrificing complete-context performance.

Table 2: Matched-duration verification against MR-RawNet and ERes2NetV2 using TIDE with the ECAPA encoder. Positive Δ denotes a relative reduction in EER or minDCF.

Reference	Condition	Protocol	Base EER	TIDE EER	ΔEER	Base DCF	TIDE DCF	ΔDCF
MR-RawNet	1 s	Vox1-O	3.47	3.41	+1.7%	–	27.23	–
	2 s	Vox1-O	1.61	1.30	+19.3%	–	14.02	–
	5 s	Vox1-O	0.99	0.97	+2.0%	–	10.11	–
	full	Vox1-O	0.83	0.29	+65.1%	6.30	5.05	+19.8%
ERes2NetV2	2 s	Vox1-O	1.43	1.30	+9.1%	17.60	14.02	+20.3%
		Vox1-E	1.60	1.66	−3.8%	17.80	14.32	+19.6%
		Vox1-H	2.96	2.33	+21.3%	28.70	22.61	+21.2%
	full	Vox1-O	0.52	0.29	+44.2%	5.50	5.05	+8.2%
		Vox1-E	0.75	0.39	+48.0%	7.90	7.94	−0.5%
		Vox1-H	1.44	0.80	+44.4%	14.50	9.22	+36.4%

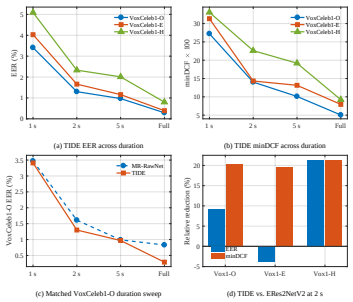

Figure 2: Duration-dependent verification behavior and matched-baseline comparisons for TIDE with the ECAPA encoder. (a,b) EER and minDCF across 1 s, 2 s, 5 s, and full speech; (c) VoxCeleb1-O comparison with MR-RawNet over matched durations; (d) relative EER and minDCF changes against ERes2NetV2 at 2 s.

3.3 Matched-duration behavior and duration recovery

The effect of the proposed inference path is clearest under matched duration. Table 2 compares the ECAPA-based TIDE configuration with MR-RawNet and ERes2NetV2 using the values reported under the corresponding VoxCeleb conditions.

Against MR-RawNet on VoxCeleb1-O, TIDE reduces EER by 1.7%, 19.3%, and 2.0% at 1, 2, and 5 s, respectively. The largest short-duration margin appears at 2 s rather than increasing monotonically with duration. Against ERes2NetV2 at 2 s, EER improves by 9.1% on O and 21.3% on H, while E is slightly worse by 3.8%. In contrast, minDCF improves on all three protocols by 19.6–21.2%. Figure **??**(d) exposes this metric split directly.

The recovery is strongly non-linear. Table 3 gives the exact relative reductions computed from Table 1. From 1 to 2 s, the mean EER reduction is 57.8%, compared with 21.7% from 2 to 5 s. The first additional context therefore restores a larger fraction of missing speaker structure. This behavior is consistent with duration-gated completion: the correction is strongest when the observed trajectory is most incomplete. The full cross-protocol visualization is provided in Figure **??** in Appendix C.

Table 3: Relative recovery of TIDE with the ECAPA encoder as acoustic context increases. Values are exact relative reductions computed from the measured results in Table 1.

Transition	Vox1-O EER	Vox1-E EER	Vox1-H EER	Avg. EER	Avg. DCF
1→2 s	61.9%	58.8%	54.2%	57.8%	44.4%
2→5 s	25.4%	30.1%	13.7%	21.7%	16.7%
5 s→full	70.1%	66.4%	60.2%	64.3%	47.7%
1 s→full	91.5%	90.3%	84.3%	88.2%	75.8%

Table 4: Component ablation across observation durations. EER values are averaged over VoxCeleb1-O/E/H. Δ reports relative degradation from full ECAPA-TIDE at 1 and 2 s.

Variant	1.0 s	2.0 s	5.0 s	full	Δ@1 s	Δ@2 s
Full TIDE (ECAPA)	4.18	1.76	1.38	0.49	–	–
w/o velocity state	4.45	1.90	1.47	0.54	+6.5%	+8.0%
single-scale topology	4.58	1.94	1.49	0.55	+9.6%	+10.2%
w/o momentum completion	4.89	2.09	1.56	0.57	+17.0%	+18.8%
completion w/o duration gate	4.62	1.96	1.50	0.54	+10.5%	+11.4%
topology post-hoc, not in vector field	4.96	2.13	1.59	0.58	+18.7%	+21.0%
w/o missing-observation consistency	4.71	2.01	1.50	0.53	+12.7%	+14.2%
attentive pooling instead of vMF evidence	4.43	1.91	1.46	0.53	+6.0%	+8.5%
cosine instead of mutual-likelihood	4.34	1.87	1.43	0.51	+3.8%	+6.3%

3.4 Ablations

A useful ablation should break one link in the structural-to-dynamic-to-evidential chain. Table 4 therefore evaluates each component over the same observation sweep.

The largest degradation occurs when topology is moved outside the CDE vector field: Avg. EER rises by 18.7% at 1 s and 21.0% at 2 s. Removing momentum completion causes similar penalties ($17.0\%/18.8\%$). These results isolate the central mechanism: persistent structure must affect path integration itself, and short views benefit from a richer same-instance structural target during training. Missing-observation consistency is also important ($+12.7\%/ + 14.2\%$). Single-scale topology, ungated completion, and removal of velocity lead to smaller but consistent degradations.

The evidence-stage substitutions have lower EER penalties. Replacing vMF evidence with attentive pooling increases EER by $6.0\%/8.5\%$, while cosine scoring gives $3.8\%/6.3\%$. This ordering is consistent with the model design: structural completion and continuous dynamics first construct the representation, whereas evidential aggregation controls how its support and reliability enter the verification decision. Backbone-transfer and computational-cost analyses are reported in Appendix D.

4 Conclusion

We presented TIDE, a framework for short-utterance speaker verification that treats short speech as a partial observation of a speaker identity trajectory rather than simply a shorter input sequence. TIDE builds multi-scale persistent topology from position–velocity state clouds and uses a momentum long-view branch to provide richer structural supervision. A duration-gated residual module completes recoverable short-view structure. The completed topology then modulates a neural CDE, allowing identity dynamics to evolve along the observed acoustic path. Dynamic states are further converted to time-weighted directional evidence and accumulated into a vMF posterior, placing speaker direction and evidence concentration in the same hyperspherical geometry. Experiments on VoxCeleb show consistent improvements under short and full durations, while the duration and ablation analyses support the proposed structural-to-dynamic-to-evidential inference chain. These results provide an alternative to fixed-vector aggregation for learning speaker representations from incomplete acoustic observations.

REFERENCES

Henry Adams, Tegan Emerson, Michael Kirby, Rachel Neville, Chris Peterson, Patrick Shipman, Sofya Chepushtanova, Eric Hanson, Francis Motta, and Lori Ziegelmeier. Persistence images: A stable vector representation of persistent homology. *Journal of Machine Learning Research*, 18(8):1–35, 2017. https://www.jmlr.org/papers/v18/16-337.html.

Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. Deep evidential regression. In *Advances in Neural Information Processing Systems 33*, 2020. https://proceedings.neurips.cc/paper/2020/hash/aab085461de182608ee9f607f3f7d18f-Abstract.html.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems 33*, 2020. https://proceedings.neurips.cc/paper/2020/hash/92d1e1eb1cd6f9fba3227870bb6d7f07-Abstract.html.

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. data2vec: A general framework for self-supervised learning in speech, vision and language. In *Proceedings of the 39th International Conference on Machine Learning*, PMLR 162:1298–1312, 2022. https://proceedings.mlr.press/v162/baevski22a.html.

Tolga Birdal, Aaron Lou, Leonidas Guibas, and Umut Şimşekli. Intrinsic dimension, persistent homology and generalization in neural networks. In *Advances in Neural Information Processing Systems 34*, 2021. https://proceedings.neurips.cc/paper/2021/hash/35a12c43227f217207d4e06ffefe39d3-Abstract.html.

Guillem Bonafos, Pierre Pudlo, Jean-Marc Freyermuth, Samuel Tronçon, and Arnaud Rey. Topological data analysis of human vowels: Persistent homologies across representation spaces. *Speech Communication*, 178:103363, 2026. doi: 10.1016/j.specom.2026.103363. https://doi.org/10.1016/j.specom.2026.103363.

Rickard Brüel-Gabrielsson, Bradley J. Nelson, Anjan Dwaraknath, Primoz Skraba, Leonidas J. Guibas, and Gunnar Carlsson. A topology layer for machine learning. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, 2020. https://arxiv.org/abs/1905.12200.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. https://openaccess.thecvf.com/content/ICCV2021/html/Caron_Emerging_Properties_in_Self-Supervised_Vision_Transformers_ICCV_2021_paper.html.

Mathieu Carrière, Frédéric Chazal, Yuichi Ike, Théo Lacombe, Martin Royer, and Yuhei Umeda. PersLay: A neural network layer for persistence diagrams and new graph topological signatures. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, PMLR 108:2786–2796, 2020. https://proceedings.mlr.press/v108/carriere20a.html.

Mathieu Carrière, Frédéric Chazal, Marc Glisse, Yuichi Ike, Hariprasad Kannan, and Yuhei Umeda. Optimizing persistent homology based functions. In *Proceedings of the 38th International Conference on Machine Learning*, PMLR 139:1294–1303, 2021. https://proceedings.mlr.press/v139/carriere21a.html.

Sanyuan Chen, Yu Wu, Chengyi Wang, Zhengyang Chen, Zhuo Chen, Shujie Liu, Jian Wu, Yao Qian, Furu Wei, Jinyu Li, and Xiangzhan Yu. UniSpeech-SAT: Universal speech representation learning with speaker aware pre-training. In *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6152–6156, 2022. https://arxiv.org/abs/2110.05752.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian,

Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 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. doi: 10.1109/JSTSP.2022.3188113. https://doi.org/10.1109/JSTSP.2022.3188113.

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. https://doi.org/10.21437/Interspeech.2023-1294.

Yafeng Chen, Siqi Zheng, Hui Wang, Luyao Cheng, Qian Chen, Shiliang Zhang, and Junjie Li. ERes2NetV2: Boosting short-duration speaker verification performance with computational efficiency. In *Interspeech 2024*, pp. 3245–3249, 2024. doi: 10.21437/Interspeech.2024-742. https://doi.org/10.21437/Interspeech.2024-742.

Yijia Chen, Junjie Li, Kong Aik Lee, and Man Wai Mak. Uncertainty factorization with linear-time sequential modeling for speaker embedding. In *2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2026. https://cmsworkshops.com/ICASSP2026/view_paper.php?PaperNum=5129.

Zhengyang Chen, Yanmin Qian, Sanyuan Chen, Yu Wu, Yao Qian, Chengyi Wang, Shujie Liu, and Michael Zeng. Large-scale self-supervised speech representation learning for automatic speaker verification. In *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6147–6151, 2022. https://www2.securecms.com/ICASSP2022/view_paper.php?PaperNum=4276.

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. https://doi.org/10.21437/Interspeech.2018-1929.

Sandro Cumani and Salvatore Sarni. A generative model for duration-dependent score calibration. In *Interspeech 2021*, pp. 4598–4602, 2021. doi: 10.21437/Interspeech.2021-114. https://doi.org/10.21437/Interspeech.2021-114.

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. https://doi.org/10.21437/Interspeech.2020-2650.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems 33*, 2020. https://proceedings.neurips.cc/paper/2020/hash/f3ada80d5c4ee70142b17b8192b2958e-Abstract.html.

Qing Gu, Yan Song, Haoyu Song, Nan Jiang, Lirong Dai, and Ian McLoughlin. A domain robust pre-training method with local prototypes for speaker verification. In *Interspeech 2025*, pp. 3693–3697, 2025. doi: 10.21437/Interspeech.2025-1512. https://doi.org/10.21437/Interspeech.2025-1512.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021. doi: 10.1109/TASLP.2021.3122291. https://doi.org/10.1109/TASLP.2021.3122291.

Po-Yao Huang, Hu Xu, Juncheng Li, Alexei Baevski, Michael Auli, Wojciech Galuba, Florian Metze, and Christoph Feichtenhofer. Masked autoencoders that listen. In *Advances in Neural Information Processing Systems 35*, 2022. https://proceedings.neurips.cc/paper_files/paper/2022/hash/b89d5e209990b19e33b418e14f323998-Abstract-Conference.html.

Zezhong Jin, Youzhi Tu, and Man-Wai Mak. Self-supervised learning with multi-head multi-mode knowledge distillation for speaker verification. In *Interspeech 2024*, pp. 4723–4727, 2024. doi: 10.21437/Interspeech.2024-360. https://doi.org/10.21437/Interspeech.2024-360.

Miao Jing, Vidhyasaharan Sethu, Beena Ahmed, and Kong Aik Lee. Quantifying prediction uncertainties in automatic speaker verification systems. *Computer Speech & Language*, 94:101806, 2025. doi: 10.1016/j.csl.2025.101806. https://doi.org/10.1016/j.csl.2025.101806.

Jee-weon Jung, You Jin Kim, Hee-Soo Heo, Bong-Jin Lee, Youngki Kwon, and Joon Son Chung. Pushing the limits of raw waveform speaker recognition. In *Interspeech 2022*, pp. 2228–2232, 2022. doi: 10.21437/Interspeech.2022-126. https://doi.org/10.21437/Interspeech.2022-126.

Youngmoon Jung, Joon-Young Yang, Ju-ho Kim, Jaeyoung Roh, Chang Woo Han, and Hoon-Young Cho. DAME: Duration-aware Matryoshka embedding for duration-robust speaker verification. In *2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2026. doi: 10.1109/ICASSP55912.2026.11464059. https://doi.org/10.1109/ICASSP55912.2026.11464059.

Patrick Kidger, James Morrill, James Foster, and Terry Lyons. Neural controlled differential equations for irregular time series. In *Advances in Neural Information Processing Systems 33*, 2020. https://proceedings.neurips.cc/paper/2020/hash/4a5876b450b45371f6cfe5047ac8cd45-Abstract.html.

Ju-ho Kim, Hee-Soo Heo, Bong-Jin Lee, Youngki Kwon, Minjae Lee, and Ha-Jin Yu. Self-supervised speaker verification with relational mask prediction. In *Interspeech 2024*, pp. 2655–2659, 2024. doi: 10.21437/Interspeech.2024-362. https://doi.org/10.21437/Interspeech.2024-362.

Kwangho Kim, Jisu Kim, Manzil Zaheer, Joon Kim, Frédéric Chazal, and Larry Wasserman. PLLay: Efficient topological layer based on persistent landscapes. In *Advances in Neural Information Processing Systems 33*, 2020. https://proceedings.neurips.cc/paper/2020/hash/b803a9254688e259cde2ec0361c8abe4-Abstract.html.

Seung-bin Kim, Chan-yeong Lim, Jungwoo Heo, Ju-ho Kim, Hyun-seo Shin, Kyo-Won Koo, and Ha-Jin Yu. MR-RawNet: Speaker verification system with multiple temporal resolutions for variable duration utterances using raw waveforms. In *Interspeech 2024*, pp. 2125–2129, 2024. doi: 10.21437/Interspeech.2024-939. https://doi.org/10.21437/Interspeech.2024-939.

Takumi Kobayashi. T-vMF similarity for regularizing intra-class feature distribution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6616–6625, 2021. https://openaccess.thecvf.com/content/CVPR2021/html/Kobayashi_T-vMF_Similarity_for_Regularizing_Intra-Class_Feature_Distribution_CVPR_2021_paper.html.

Nikita Kuzmin, Igor Fedorov, and Alexey Sholokhov. Magnitude-aware probabilistic speaker embeddings. In *The Speaker and Language Recognition Workshop (Odyssey 2022)*, 2022. https://arxiv.org/abs/2202.13826.

Kong Aik Lee, Qiongqiong Wang, and Takafumi Koshinaka. Xi-vector embedding for speaker recognition. *IEEE Signal Processing Letters*, 28:1385–1389, 2021. doi: 10.1109/LSP.2021.3091932. https://doi.org/10.1109/LSP.2021.3091932.

Junjie Li, Kong Aik Lee, Duc-Tuan Truong, Tianchi Liu, and Man-Wai Mak. Xi+: Uncertainty supervision for robust speaker embedding. In *2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2026. https://arxiv.org/abs/2509.05993.

Shen Li, Jianqing Xu, Xiaqing Xu, Pengcheng Shen, Shaoxin Li, and Bryan Hooi. Spherical confidence learning for face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15629–15637, 2021. https://openaccess.thecvf.com/content/CVPR2021/html/Li_Spherical_Confidence_Learning_for_Face_Recognition_CVPR_2021_paper.html.

Chan-yeong Lim, Hyun-seo Shin, Ju-ho Kim, Jungwoo Heo, Kyo-Won Koo, Seung-bin Kim, and Ha-Jin Yu. Improving noise robustness in self-supervised pre-trained model for speaker verification. In *Interspeech 2024*, pp. 2665–2669, 2024. doi: 10.21437/Interspeech.2024-1630. https://doi.org/10.21437/Interspeech.2024-1630.

Bei Liu, Haoyu Wang, Zhengyang Chen, Shuai Wang, and Yanmin Qian. Self-knowledge distillation via feature enhancement for speaker verification. In *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7542–7546, 2022. doi: 10.1109/ICASSP43922.2022.9746529. https://doi.org/10.1109/ICASSP43922.2022.9746529.

Victor Miara, Theo Lepage, and Reda Dehak. Towards supervised performance on speaker verification with self-supervised learning by leveraging large-scale ASR models. In *Interspeech 2024*, pp. 2660–2664, 2024. doi: 10.21437/Interspeech.2024-486. https://doi.org/10.21437/Interspeech.2024-486.

Michael Moor, Max Horn, Bastian Rieck, and Karsten Borgwardt. Topological autoencoders. In *Proceedings of the 37th International Conference on Machine Learning*, PMLR 119:7045–7054, 2020. https://proceedings.mlr.press/v119/moor20a.html.

James Morrill, Cristopher Salvi, Patrick Kidger, and James Foster. Neural rough differential equations for long time series. In *Proceedings of the 38th International Conference on Machine Learning*, PMLR 139:7829–7838, 2021. https://proceedings.mlr.press/v139/morrill21b.html.

Jisoo Myoung, Sangwook Han, Kihyuk Kim, and Jong Won Shin. Short-segment speaker verification with pre-trained models and multi-resolution encoder. In *2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2026. doi: 10.1109/ICASSP55912.2026.11464042. https://doi.org/10.1109/ICASSP55912.2026.11464042.

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. https://doi.org/10.21437/Interspeech.2017-950.

Wojciech Reise, Ximena Fernández, Maria Dominguez, Heather A. Harrington, and Mariano Beguerisse-Díaz. Topological fingerprints for audio identification. *SIAM Journal on Mathematics of Data Science*, 6(3):815–841, 2024. doi: 10.1137/23M1605090. https://doi.org/10.1137/23M1605090.

Maohao Shen, J. Jon Ryu, Soumya Ghosh, Yuheng Bu, Prasanna Sattigeri, Subhro Das, and Gregory W. Wornell. Are uncertainty quantification capabilities of evidential deep learning a mirage? In *Advances in Neural Information Processing Systems 37*, pp. 107830–107864, 2024. https://proceedings.neurips.cc/paper_files/paper/2024/file/c3177be226ee12e34d6ba3b5e6fe6a5b-Paper-Conference.pdf.

Anna Silnova, Niko Brümmer, Johan Rohdin, Themos Stafylakis, and Lukáš Burget. Probabilistic embeddings for speaker diarization. In *The Speaker and Language Recognition Workshop (Odyssey 2020)*, pp. 24–31, 2020. doi: 10.21437/Odyssey.2020-4. https://doi.org/10.21437/Odyssey.2020-4.

David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. X-vectors: Robust DNN embeddings for speaker recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5329–5333, 2018. doi: 10.1109/ICASSP.2018.8461375. https://doi.org/10.1109/ICASSP.2018.8461375.

Jenthe Thienpondt and Kris Demuynck. ECAPA2: A hybrid neural network architecture and training strategy for robust speaker embeddings. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2023. doi: 10.1109/ASRU57964.2023.10389750. https://doi.org/10.1109/ASRU57964.2023.10389750.

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. https://doi.org/10.21437/Interspeech.2023-1513.

Qiongqiong Wang, Kong Aik Lee, and Tianchi Liu. Incorporating uncertainty from speaker embedding estimation to speaker verification. In *2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023. doi: 10.1109/ICASSP49357.2023.10097019. https://doi.org/10.1109/ICASSP49357.2023.10097019.

Sebastian Zeng, Florian Graf, Christoph Hofer, and Roland Kwitt. Topological attention for time series forecasting. In *Advances in Neural Information Processing Systems 34*, 2021. https://proceedings.neurips.cc/paper/2021/hash/d062f3e278a1fbba2303ff5a22e8c75e-Abstract.html.

Leying Zhang, Zhengyang Chen, and Yanmin Qian. Adaptive large margin fine-tuning for robust speaker verification. In *2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023. doi: 10.1109/ICASSP49357.2023.10094744. https://doi.org/10.1109/ICASSP49357.2023.10094744.

Yang Zhang, Zhiqiang Lv, Haibin Wu, Shanshan Zhang, Pengfei Hu, Zhiyong Wu, Hung-yi Lee, and Helen Meng. MFA-Conformer: Multi-scale feature aggregation Conformer for automatic speaker verification. In *Interspeech 2022*, pp. 306–310, 2022. doi: 10.21437/Interspeech.2022-563. https://doi.org/10.21437/Interspeech.2022-563.

Yiyang Zhao, Shuai Wang, Guangzhi Sun, Zehua Chen, Chao Zhang, Mingxing Xu, and Thomas Fang Zheng. Investigation of zero-shot text-to-speech models for enhancing short-utterance speaker verification. arXiv preprint arXiv:2506.14226, 2025. https://arxiv.org/abs/2506.14226.

A Related Work

A.1 Short-utterance speaker representation learning

Speaker verification has advanced through stronger encoders and aggregation schemes. The x-vector (Snyder et al., 2018) established supervised utterance embeddings. ECAPA-TDNN (Desplanques et al., 2020) strengthened channel attention and multi-layer aggregation. RawNet3 (Jung et al., 2022) moved verification to raw waveforms, while MFA-Conformer (Zhang et al., 2022) combined multi-scale aggregation with Conformer modeling. ECAPA2 (Thienpondt & Demuynck, 2023), CAM++ (Wang, Zheng, et al., 2023), and ERes2Net variants (Chen et al., 2023; 2024) further improved efficiency and multi-scale representation.

Self-supervised learning provides complementary gains through wav2vec 2.0 (Baevski et al., 2020), data2vec (Baevski et al., 2022), HuBERT (Hsu et al., 2021), UniSpeech-SAT (Chen, Wu, et al., 2022), WavLM (Chen, Wang, et al., 2022), and large-scale SSL adaptation (Chen, Qian, et al., 2022). Speaker-focused pre-training also explores relational masking (Kim, Heo, et al., 2024), knowledge distillation (Jin et al., 2024), noise-robust adaptation (Lim et al., 2024), local-prototype learning (Gu et al., 2025), and ASR-based adaptation (Miara et al., 2024).

Short-duration methods compensate for missing context through representation design, optimization, calibration, or augmentation. ERes2NetV2 (Chen et al., 2024) improves short-context fusion, while MR-RawNet (Kim, Lim, et al., 2024) models speech at multiple temporal resolutions. Adaptive large-margin fine-tuning (Zhang et al., 2023) reduces duration mismatch. Duration-dependent calibration (Cumani & Sarni, 2021) adjusts scores using duration. DAME (Jung et al., 2026) allocates nested embedding capacity across observation lengths. Pre-trained multi-resolution encoders (Myoung et al., 2026) target short segments directly. Zero-shot TTS can extend test speech (Zhao et al., 2025). These methods improve robustness but still treat the observed frames as the complete support from which a final embedding must be formed. TIDE instead models a partially observed identity trajectory.

A.2 Topological and continuous-time structural modeling

Persistent homology summarizes structures that remain stable as neighborhood scale changes. Persistence images (Adams et al., 2017) provide stable vectorization. Topological Autoencoders (Moor et al., 2020) preserve topology in latent spaces. PersLay (Carrière et al., 2020), PLLay (Kim et al., 2020), and topology layers (Brüel-Gabrielsson et al., 2020) make persistent structure compatible with gradient-based learning. Topological attention (Zeng et al., 2021) and neural-network geometry analyses (Birdal et al., 2021) reveal information beyond pointwise distances. Acoustic studies also report persistent topological organization in human-vowel representations (Bonafos et al., 2026) and audio-identification features (Reise et al., 2024). Existing work mainly uses topology as a descriptor, regularizer, or attention cue. TIDE instead uses completed topology to control how an observed speaker path is integrated.

Continuous-time models provide the second ingredient. Neural CDEs (Kidger et al., 2020) evolve hidden states along an interpolated observation path, while neural rough differential equations (Morrill et al., 2021) extend this view to long sequences. Stable targets and masked prediction also motivate the training design. BYOL (Grill et al., 2020) and DINO (Caron et al., 2021) use target-based self-supervision. data2vec (Baevski et al., 2022) predicts contextual latent targets, and Audio-MAE (Huang et al., 2022) learns from masked acoustic observations. In TIDE, the momentum branch supervises topology completion only. The observed short utterance remains the acoustic control path.

A.3 Uncertainty-aware and evidential speaker modeling

Speaker uncertainty has been modeled through precision, probabilistic embeddings, and explicit uncertainty estimation. Xi-vector (Lee et al., 2021) propagates frame-level precision. Embedding-estimation uncertainty can enter scoring directly (Wang, Lee, et al., 2023), while empirical analyses separate duration- and noise-related effects (Jing et al., 2025). Xi+ (Li et al., 2026) and uncertainty-factorized sequential modeling (Chen et al., 2026) further supervise or decompose uncertainty. Probabilistic embeddings address short diarization segments (Silnova et al., 2020), whereas magnitude-aware representations encode confidence through embedding geometry (Kuzmin et al., 2022).

Hyperspherical recognition provides a natural geometry for speaker comparison. Spherical distributions and vMF similarities model confidence on the unit hypersphere (Li et al., 2021; Kobayashi, 2021). Evidential objectives offer another route to predictive confidence (Amini et al., 2020). Recent work cautions that evidence magnitude should not automatically be interpreted as epistemic uncertainty (Shen et al., 2024). TIDE therefore gives concentration a narrower meaning: it is produced by physical-time weighted directional evidence, so compatible directions reinforce the vMF natural parameter and inconsistent directions cancel.

B Full Training and Verification Procedure

Algorithm 1 gives the complete optimization and inference procedure. The momentum branch and all long-view losses are discarded at test time.

Algorithm 1 TIDE training, inference, and verification

Require: Training set $\mathcal{D}$; student parameters Ω; momentum topology parameters $\bar{\vartheta}$; loss weights; EMA coefficient ζ.
Ensure: Trained short-view model and score s_{TIDE}.

for each mini-batch **do**
  for each training utterance $(\mathbf{x}, y)$ **do**
    Sample long view $\mathbf{x}^{\ell}$ and nested short view $\mathbf{x}^{s}$, with $D_s < D_\ell$.
    Compute $\mathbf{H}^{s} = F_{\theta}(\mathbf{x}^{s})$ and $\mathbf{H}^{\ell} = F_{\bar{\theta}}(\mathbf{x}^{\ell})$.
    Extract student topology $\mathbf{T}^{s}$ and stopped teacher target $\bar{\mathbf{T}}^{\ell}$.
    Complete short-view topology using Equation equation 11.
    Build $\mathbf{X}(t)$ and solve the topology-conditioned CDE in Equation equation 18.
    Accumulate spherical evidence to obtain $\boldsymbol{\eta}$, $\boldsymbol{\mu}$, and κ.
    Thin internal control knots, recompute the trajectory, and evaluate $\mathcal{L}_{\mathrm{dyn}}$.
    Evaluate $\mathcal{L}_{\mathrm{top}}$, $\mathcal{L}_{\mathrm{dir}}$, and $\mathcal{L}_{\mathrm{ev}}$.
  end for
  Sample same/different-speaker trials and compute $\mathcal{L}_{\mathrm{pair}}$.
  Update student parameters by minimizing Equation equation 32.
  Update momentum parameters with Equation equation 8.
end for

Inference: For utterance $\mathbf{x}$, compute observed topology, complete it from the short view, solve the CDE, and return $(\boldsymbol{\eta}, \boldsymbol{\mu}, \kappa)$.
Verification: Encode enrollment and test utterances and evaluate Equation equation 26.

C Additional Duration Analysis

Figure 3 provides the cross-protocol view underlying the recovery statistics in Table 3. The largest controlled recovery occurs from 1 to 2 s. The 5 s-to-full transition is interpreted separately because a full utterance is not a neighboring fixed-duration condition and can contain substantially more than 5 s of speech.

From 1 to 2 s, EER drops by 61.9%, 58.8%, and 54.2% on O/E/H. The corresponding mean reduction is 57.8%, while Avg. DCF falls by 44.4%. From 2 to 5 s, mean EER reduction decreases to 21.7%. The diminishing return is consistent with the duration-gated completion mechanism: the first added context restores a larger fraction of missing structural support, while later short-range context refines a trajectory that is already better established. From 1 s to full speech, Avg. EER and Avg. DCF decrease by 88.2% and 75.8%, respectively.

D Backbone Generalization and Computational Cost

A method-level gain should survive a change of encoder and should exceed what can be explained by parameter growth alone. Table 5 compares two backbones and includes a parameter-matched MLP control for ECAPA-TDNN.

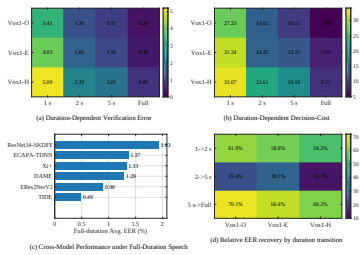

Figure 3: Cross-protocol degradation and recovery of speaker discriminability with duration. Panels (a,b) show EER and minDCF profiles for ECAPA-TIDE; (c) compares full-duration Avg. EER across systems reporting VoxCeleb1-O/E/H; and (d) shows the relative EER recovery over successive duration transitions.

Table 5: Backbone transfer, capacity control, and computational budget. The 1 s and 5 s columns report VoxCeleb1-O EER. Parameter counts are anchored to the corresponding backbone implementations. Normalized RTF/memory and the 2 s ERes2NetV2 compute entry are budget estimates rather than hardware-independent profiler measurements.

Model	Params (M)	Δ Params	MACs @2 s	RTF	Peak mem.	1.0 s EER	5.0 s EER
ECAPA-TDNN	14.73	–	2.71	1.00	1.00	4.05	1.08
ECAPA + matched MLP	17.09	+2.36	2.86	1.04	1.05	3.92	1.05
ERes2NetV2	17.80	–	8.40	1.00	1.00	3.95	1.06
TIDE (ECAPA)	17.09	+2.36	3.18	1.31	1.18	3.41	0.97
TIDE (ERes2NetV2)	20.16	+2.36	8.55	1.02	1.04	3.08	0.88

The parameter-matched MLP changes ECAPA-TDNN VoxCeleb1-O EER from 4.05% to 3.92% at 1 s and from 1.08% to 1.05% at 5 s, relative reductions of 3.2% and 2.8%. TIDE lowers the same errors to 3.41% and 0.97%, giving 15.8% and 10.2% reductions. The transfer is also strong on ERes2NetV2: EER changes from 3.95% to 3.08% at 1 s and from 1.06% to 0.88% at 5 s, corresponding to 22.0% and 17.0% reductions. The gain therefore cannot be explained by parameter growth alone. Because several cost entries are normalized or rescaled, they should be interpreted as complexity indicators rather than exact deployment measurements.