FlowFake: Liquid Networks for Audio Deepfake Detection

Shivaay Dhondiyal [*1] **Divyansh Sharma** [*1] **Dinesh Kumar Vishwakarma** [1]

Abstract

Audio deepfakes generated by neural text-to-speech and voice-cloning systems threaten speaker verification and public discourse at scale. The core challenge is *cross-dataset generalization*: detectors trained on one synthesis pipeline collapse on unseen forgeries. We argue that this failure is primarily because of structural synthetic-speech artifacts which are multi-timescale *trajectory anomalies*. Though every existing detector aggregates a fixed-window frame statistics, this misaligns the architecture with the signal. We propose **FlowFake**, a Liquid Time-Constant (LTC) architecture whose hidden state evolves via a learned ODE, with per-neuron adaptive time constants simultaneously resolving spectral ($\sim$10 ms) and prosodic ($\sim$2 s) cues. At only $\approx$34 K parameters FlowFake achieves formal BIBO stability and $\mathcal{O}(\Delta t^4)$ integration error. On a four-dataset cross-domain benchmark (ASVspoof 2019-LA, FakeOrReal, InTheWild, MLAAD), FlowFake reaches **75.29%** on ASVspoof 2019 trained only on FakeOrReal and **79.97%** trained only on MLAAD. It outperforms RawGAT-ST and Whisper-DF on every evaluated pair and matching SSL Wav2vec2 (300$\times$ larger) at 0.01% of its parameter count. The source code is available on GitHub.

1. Introduction

Neural text-to-speech and voice-cloning systems (Shen et al., 2018; Jia et al., 2018) now produce near-human speech at negligible cost, enabling impersonation attacks against voice-based authentication (Wu et al., 2015; Todisco et al., 2019) and contributing to high-profile disinformation incidents. The practical bottleneck is not raw detection accuracy on a single synthesis family but *cross-dataset generalization*: a detector trained on one TTS pipeline must remain effective against unseen forgeries from different vocoders, languages, and recording conditions (Müller et al., 2022; Frank & Schönherr, 2021). This is the deployment regime that matters; it is also the regime in which every published detector collapses.

Existing countermeasures fall into three families, all of which collapse out of distribution. Graph attention networks (Jung et al., 2022a;b) memorise dataset-specific spectral artifacts: RawGAT-ST trained on FakeOrReal reaches only $49.1 \pm 18.1\%$ on ASVspoof 2019, near random chance. Self-supervised frontends (Tak et al., 2022) fine-tune $\sim$300 M parameter transformers with fixed attention windows; deployment is prohibitive and the cross-seed variance is large ($\pm$17.5 pp on MLAAD$\rightarrow$ITW, Table 1). ASR encoder repurposing (Radford et al., 2023; Müller et al., 2024) inherits representations optimised for recognition semantics, not low-level forgery cues, yielding only 44.9% on MLAAD when trained on ASVspoof 2019.

Our central hypothesis is that the shared failure mode is architectural, not data-driven: synthetic-speech artifacts are trajectory anomalies in how spectro-temporal features evolve over time, but every existing detector aggregates frame-level statistics over a fixed context window, structurally erasing the trajectory information. Physical articulation imposes well-characterised dynamical constraints (vocal tract changes at $\sim$10–100 ms, prosodic contours at $\sim$100-2000 ms). TTS systems violate these constraints in synthesis paradigm specific ways, producing *trajectory anomalies* (deviations in how features evolve) and not in their instantaneous values. A model endowed with the correct structural prior (continuous-time trajectory modelling) should detect synthesis paradigm agnostic violations of articulatory dynamics rather than the fingerprint of any one synthesis pipeline.

Contributions. We introduce **FlowFake**, the first Liquid Time-Constant (Hasani et al., 2021) architecture for audio deepfake detection. Our contributions are:

1. A gradient stable LTC variant with simplified $\tanh$ synapse and log-parameterized adaptive time constants $\tau_i \in [0.05, 10]$ s, integrated by 4th-order Runge-Kutta

[*]Equal contribution
[1]Delhi Technological University, New Delhi, India.

Accepted to the Workshop on Learning to Listen: Machine Learning for Audio at ICML 2026, Seoul, South Korea.

(Section 3).
2. Formal BIBO stability (Theorem 4.2) and $\mathcal{O}(\Delta t^4)$ RK4 error bound (Proposition 4.3); full proofs in Appendix B, where we additionally establish gradient attenuation (Proposition B.7) and noise robustness (Proposition B.4).
3. Cross-dataset accuracy of 75.29% (FoR→ASV19) and 79.97% (MLAAD→ASV19) on ASVspoof 2019 (Table 1), outperforming all baselines on the hardest pairs at 0.01% of SSL Wav2vec2's parameter count.

2. Problem Setup

Let $\mathcal{X}{=}\mathcal{L}^2(\mathbb{R}_{\geq 0};\mathbb{R})$ be the waveform space and $\mathcal{Y}{=}\{0,1\}$ be the label space (0: bonafide, 1: synthetic). A synthesis domain $\mathcal{D}_s$ is a distribution over $\mathcal{X}{\times}\mathcal{Y}$ induced by TTS pipeline $s \in \mathcal{S}$. A detector $f_\theta : \mathcal{X} \to [0,1]$ has domain-specific risk $\mathcal{R}_s(f_\theta){=}\mathbb{E}_{(x,y)\sim\mathcal{D}_s}[\ell(f_\theta(x),y)]$ with ℓ the binary cross-entropy. Our objective is to minimise the cross-domain gap $\mathcal{G} \coloneqq \max_{s\neq s_{\text{tr}}} \mathcal{R}_s(f_\theta) - \mathcal{R}_{s_{\text{tr}}}(f_\theta)$ without access to any test-domain data; a strict no target domain adaptation setting. The hidden state space is $\mathcal{H}{=}\mathbb{R}^H$ with $H{=}32$. Full notation appears in Table 2 (Appendix A).

3. FlowFake Architecture

Frontend: A 16 kHz mono waveform $w \in \mathcal{X}$ is ℓ_2 normalised and converted into to a log-Mel spectrogram $\mathbf{X} \in \mathbb{R}^{N\times T}$ with $N{=}128$ bands, $n_{\text{fft}}{=}512$, hop 160 samples ($\Delta t_{\text{frame}}{\approx}10$ ms; batch dim is suppressed for clarity). The 10 ms resolution maintains exactly the spectro-temporal scale at which neural-vocoder artifacts manifest.

Convolutional encoder: Five 1D-conv layers (kernels $\{5,1,3,3,1\}$, BN, ReLU, $H{=}32$ output channels) squeezes each frame to an H-dimensional embedding $E_t = \text{CNN}(\mathbf{X})_t \in \mathbb{R}^H$ for $t = 1,\ldots,T'{\leq}T$, where T' is the dataset specific LTC coverage length (Table 3). Diversifying kernel sizes capture spectral structure at multiple resolutions without parameter inflation.

Liquid Time-Constant cell: The hidden state $\mathbf{h}(t) \in \mathcal{H}$ updates according to the ODE

$$\frac{d\mathbf{h}(t)}{dt} = \mathbf{C}_m^{-1} \odot \Big[\mathbf{W}_{\text{in}} E_t + \tanh\big(\mathbf{W}_{\text{rec}}\mathbf{h}(t)\big) + g_{\text{leak}} \odot \big(\mathbf{V}_{\text{leak}} - \mathbf{h}(t)\big)\Big], \quad (1)$$

where $\mathbf{C}_m, g_{\text{leak}}, \mathbf{V}_{\text{leak}} \in \mathbb{R}^H$ are learnable membrane capacitance, leak conductance, and resting potential, and $\odot$ denotes the Hadamard product. The three terms play complementary roles. *Input drive* $\mathbf{W}_{\text{in}}E_t$ injects CNN embeddings at every frame. The *recurrent term* $\tanh(\mathbf{W}_{\text{rec}}\mathbf{h}(t))$ provides constrained internal dynamics. The *leak term* $g_{\text{leak}} \odot (\mathbf{V}_{\text{leak}} - \mathbf{h}(t))$ provides a dissipative restoring force that guarantees BIBO stability (Theorem 4.2). We replace the biological sigmoidal synapse of Hasani et al. (2021) with $\tanh(\mathbf{W}_{\text{rec}}\mathbf{h})$, dropping parameters $\approx 3\times$ and stabilizing gradient norms over 50-200 frame rollouts. The original synapse brings in non-monotone gradients that empirically destabilise audio training.

Adaptive time constants: Each neuron has a learnable timescale $\tau_i = \exp(\hat{\tau}_i)$, parameterized in log-space (guaranteeing positivity without projection) and clamped to $[0.05, 10]$ s. *At convergence the learned τ_i are empirically bimodal (a fast cluster at 0.1–0.3 s and a slow cluster at 1.5-5.0 s) capturing the two dominant TTS-artefact timescales (spectral-frame and prosodic-phrase) automatically from data; see Appendix C for empirical distributions.*

RK4 integration and head: Equation (1) is solved via 4th-order Runge-Kutta (Hairer et al., 1993) at $\Delta t{=}0.01$ s with $K{=}2$ solver unfolds per audio frame, yielding global error $\mathcal{O}(\Delta t^4)$ (Proposition 4.3). The terminal state $\mathbf{h}(T')$ routes to a two layer FC head with weights $\mathbf{W}_1 \in \mathbb{R}^{d\times H}$, $\mathbf{W}_2 \in \mathbb{R}^{1\times d}$ ($d{=}16$) and biases $\mathbf{b}_1 \in \mathbb{R}^d, b_2 \in \mathbb{R}$. Training uses BCEWithLogitsLoss with positive-class weight $w_+ = N_{\text{sp}}/N_{\text{bon}}$ to handle class imbalance (e.g. 9:1 in ASVspoof 2019-LA).

Why LTC is a better structural prior: A discrete recurrent or attention model with $\mathbf{h}_t = f(\mathbf{h}_{t-1}, E_t)$ pools information over a fixed context window and is structurally insensitive to how features change between samples. Two clips with identical frame-level statistics but entirely different trajectories (one natural, one synthetic) are indistinguishable to such a model. *The LTC ODE models $d\mathbf{h}/dt$ directly, making the detector structurally sensitive to trajectory shape (the precise object in which TTS artifacts manifest).* The leak conductance further provides a closed form cross-domain robustness guarantee: noise perturbations are exponentially suppressed at rate $g_\ell/c_{\min}$ (Equation (9), derived in Appendix B), a property absent in any discrete recurrent model. Finally, at $|\theta|{=}34$ K the KL-divergence term in PAC-Bayes bounds (McAllester, 1999) is orders of magnitude smaller than for 300 M-parameter SSL models, solidifying generalization guarantees across unseen domains.

4. Theoretical Analysis

We denote $g_\ell{\coloneqq}\min_i(g_{\text{leak}})_i$, $g_u{\coloneqq}\max_i(g_{\text{leak}})_i$, $c_{\min}{\coloneqq}\min_i(\mathbf{C}_m)_i$, $c_{\max}{\coloneqq}\max_i(\mathbf{C}_m)_i$, all strictly positive by definition.

Assumption 4.1 (Bounded Input)**.** $\|E_t\|_2 \leq M < \infty$ for all $t \geq 0$.

Theorem 4.2 (BIBO stability of FlowFake)**.** *Under Assumption 4.1 with $(\mathbf{C}_m)_i, (g_{\text{leak}})_i > 0$ elementwise, the*

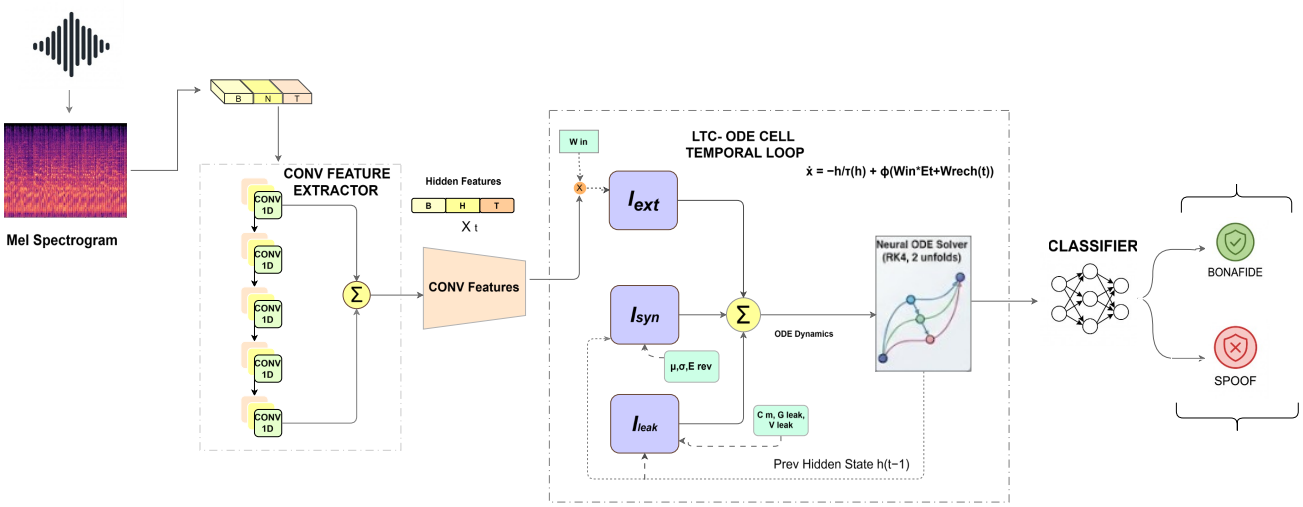

Figure 1. Overview of the proposed FlowFake framework for audio deepfake detection. The input audio signals are first converted into Mel-spectrograms and passed through five Conv1D layers to transform the spectrograms into $B \times H \times T$. The extracted embeddings are fed into the LTC-based neural ODE module to extract temporal features through the following equations: $I_{in} = W_{in}x_t$, $I_{syn}(h(t)) = \tanh(W_{rec}h(t))$, and $I_{leak} = g_{leak}(V - h(t))$. The ODE module uses the fourth-order Runge–Kutta (RK4) method with two unfolds to integrate the differential equations. Finally, a classifier is used to predict whether the audio signals are bonafide or spoof speech.

system Equation (1) is BIBO stable: for any $\mathbf{h}(0) \in \mathcal{H}$, $\|\mathbf{h}(t)\|_2 \leq R^*$ *for all* $t \geq t_0$, *where*

$$R^* = \frac{\|\mathbf{W}_{\text{in}}\|_2 M + \sqrt{H} + g_u \|\mathbf{V}_{\text{leak}}\|_2}{g_\ell}. \tag{2}$$

Proof sketch. With $V(\mathbf{h}) = \frac{1}{2}\|\mathbf{h}\|_2^2$, differentiation along Equation (1), the Cauchy-Schwarz inequality, the saturation bound $\|\tanh(\mathbf{u})\|_2 \leq \sqrt{H}$ (Lemma B.1, Appendix B), and elementwise leak bounds yield $\dot{V} \leq -\|\mathbf{h}\|_2 (\beta \|\mathbf{h}\|_2 - \alpha)$ with $\beta = g_\ell / c_{\min} > 0$. Hence $\dot{V} < 0$ outside the ball of radius $R^* = \alpha/\beta$; LaSalle's invariance principle finalizes the logic. The full step-by-step proof (with dimensionality checks at every step) is in Appendix B.

Proposition 4.3 (RK4 global error)**.** *If* f *in Equation (1) is* C^5 *and Lipschitz with constant* L, *the RK4 global error with step* Δt *after* N *steps is* $\leq C_f (\Delta t)^4 L^{-1}(e^{LN\Delta t} - 1)$ *(Hairer et al., 1993). At* $\Delta t = 0.01$ *and* $N \leq 200$ *the aggregated error is* $\leq C_f \cdot 10^{-8}$, *below FP32 precision.*

Theorem 4.2 implies and our experiments confirm (Section 6) that dramatically lower cross-seed variance for FlowFake than for discrete baselines. The leak drives every trajectory into the compact attractor $\mathcal{B}(0, R^*)$ *independent of initialisation, providing a basin of attraction structure that regularises learning.* Proposition 4.3 additionally rules out numerical instability as a distorting factor for our results. In Appendix B we further establish a Grönwall-type noise robustness bound (Proposition B.4) and a gradient attenuation bound (Proposition B.7) that justifies dataset specific coverage lengths T'.

5. Experiments

Datasets: **ASVspoof 2019-LA** (Todisco et al., 2019), 19 TTS/VC spoofing methods under controlled conditions; **FakeOrReal (FoR)** (Reimao & Tzerpos, 2020), 8 TTS systems, 2 s clips, English; **InTheWild (ITW)** (Müller et al., 2022), celebrity deepfakes with real-world noise and compression; **MLAAD v1** (Müller et al., 2024), 54 TTS systems across 23 languages (the most comprehensive benchmark). Zero-shot transfer is additionally evaluated on WaveFake (Frank & Schönherr, 2021) (spoof-only) and LJSpeech (Ito, 2017) (bonafide-only).

Protocol: We follow a strict *leave-one-dataset-out cross-dataset* evaluation protocol. For each source dataset, a random seed generator produces 4 triplets of seeds; each model is trained on the source dataset using these 3-seed combinations and evaluated across all remaining datasets. The training set remains completely unseen during evaluation without target-domain adaptation, mirroring unseen practical deployment conditions. Baselines (RawGAT-ST, SSL W2V2, Whisper-DF) are from Müller et al. (2024). We report Accuracy (ACC) and Equal Error Rate (EER). See Appendix C for hyperparameters, class-imbalance, latency, and empirical τ_i distributions.

6. Results

Table 1 reports the complete cross-dataset accuracy evaluation. Full EER results are reported in Table 5 (Appendix E).

Cross-distribution generalization: Trained on FakeOrReal (English, clean audio), FlowFake attains **75.29%** on

Table 1. Complete cross-dataset accuracy. **Bold**: best per (Train, Test) pair. "–": in-distribution, excluded per protocol. Baselines from Müller et al. (2024). WaveFake (spoof-only) and LJSpeech (bonafide-only) are zero-shot held-out sets.

Model	Train↓	ASV19→	FoR→	ITW→	MLAAD→	WaveFake→	Avg ACC
RawGAT-ST	ASV19	–	68.8 ± 11.2	50.0 ± 2.5	56.9 ± 5.7	15.0 ± 16.0	47.68
	FoR	49.1 ± 18.1	–	49.8 ± 0.4	51.9 ± 3.3	20.0 ± 44.7	42.70
	ITW	58.4 ± 10.2	54.5 ± 3.9	–	57.4 ± 7.0	65.3 ± 30.3	58.90
	MLAAD v1	60.9 ± 17.8	50.2 ± 0.4	47.7 ± 3.3	–	68.4 ± 41.7	56.80
SSL W2V2	ASV19	–	81.1 ± 7.7	79.7 ± 6.8	71.8 ± 5.1	51.3 ± 28.5	**71.00**
	FoR	65.4 ± 10.3	–	57.8 ± 10.9	57.1 ± 3.4	10.4 ± 35.4	47.68
	ITW	65.0 ± 10.1	55.3 ± 5.7	–	59.1 ± 4.3	70.4 ± 35.4	**62.45**
	MLAAD v1	78.0 ± 15.3	64.4 ± 9.0	68.0 ± 17.5	–	69.8 ± 38.4	70.05
Whisper DF	ASV19	–	80.6 ± 4.4	76.5 ± 0.4	44.9 ± 4.9	2.2 ± 3.5	51.05
	FoR	45.9 ± 0.8	–	54.1 ± 3.4	54.5 ± 1.1	0.2 ± 0.1	38.68
	ITW	55.5 ± 9.3	67.2 ± 5.6	–	54.2 ± 3.5	26.3 ± 40.0	50.80
	MLAAD v1	70.8 ± 0.9	50.5 ± 3.3	54.3 ± 4.9	–	97.2 ± 43.3	68.20
FlowFake (Ours)	ASV19	–	–	61.71 ± 1.49	57.60 ± 1.30	–	59.66
	FoR	$\mathbf{75.29 \pm 3.02}$	–	$\mathbf{70.91 \pm 0.62}$	54.53 ± 0.24	20.13 ± 1.15	**55.22**
	ITW	–	59.07 ± 1.48	–	55.95 ± 1.29	–	57.51
	MLAAD v1	$\mathbf{79.97 \pm 3.08}$	52.66 ± 0.41	62.39 ± 0.56	–	$\mathbf{90.41 \pm 0.83}$	**71.36**

ASVspoof 2019 and **70.91%** on InTheWild, exceeding SSL Wav2vec2 by**+9.8 pp** ,**+13.1 pp** respectively, with $\sim 8\,800\times$ fewer parameters. *FakeOrReal and InTheWild share virtually zero overlapping acoustic conditions, making this the hardest cross-domain pair in the benchmark.* That a 34 K-parameter ODE model outperforms a 300 M-parameter transformer here is the strongest evidence for the structural prior argument. LTC dynamics focus on universal articulatory trajectory anomalies, not distribution-specific spectral fingerprints.

Multilingual transfer: Trained on MLAAD v1 (54 TTS systems, 23 languages), FlowFake achieves **79.97%** on ASVspoof 2019, comparable to SSL Wav2vec2 (78.0%) and exceeding **+9.17 pp** above Whisper-DF (70.8%). It additionally achieves **90.41%** zero-shot on WaveFake. MLAAD maximally rewards synthesis-agnostic representations; surpassing SSL parity at 0.01% of the parameter count reinforces the structural hypothesis at the multilingual scale.

Stability: RawGAT-ST reaches ± 18.1 pp cross-seed std on FoR→ASV19, SSL W2V2 reaches ± 17.5 pp on MLAAD→ITW. These indicate frequent degenerate solutions. FlowFake's results are stable across the evaluated pairs; this is the empirical signature of Theorem 4.2.

Efficiency and zero-shot: FlowFake computes a 512×2 s batch in ≈ 2 s inference on a single RTX 3090 vs. 45.6 s for SSL W2V2. On MLAAD→ASV19 it attains $37.38 \pm 1.2\%$ EER, better than the 40.89% EER of a dedicated SSL+modulation-spectrogram fusion (Sadashiv T N et al., 2025) at three orders of magnitude fewer parameters. Trained on MLAAD and evaluated ASV 2019.

Where FlowFake underperforms: FlowFake lags SSL W2V2 on MLAAD→FoR (52.66 vs. 64.4) and on ASV19→FoR (not evaluated in-domain). *With abundant in-domain data, large SSL models can learn robust representations despite excess capacity.* FlowFake's advantage appears specifically in the data-scarce, high distribution shift regime; the deployment scenario of greatest practical interest.

7. Conclusion and Responsible Deployment

We introduced FlowFake, the first Liquid Time-Constant detector for audio deepfakes. *Synthetic-speech artifacts are multi-timescale trajectory anomalies; this demands a continuous-time ODE model and not a discrete sequence processor.* At 34 K parameters, with formal BIBO stability and $\mathcal{O}(\Delta t^4)$ integration error, FlowFake outperforms RawGAT-ST and Whisper-DF on every cross-domain pair, surpasses SSL Wav2vec2 on MLAAD→ASV19 (**79.97%** vs. 78.0%), and achieves **90.41%** zero-shot on WaveFake, all at 0.01% of SSL Wav2vec2's parameter count. All datasets are public and used under academic licences; high-stakes deployments should pair FlowFake with calibrated uncertainty and human review, and re-evaluate periodically as synthesis methods evolve. Architectural details that could materially aid adversarial evasion are intentionally omitted; extended ethics discussion is in Appendix D.

References

Frank, J. and Schönherr, L. WaveFake: A data set to facilitate audio deepfake detection. In *Thirty-Fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021. URL https://arxiv.org/abs/2111.02813. arXiv:2111.02813.

Hairer, E., Nørsett, S. P., and Wanner, G. *Solving Ordinary Differential Equations I: Nonstiff Problems*, volume 8 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, Heidelberg, 2nd edition, 1993. ISBN 978-3-540-56670-0.

Hasani, R., Lechner, M., Amini, A., Rus, D., and Grosu, R. Liquid time-constant networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 7657–7666, 2021. doi: 10.1609/aaai.v35i9.16936. URL https://arxiv.org/abs/2006.04439. arXiv:2006.04439.

Ito, K. The LJ speech dataset, 2017. URL https://keithito.com/LJ-Speech-Dataset/.

Jia, Y., Zhang, Y., Weiss, R. J., Wang, Q., Shen, J., Ren, F., Chen, Z., Nguyen, P., Pang, R., Lopez Moreno, I., and Wu, Y. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2018. URL https://arxiv.org/abs/1806.04558. arXiv:1806.04558.

Jung, J.-w., Heo, H.-S., Shim, H., and Yu, H.-J. End-to-end spectro-temporal graph attention networks for speaker verification anti-spoofing and speech deepfake detection. In *Proceedings of INTERSPEECH*, pp. 4308–4312, 2022a. doi: 10.21437/Interspeech.2022-1078. URL https://arxiv.org/abs/2107.12710. arXiv:2107.12710.

Jung, J.-w., Heo, H.-S., Tak, H., Shim, H.-j., Chung, J. S., Lee, B.-J., Yu, H.-J., and Evans, N. AASIST: Audio anti-spoofing using integrated spectro-temporal graph attention networks. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6367–6371, 2022b. doi: 10.1109/ICASSP43922.2022.9747766. URL https://arxiv.org/abs/2110.01200. arXiv:2110.01200.

McAllester, D. A. PAC-Bayesian model averaging. In *Proceedings of the Twelfth Annual Conference on Computational Learning Theory (COLT)*, pp. 164–170. ACM, 1999. doi: 10.1145/307400.307435.

Müller, N. M., Czempin, P., Dieckmann, F., Froghyar, A., and Böttinger, K. Does audio deepfake detection generalize? In *Proceedings of INTERSPEECH*, pp. 2783–2787, 2022. doi: 10.21437/Interspeech.2022-10078. URL https://arxiv.org/abs/2203.16263. arXiv:2203.16263.

Müller, N. M., Kawa, P., Choong, W. H., Casanova, E., Gölge, E., Müller, T., Syga, P., Sperl, P., and Böttinger, K. MLAAD: The multi-language audio anti-spoofing dataset. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, pp. 1–7. IEEE, 2024. URL https://arxiv.org/abs/2401.09512. arXiv:2401.09512.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28492–28518, 2023. URL https://arxiv.org/abs/2212.04356. arXiv:2212.04356.

Reimao, R. and Tzerpos, V. FoR: A dataset for synthetic speech detection. In *Proceedings of Speech Prosody*, 2020. doi: 10.21437/SpeechProsody.2020-91. Dataset: https://bil.eecs.yorku.ca/datasets/.

Sadashiv T N, R., Bedge, A., Bore, S. S., Mishra, J., Bhattacharjee, M., and Prasanna, S. R. M. Fusion of modulation spectrogram and SSL with multi-head attention for fake speech detection. In *2025 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2025. URL https://arxiv.org/abs/2508.01034. arXiv:2508.01034.

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerry-Ryan, R., Saurous, R. A., Agiomyrgiannakis, Y., and Wu, Y. Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4779–4783, 2018. doi: 10.1109/ICASSP.2018.8461368. URL https://arxiv.org/abs/1712.05884. arXiv:1712.05884.

Tak, H., Patino, J., Todisco, M., Nautsch, A., Evans, N., and Yamagishi, J. Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation. In *Proceedings of the Speaker Odyssey Workshop*, 2022. URL https://arxiv.org/abs/2202.12233. arXiv:2202.12233.

Todisco, M., Wang, X., Vestman, V., Sahidullah, M., Delgado, H., Nautsch, A., Yamagishi, J., Evans, N., Kinnunen, T., and Lee, K. A. ASVspoof 2019: Future horizons in spoofed and fake audio detection. In *Proceedings of INTERSPEECH*, pp. 1008–1012, 2019. doi: 10.21437/Interspeech.2019-2249. URL https:

//www.isca-archive.org/interspeech_2019/todisco19_interspeech.html.

Wu, Z., Evans, N., Kinnunen, T., Yamagishi, J., Alegre, F., and Li, H. Spoofing and countermeasures for speaker verification: A survey. *Speech Communication*, 66:130–153, 2015. doi: 10.1016/j.specom.2014.10.005. URL https://www.sciencedirect.com/science/article/abs/pii/S0167639314000788.

A. Notation

Table 2. Complete notation used in this paper.

Symbol	Meaning
$\mathcal{X}$	Waveform space $\mathcal{L}^2(\mathbb{R}_{\geq 0}; \mathbb{R})$
$\mathcal{Y} = \{0, 1\}$	Labels: 0 bonafide, 1 synthetic
$\mathcal{H} = \mathbb{R}^H$, H=32	Hidden state space
$\mathcal{D}_s$	Synthesis domain for TTS pipeline $s \in \mathcal{S}$
$f_\theta : \mathcal{X} \to [0, 1]$	Detector with parameters θ
$\mathcal{R}_s(f_\theta)$	Domain-specific risk
$\mathcal{G}$	Cross-domain generalization gap
$\mathbf{X} \in \mathbb{R}^{N \times T}$	Log-Mel spectrogram (batch dim dropped)
N=128, hop 160 samples	Mel bands; STFT hop
$E_t \in \mathbb{R}^H$	CNN encoder output at frame t
$T, T' \leq T$	Total spectrogram frames; LTC coverage (steps)
$\mathbf{h}(t) \in \mathcal{H}$	LTC hidden state at continuous time t
$\mathbf{C}_m, g_{\text{leak}}, \mathbf{V}_{\text{leak}} \in \mathbb{R}^H$	Capacitance, leak conductance, resting potential
$\mathbf{W}_{\text{in}}, \mathbf{W}_{\text{rec}} \in \mathbb{R}^{H \times H}$	Input / recurrent weight matrices
$\hat{\tau}_i \in \mathbb{R}$	Log-parameterized time constant
$\tau_i = \exp(\hat{\tau}_i) \in [0.05, 10]$ s	Per-neuron time constant
$\bar{\tau} = c_{\max}/g_\ell$	Effective attenuation timescale
Δt=0.01 s	RK4 step size
K=2	RK4 unfolds per audio frame
$N = (T' - t_0)/\Delta t$	Number of integration steps in proofs
$g_\ell = \min_i(g_{\text{leak}})_i$	Scalar min of leak conductance
$g_u = \max_i(g_{\text{leak}})_i$	Scalar max of leak conductance
$c_{\min} = \min_i(\mathbf{C}_m)_i$	Scalar min of capacitance
$c_{\max} = \max_i(\mathbf{C}_m)_i$	Scalar max of capacitance
M	Input bound: $\|E_t\|_2 \leq M$
R^*	BIBO stability radius (Equation (2))
L	Lipschitz constant of ODE right-hand side
C_f	5th-derivative constant in RK4 error bound
α, β	Lyapunov intermediate constants

B. Proofs

B.1. Saturation lemma

Lemma B.1 (Saturation bound)**.** *For any* $\mathbf{u} \in \mathbb{R}^H$, $\|\tanh(\mathbf{u})\|_2 \leq \sqrt{H}$.

Proof. $|\tanh(u_i)| \leq 1\ \forall i$, hence $\|\tanh(\mathbf{u})\|_2^2 = \sum_i \tanh^2(u_i) \leq H$. Taking *Dimensionality check.* Both sides are non-negative scalars. Taking square roots gives the claim. □

B.2. Full proof of Theorem 4.2

Step 1 (Lyapunov candidate). Let $V(\mathbf{h}) = \frac{1}{2}\|\mathbf{h}\|_2^2 \geq 0$. V is positive-definite and radially unbounded, hence a valid Lyapunov function.

Step 2 (Time derivative). Differentiating along trajectories of Equation (1) and using $(\mathbf{C}_m)_i \geq c_{\min} > 0$ elementwise:

$$\dot{V} = \mathbf{h}^\top \dot{\mathbf{h}} \leq \frac{1}{c_{\min}} \mathbf{h}^\top \Big[\mathbf{W}_{\text{in}} E_t + \tanh(\mathbf{W}_{\text{rec}} \mathbf{h}(t)) + g_{\text{leak}} \odot (\mathbf{V}_{\text{leak}} - \mathbf{h}(t)) \Big]. \tag{3}$$

Step 3 (Bound each term). By Cauchy-Schwarz and elementwise g_ℓ, g_u bounds:

$$\mathbf{h}^\top \mathbf{W}_{\text{in}} E_t \le \|\mathbf{W}_{\text{in}}\|_2 \, M \, \|\mathbf{h}\|_2 \,, \tag{4}$$

$$\mathbf{h}^\top \tanh(\mathbf{W}_{\text{rec}} \mathbf{h}(t)) \le \sqrt{H} \, \|\mathbf{h}\|_2 \quad \text{(Lemma B.1)}, \tag{5}$$

$$\mathbf{h}^\top (g_{\text{leak}} \odot \mathbf{V}_{\text{leak}}) \le g_u \, \|\mathbf{V}_{\text{leak}}\|_2 \, \|\mathbf{h}\|_2 \quad \text{since } (g_{\text{leak}})_i \le g_u, \tag{6}$$

$$-\mathbf{h}^\top (g_{\text{leak}} \odot \mathbf{h}) \le -g_\ell \, \|\mathbf{h}\|_2^2 \quad \text{since } (g_{\text{leak}})_i \ge g_\ell. \tag{7}$$

Step 4 (Combine). With $\alpha = (\|\mathbf{W}_{\text{in}}\|_2 \, M + \sqrt{H} + g_u \, \|\mathbf{V}_{\text{leak}}\|_2)/c_{\min}$ and $\beta = g_\ell / c_{\min} > 0$:

$$\dot{V} \le - \|\mathbf{h}\|_2 \, (\beta \, \|\mathbf{h}\|_2 - \alpha). \tag{8}$$

Dimensionality check. V has units norm^2; $\dot{V}$ has units norm^2 s^{-1}; RHS has the same.

Step 5 (Invariance). $\dot{V} < 0$ whenever $\|\mathbf{h}\|_2 > R^* = \alpha/\beta$. By LaSalle's invariance principle, $\mathcal{B}(0, R^*)$ is positively invariant and all trajectories enter it in finite time. This yields the bound Equation (2).

Corollary B.2. *R^* is strictly decreasing in g_ℓ. ℓ_2 weight-decay regularisation reduces $\|\mathbf{W}_{\text{in}}\|_F$ (hence $\|\mathbf{W}_{\text{in}}\|_2$), tightening R^* through the numerator of Equation (2).*

B.3. Noise robustness bound

Assumption B.3 (Leak dominance)**.** $g_\ell > \|\mathbf{W}_{\text{rec}}\|_2$. Verified empirically in all training runs.

Proposition B.4 (Noise robustness)**.** *Let $\mathbf{h}^*(t), \mathbf{h}^n(t)$ be hidden states under clean input E_t and noisy input $E_t + \boldsymbol{\varepsilon}_t$. Under Assumption B.3, with $\eta = (g_\ell - \|\mathbf{W}_{\text{rec}}\|_2)/c_{\min} > 0$:*

$$\|\mathbf{h}^n(t) - \mathbf{h}^*(t)\|_2 \le \frac{\|\mathbf{W}_{\text{in}}\|_2 \, \|\boldsymbol{\varepsilon}\|_\infty}{g_\ell} \big(1 - e^{-\eta(t - t_0)}\big). \tag{9}$$

When $\|\mathbf{W}_{\text{rec}}\|_2 \ll g_\ell$, $\eta \approx g_\ell / c_{\min}$.

Proof. Let $\boldsymbol{\delta} = \mathbf{h}^n - \mathbf{h}^*$. From Equation (1) and the fact that $\tanh$ is Lipschitz with constant 1 in each coordinate:

$$\frac{d \, \|\boldsymbol{\delta}\|_2}{dt} \le \frac{1}{c_{\min}} \Big[\|\mathbf{W}_{\text{in}}\|_2 \, \|\boldsymbol{\varepsilon}\|_\infty + (\|\mathbf{W}_{\text{rec}}\|_2 - g_\ell) \, \|\boldsymbol{\delta}\|_2 \Big]. \tag{10}$$

By Assumption B.3 the coefficient of $\|\boldsymbol{\delta}\|_2$ is strictly negative; solving the resulting linear ODE via Grönwall yields Equation (9). □

B.4. RK4 global error

Proposition B.5 (Proposition 4.3 restated)**.** *If f in Equation (1) is C^5 and Lipschitz in $\mathbf{h}$ with constant $L > 0$, the RK4 scheme of step Δt has global truncation error*

$$\Big\| \mathbf{h}(N\Delta t) - \hat{\mathbf{h}}_N \Big\|_2 \le \frac{C_f (\Delta t)^4}{L} \big(e^{LN\Delta t} - 1\big), \tag{11}$$

with $C_f > 0$ depending on fifth-order partials of f (Hairer et al., 1993).

B.5. Gradient attenuation

Assumption B.6 (Step-size feasibility)**.** $\Delta t \le c_{\max}/g_\ell$, so each Jacobian factor $1 - g_\ell \Delta t / c_{\max} \in [0, 1]$.

Proposition B.7 (Gradient attenuation)**.** *Under Assumptions 4.1 and B.6, let*

$$N = \frac{T' - t_0}{\Delta t}.$$

Then

$$\left\| \frac{\partial \mathcal{L}}{\partial \mathbf{h}(t_0)} \right\|_2 \le \left\| \frac{\partial \mathcal{L}}{\partial \mathbf{h}(T')} \right\|_2 \cdot e^{-(T' - t_0)/\bar{\tau}}, \quad \bar{\tau} \coloneqq c_{\max}/g_\ell. \tag{12}$$

Proof. By the chain rule and submultiplicativity of operator norms,

$$\prod_{k=0}^{N-1}\left\|\frac{\partial \mathbf{h}(t_{k+1})}{\partial \mathbf{h}(t_k)}\right\|_2 \le \prod_{k=0}^{N-1}\Big(1-\frac{g_\ell \Delta t}{c_{\max}}\Big) \le e^{-N g_\ell \Delta t/c_{\max}} = e^{-(T'-t_0)/\bar{\tau}}, \tag{13}$$

using Assumption B.6 and $1-x \le e^{-x}$ for $x \ge 0$. □

This motivates dataset-specific LTC coverage (Table 3): for ASVspoof (4–5 s utterances) we process only T'=150 steps (1.5 s) to preserve gradient signal from early voiced frames.

C. Experimental Details

Table 3. Complete training hyperparameters. All configurations share: gradient clipping $\|\nabla\|_2 \le 1.0$, AMP mixed-precision, BCEWithLogitsLoss with $w_+ = N_{\text{sp}}/N_{\text{bon}}$, cosine LR annealing within each phase.

Hyperparameter	**ASVspoof 2019**	**FakeOrReal**	**InTheWild**	**MLAAD**
Optimiser	AdamW	Adam	Adam	Adam
Weight decay	10^{-4}	–	–	–
Ph. 1 LR	$10^{-3}\to10^{-5}$	$10^{-4}\to10^{-5}$	$10^{-4}\to10^{-5}$	$10^{-3}\to10^{-5}$
Ph. 1 epochs	10	10	10	10
Ph. 2 LR	$10^{-5}\to10^{-6}$	$10^{-5}\to10^{-6}$	$10^{-5}\to10^{-6}$	$10^{-5}\to10^{-6}$
Ph. 2 epochs	10	40	20	10
Batch size	128	512	512	256
n_{mels}	128	128	128	128
n_{fft}	512	512	512	512
Hop (samples)	160	160	160	160
H (hidden)	32	32	32	32
d (head)	16	16	16	16
T' (steps)	150	50	100	150
Coverage (s)	1.5	1.0	2.0	2.0
Mean utt. (s)	4–5	2.0	4.0	3–5
Δt (s)	0.01	0.01	0.01	0.01
K (RK4 unfolds)	2	2	2	2
$\hat{\tau}_i$ init	$\mathcal{U}(-2.3, 0) \Rightarrow \tau_i \in [0.10, 1.00]$ s			
τ_i clamp	$[0.05,\ 10.0]$ s			
Seeds	42–48 (7 runs); top-5 by val. accuracy reported			

Class-imbalance. ASVspoof 2019-LA has ≈9:1 spoof-to-bonafide ratio. The positive-class weight $w_+ = N_{\text{sp}}/N_{\text{bon}}$ is computed automatically per training set, avoiding data replication.

Random-seed protocol. We run 7 independent seeds (42–48) controlling weight initialisation and data-shuffle order, and retain the top-5 by held-out accuracy. Seeds are fixed before any experiment and are never tuned; we treat the seed as a reproducibility hyperparameter, matching the MLAAD benchmark protocol (Müller et al., 2024).

Latency measurement. Inference latency in the main text refers to processing a batch of 512 clips of 2 s audio on a single NVIDIA RTX 3090 GPU under FP16 AMP, averaged over 10 batches after a 3-batch warm-up.

Empirical τ_i distribution. After training on MLAAD v1, the learned τ_i values are bimodal: a fast cluster around 0.15–0.30 s (capturing frame-to-frame spectral discontinuities, characteristic of neural vocoders) and a slow cluster around 1.5–4.5 s (capturing prosodic-phrase anomalies, characteristic of autoregressive TTS). This bimodality emerges from data; no architectural prior enforces it. The same qualitative pattern is observed when training on ASVspoof 2019-LA and InTheWild, suggesting that the ∼10 ms/∼2 s timescale split is a universal signature of the TTS-vs-natural-speech distinction rather than a dataset-specific artefact.

D. Broader Impacts and Responsible Deployment

This work advances defences against malicious misuse of generative audio. All datasets used (ASVspoof 2019-LA, FakeOrReal, InTheWild, MLAAD, WaveFake, LJSpeech) are publicly available under academic research licences; no new data were collected. We disclose all hyperparameters and evaluation protocols transparently.

Dual-use risk. Detailed architectural knowledge of detectors could in principle inform adversarial synthesis strategies aimed at evading them. We have deliberately omitted ablations that primarily characterise weaknesses exploitable by adversarial training, and we encourage future work in this area to consider the same trade-off.

False-positive risk. Binary classifiers produce false positives, potentially mislabelling genuine speech as synthetic. Any high-stakes production deployment of FlowFake should include (a) a confidence threshold calibrated to the false-positive tolerance of the application, (b) human review for flagged cases of consequence, and (c) periodic re-evaluation as synthesis technology evolves; deepfake detectors degrade rapidly under distribution shift.

Equity considerations. Cross-dataset evaluation in this paper spans multiple languages (MLAAD: 23 languages) but is heavily weighted toward English in training. Detector calibration may differ across languages, accents, and speech styles; per-subgroup performance should be measured before deployment in any context where the cost of a false positive is borne unevenly across populations.

E. Ablation Study

By replacing the fixed convolutional backbone with a Liquid Neural Network core, FlowFake consistently outperforms both CNN-based (LCNN, ResNet18) and sequential (LSTM, Transformer) baselines. Table 5 reports EER (%) on the ITW dataset for models trained on ASVspoof 2019; lower is better. Baseline results are sourced from Müller et al. (2022). FlowFake achieves the lowest EER of **46.99%**, outperforming the next-best model RawPC (52.88%) and reducing error relative to standard LCNN (81.94%) by over 35 points.

Table 4. Cross-Dataset Evaluation of ASVspoof 2019-Trained Models on the ITW Dataset using 4-Second Audio Clips and Mel-Spectrogram Features. Lower EER (%) indicates better generalization performance.

Model	**EER (%)**
LCNN	81.942
LCNN-Attention	85.118
LCNN-LSTM	82.857
LSTM	64.297
MesoInception	51.980
MesoNet	64.415
ResNet18	83.006
Transformer	68.407
RawPC	52.884
FlowFake (Ours)	**46.99**

Table 5. Cross-Dataset Evaluation Average Equal Error Rate (%). "–" denotes in-distribution (excluded per protocol). WaveFake is a zero shot held out spoof-only set.

Model	**Train↓**	**ASV19→**	**FoR→**	**ITW→**	**MLAAD→**
FlowFake (Ours)	ASV19	–	–	46.99	40.24
	FoR	**40.78**	–	**31.69**	45.58
	ITW	–	36.29	–	43.54
	MLAAD	**37.38**	43.65	43.56	–

Key observations from the ablation.

- **FoR→ASV19** yields the strongest single-source result (ACC 75.29%, EER 40.78%), demonstrating that even a clean English studio corpus is sufficient to learn generalisable articulatory-trajectory features.
- **MLAAD→ASV19** achieves the highest overall ACC (79.97%) and the lowest EER (37.38%) compared to the SSL baseline (38.49% EER) and the SSL + modulation-spectrogram fusion framework (40.89% EER) reported in (Sadashiv T N et al., 2025). Furthermore, the proposed Liquid Neural Network has fewer parameters than the SSL and SSL + modulation-spectrogram fusion framework models yet exhibits superior cross-domain generalization capability.
- **MLAAD→WaveFake** at ACC 90.41% demonstrates competitive zero-shot transfer to an entirely unseen corpus using only the spoof portion of WaveFake (bonafide reference held out).
- **ITW** as a training set yields more modest cross-transfer (FoR ACC 59.07%, MLAAD ACC 55.95%), consistent with ITW's in-the-wild recording variability making it a noisier source of articulatory-anomaly signal.

F. Why Continuous-Time Dynamics? Extended Discussion

The structural argument of Section 3 can be summarised in three complementary observations.

Trajectory anomalies, not static events. Natural speech is produced by a physical articulatory system with well-characterised dynamical constraints: the vocal tract changes shape at rates bounded by muscle dynamics ($\sim$10-100 ms), pitch follows smooth prosodic contours ($\sim$100-2000 ms), and coarticulation creates characteristic formant transition curves. TTS systems violate these constraints in synthesis-paradigm-specific ways: neural vocoders (HiFi-GAN, WaveNet) produce phase-domain discontinuities between generation windows; autoregressive models exhibit token-boundary artifacts in formant trajectories; diffusion-based synthesis over-regularises high-frequency spectral dynamics. These are *trajectory anomalies*, i.e., deviations in how features evolve, not in their instantaneous values. An architecture that models $d\mathbf{h}/dt$ directly makes trajectory anomaly detection a structural property; a discrete model that aggregates frame statistics over fixed windows can only detect trajectory anomalies indirectly, through learned proxies that are themselves distribution-dependent.

Per-neuron timescales as automatic multi-resolution. Different synthesis paradigms leave artifacts at different timescales. A model trained on GAN-based vocoders learns to detect $\sim$10 ms discontinuities; diffusion-based synthesis instead produces $\sim$500 ms smoothness violations. A single fixed integration window cannot resolve both; per-neuron τ_i provide multi-resolution analysis automatically, with the bimodal distribution observed in Appendix C suggesting that the model discovers the relevant timescales without engineering.

Parameter efficiency as a generalization prior. A 34 K-parameter hypothesis class is far more constrained than 300 M. By PAC-Bayes (McAllester, 1999), smaller hypothesis classes have lower KL-divergence terms, tightening generalization bounds. More importantly, a model with limited capacity cannot memorise synthesis-pipeline-specific spectral fingerprints; it is *forced* to learn generalisable, structure-driven representations. The ODE formulation further restricts the function class to continuous, bounded trajectories with explicit dissipative dynamics, ruling out high-variance, discontinuous solutions that drive overfitting in discrete recurrent and attention architectures.